Bioinformatic approaches to determine pathogenicity and function of clinical genetic variants across ion channels and neurodevelopmental disorder associated genes

Inaugural-Dissertation

Zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Tobias Brünger

aus Brühl

Köln, 2023

Berichterstatter:

Prof. Dr. Michael Nothnagel

Prof. Dr. Kay Hofmann

Prof. Dr. Peter Krawitz

Vorsitz der Prüfung:

Prof. Dr. Günter Schwarz

Beisitzerin:

Dr. Birgit Budde

Tag der mündlichen Prüfung:

25.09.2023

Contents

1. List of abbreviations 3

2. Abstract 4

2.1 Zusammenfassung 7

3. Introduction 11

3.1 Clinical genetic testing and variant interpretation in rare monogenic diseases: current state, and its limitations 11

3.1.1 Guideline-based classification of the genetic variant pathogenicity 12

3.1.2 Clinical classification of variant pathogenicity for missense variants remains a major challenge. 14

3.1.3 For successfully pathogenic classified variants, the next frontier presents the classification of the variant effect on protein function 15

3.2 Bioinformatic approaches to overcome challenges in missense variant pathogenicity classification and function prediction. 16

3.2.1 Variants from large patient and population databases can identify functionally important protein regions and inform variant pathogenicity 17

3.2.2 Evolutionary conservation and molecular functional redundancy across paralogous genes can inform variant pathogenicity and function 18

3.2.3 Systematic assessment of genetic variants in a 3D-structural context can identify functionally important protein regions and inform variant pathogenicity 20

3.3 Selection of studied disease cohorts to develop and test novel bioinformatic approaches 21

3.3.1 Challenges in the interpretation of genetic variants in ion channels 22

3.3.2 Challenges in the interpretation of genetic variants in voltage-gated sodium channels 24

3.3.3 Challenges in the interpretation of genetic variants associated with neurodevelopmental disorders 24

4. Goals and objectives 26

5. Preface 29

6. Publications 32

6.1 Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria 32

6.2 Gene variant effects across sodium channelopathies predict function and guide precision therapy 71

6.3 Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes 97

6.4 Delineation of functionally essential protein regions for 242 neurodevelopmental genes 121

7. Discussion 182

7.1 Conserved structural regions inform variant pathogenicity and protein function 184

7.2 Knowledge transfer from established pathogenic and functionally tested variants to variants in paralogous genes informs variant pathogenicity and protein function 185

7.3 Essential sites identified through a combination of evolutionary and structural features inform variant pathogenicity and provide insights into molecular disease mechanisms 187

7.4 Web-based tools improve access to novel scores and annotations 187

7.5 Summary of the conducted research outlined in this thesis 188

8. References 190

9. Appendix 201

9.1 Statutory declaration and statement 201

9.2 List of publications 202

1. List of abbreviations

3D = Three dimensional

ACMG = The American College of Medical Genetics and Genomics

DMS = Deep mutational scanning

GABA = gamma-Aminobutyric acid

gnomAD = Genome Aggregation Database

GoF = Gain of function

HGMD = Human Gene Mutation Database

LoF = Loss of function

NDD = Neurodevelopmental disorder

OR = Odds ratio

PDB = Protein data bank

SNV = Single nucleotide variant

VGSC = Voltage-gated sodium channel

VUS = Variants of uncertain significance

2. Abstract

Clinical genetic testing for rare monogenic diseases has the scope of identifying the disease-causing variants. Identification of the molecular etiology of the disease can already today improve clinical care and is essential for the administration of precision medicines that are currently in development for many disorders. However, distinguishing pathogenic variants from benign genetic variants remains a challenge – in particular for missense variants where a single amino acid is substituted. The effects of a pathogenic variant on the protein function, for example, whether it causes a gain (GoF) or a loss (LoF) of the protein function, is most of the time not understood since most genetic variants are ultra-rare and have not been molecularly tested. In particular, for genes associated with severe developmental disorders, first-generation symptomatic treatments offer often only limited relief. Consequently, the development and application of targeted treatments that promise improvement is urgently needed. Identifying the disease-causing pathogenic and predicting their function is crucial as targeted therapies can only be administered to patients with classified pathogenic variants whose functional effects are known to avoid adverse treatment outcomes. In this dissertation, I present bioinformatic approaches to enhance the assessment of variant pathogenicity and understanding of the functional effects of genetic variants. The developed approaches were applied on an exome-wide scale using public datasets and for selected disorders for which I had expert-curated clinical-genetic data available from collaborators. The major focus of this thesis is on genes implicated in neurodevelopmental disorders and diseases associated with ion channel dysfunction for which collaboration with other research groups enabled the aggregation of required genetic, clinical, and functional datasets to develop and test the bioinformatic approaches.

In the first study (Bruenger and Ivaniuk et al., in preparation for submission to Genetics in Medicine), we developed a novel approach to extend the application of current variant interpretation guidelines as proposed by the American College of Medical Genetics and Genomics (ACMG). Currently, a major limitation of interpreting the pathogenicity of variants with the ACMG guidelines presents the rare applicability of some of the proposed evidence criteria. We evaluated the potential of incorporating individual pathogenic variants observed in paralogous genes to extend the applicability of two criteria of the guidelines. Our results demonstrated that pathogenic variants in evolutionarily conserved paralogous genes can serve as evidence for a variant's pathogenicity and thus extend the current criteria's applicability by more than four times. We further explored whether the selection of the paralogous pathogenic variants can be improved by incorporating phenotype information. We assembled a clinically well-defined cohort of patients with variants in voltage-gated sodium channels (VGSC) and identified phenotype correlations among paralogous genes based on the shared variant

properties. By integrating these phenotype correlations into our proposed extension of the ACMG criteria, we demonstrated an enhanced ability to provide evidence for the pathogenicity of genetic variants in VGSC-encoding genes.

In the second study (Brunklaus, Feng, and Bruenger et al., Brain, 2022), we examined whether experimentally obtained functional effects of variants in one VGSC encoding gene could predict function in conserved variants in paralogous genes with high sequence similarity. We aggregated 437 *in-vitro* functionally tested variants from an intensive literature search and found that the functional effect across conserved variants in paralogous genes was conserved in 94% of cases. Our findings represent the first GoF versus LoF topological map of VGSC proteins, which could guide precision therapy as functionally tested variants are rare across VGSC. We integrated our findings into a publicly accessible webtool (http://SCN-viewer.broadinstitute.org) to facilitate functional variant interpretation across VGSC.

In the third study (Bruenger et al., Brain, 2022), we systematically identified biological properties associated with variant pathogenicity across all major voltage and ligand-gated ion-channel families. We discovered and independently replicated that several pore residue properties and proximity to the pore axis were significantly enriched for pathogenic variants compared to population variants across all ion channels. Using a newly developed structural framework, we provide quantitative evidence that variants at the pore showed the strongest pathogenic variant enrichment. Moreover, we found that a hydrophobic pore environment was most strongly associated with variant pathogenicity. Finally, we showed that the identified biological properties correlated with *in-vitro* functional readouts from 679 variants and clinical phenotypes in 1,422 patients with neurodevelopmental disorders which were collected through collaboration with other research groups. In summary, we identified biological properties associated with ion-channel malfunction and show that these are correlated with in vitro functional readouts and clinical phenotypes in patients with neurodevelopmental disorders. Our results suggest that clinical decision support algorithms that predict variant pathogenicity and function are feasible in the future.

In the fourth study (Iqbal and Bruenger et al., Brain, 2022), we developed a novel consensus approach that combines evolutionary and population-based genomic scores to identify 3D essential sites (Essential3D) on protein structures encoded by genes associated with neurodevelopmental disorders (NDDs). NDDs encompass severe clinical conditions caused by pathogenic variants in different genes. However, many of those genes were just recently associated with NDDs and are not well studied. We identified 14,377 Essential3D sites on protein structures encoded by 189 genes and found that these sites were eight-fold enriched for pathogenic versus population controls in an independent cohort of over 360,000 patient and population variants. The Essential3D sites offer insights into molecular mechanisms of

protein function, such as key protein-protein interaction sites. The provided annotations are available at https://es-ndd.broadinstitute.org and will guide clinical variant interpretation.

In summary, within these major studies in my Ph.D., we aggregated genetic, clinical, and functional datasets and developed bioinformatic approaches to enhance the assessment of variant pathogenicity and improve understanding of the functional effects of genetic variants on protein function. The advances made during my Ph.D. research demonstrate the power of integrating multiple data sources to study novel genetic variants and their implication for rare monogenic diseases. Our approaches specifically improve variant function and pathogenicity assessment in genes implicated in several severe diseases for which currently applied first-generation therapies cannot adequately lower the disease burden. Thus, our results contribute to a new era in precision medicine, where personalized treatments and improved clinical care become increasingly accessible to patients. Finally, the annotations developed in these can serve as a foundation for further studies, including the application of machine learning methods to predict variant pathogenicity and protein functional effects more accurately.

2.1 Zusammenfassung

Das Ziel der genetischen Sequenzierung von Patienten, welche womöglich von einer seltenen monogenen Krankheit betroffen sind, ist die Identifizierung der krankheitsverursachenden genetischen Varianten. Die Erkennung der pathogenen, krankheisterregenden, Varianten und weitergehend das Verständnis der molekularen Ursachen der Krankheit kann bereits heute die klinische Versorgung der Patienten verbessern und ist weiterhin für die Anwendung von zielgerichteten Therapien, die derzeit für viele Krankheiten entwickelt werden, unerlässlich. Zwischen pathogenen und gutartigen genetischen Varianten, welche die Mehrheit aller Identifizierten Varianten darstellt, zu unterscheiden stellt jedoch nach wie vor eine Herausforderung dar. Dies gilt insbesondere für Missense Varianten, bei denen im nur eine einzelne Aminosäure ausgetauscht ist. Die Auswirkungen einer pathogenen Variante auf die Proteinfunktion, z.B. ob sie zu einer Steigerung oder eine Verringerung der Proteinfunktion führt, sind meist nicht bekannt da die molekularen Effekte der meisten genetischen Varianten noch nicht untersucht wurden. Insbesondere bei Genen, bei denen einzelne genetische Varianten schweren Entwicklungsstörungen hervorrufen können, bieten symptomatische Behandlungen der ersten Generation allerdings oft nur begrenzte Linderung der Krankheit. Daher sind die Entwicklung und Anwendung gezielter Behandlungen, die signifikante Verbesserungen im Krankheitsbild versprechen, dringend erforderlich. Zielgerichtete Therapien kommen nur für Patienten mit klassifizierten pathogenen Varianten in Frage deren funktionelle Auswirkungen bekannt sind, um negative Behandlungsergebnisse zu vermeiden. In dieser Dissertation stelle ich bioinformatische Ansätze vor, die die Bewertung der Pathogenität von Varianten unterstützen und das Verständnis der funktionellen Auswirkungen verbessern. Die entwickelten Ansätze wurden sowohl Exom weit mit öffentlichen zugänglichen Datensätzen als auch begrenzt für ausgewählte Erkrankungen angewandt, für die mir von Experten kuratierte klinisch-genetische Daten zur Verfügung standen. Der Schwerpunkt in dieser Arbeit liegt dabei auf Erkrankungen, die mit neurologischen Entwicklungsstörungen und einer Ionenkanaldysfunktion in Verbindung stehen. Hierbei ermöglichte die Zusammenarbeit mit anderen Forschungsgruppen die Zusammenstellung der erforderlichen genetischen, klinischen und funktionellen Datensätze die zur Entwicklung und Validierung der bioinformatischen Ansätze notwendig waren.

In der ersten Studie (Bruenger und Ivaniuk et al., in Vorbereitung für die Veröffentlichung in Genetics in Medicine) entwickelten wir einen neuartigen Ansatz, um die Anwendung der aktuellen Richtlinien zur Interpretation von Varianten, wie sie vom American College of Medical Genetics and Genomics (ACMG) vorgeschlagen wurden, zu erweitern. Eine wesentliche Einschränkung bei der Interpretation der Pathogenität von Varianten anhand der ACMG-Richtlinien ist die seltene Anwendbarkeit einiger der vorgeschlagenen Evidenzkriterien. In

unserer Studie prüften wir, ob die Einbeziehung einzelner pathogener Varianten, die in paralogen Genen gefunden wurden, die Anwendbarkeit von zwei dieser Kriterien erhöhen. Dabei zeigten unsere Ergebnisse, dass pathogene Varianten in evolutionär konservierten paralogen Genen als Evidenzkriterium für die Pathogenität einer Variante dienen und die Anwendbarkeit der aktuellen ACMG-Kriterien um mehr als das Vierfache erweitern. Darüber hinaus haben wir untersucht, ob die Auswahl der paralogen pathogenen Varianten, die für die erweiterte Anwendung der ACMG-Kriterien verwendet werden, durch die Einbeziehung von Informationen über den Phänotyp der mit den jeweiligen Varianten assoziiert ist verbessert werden kann. Dafür stellten wir eine klinische Kohorte von Patienten mit Varianten in spannungsabhängigen Natriumkanälen (VGSC) zusammen und identifizierten Phänotyp-Korrelationen zwischen paralogen Genen auf der Grundlage von gemeinsamen Varianteneigenschaften. Die Integration dieser Phänotyp-Korrelationen in die von uns vorgeschlagene Erweiterung der ACMG-Kriterien zeigte, dass dies die Bestimmung der Pathogenität von genetischen Varianten in VGSC-kodierenden Genen verbessern kann.

In der zweiten Studie (Brunklaus, Feng und Bruenger et al., Brain, 2022) untersuchten wir, ob eine Reihe von experimentell ermittelten funktionelle Effekte von Varianten in einem VGSC-kodierenden Gen die Funktion von konservierten Varianten in paralogen Genen, welche nicht zuvor funktionell getestet wurden, vorhersagen können. Dafür aggregierten wir 437 *in vitro* funktionell getestete Varianten durch intensive Literaturrecherche. Wir fanden heraus, dass der funktionelle Effekt, den genetische Varianten in einem Protein hervorrufen können, zwischen evolutionär-konservierten Varianten in paralogen Genen in 94 % der Fälle konserviert war. Des Weiteren, erstellten wir in dieser Studie das erste Mal eine auf funktionell getestete Varianten basierte Übersicht. Diese Übersicht stellt dar, dass genetische Varianten in Abhängigkeit ihres funktionellen Effektes in den verschiedenen Proteinregionen unterschiedlich verteilt sind. Unsere Ergebnisse können potentiell als Leitfaden für die Bestimmung der Funktion von funktionell ungetesteten Varianten dienen. Abschließend, haben wir unsere Ergebnisse durch in ein öffentlich zugängliches Tool (http://SCN-viewer.broadinstitute.org) einer breiten Gruppe von Nutzern zugänglich gemacht.

In der dritten Studie (Bruenger et al., Brain, 2022) identifizierten wir systematisch biologische Proteineigenschaften die über allen wichtigen Spannungsabhängigen und Ligandengesteuerten Ionenkanalfamilien hinweg, mit der Pathogenität von genetischen Varianten in Verbindung stehen. Unser Ergebnisse zeigten, dass bezogen auf die Struktur eines Ionenkanals, pathogenen Varianten in der Nähe zur Pore Achse, sowie in Regionen welche durch bestimmten Eigenschaften der Pore gekennzeichnet sind, im Vergleich zu Populationsvarianten über alle Ionenkanäle hinweg signifikant angereichert sind. Mithilfe einer neuen Protein Strukturbasierten Methode konnten wir quantitativ nachweisen, dass Varianten

an der Pore die stärkste Anreicherung mit pathogenen Varianten aufweisen. Außerdem zeigten unsere Ergebnisse, dass Proteinregion in denen die Pore hydrophobe Eigenschaften aufweist am stärksten mit pathogenen Varianten angereichert war. Schließlich konnten wir zeigen, dass die identifizierten Eigenschaften sowohl mit der Funktion von 679 funktionell getesteten Varianten als auch den klinischen Phänotypen von 1.422 Patienten, welche neurologischen Entwicklungsstörungen aufweisen, korrelierten. Zusammenfassend ist festzuhalten, dass wir biologische Eigenschaften identifiziert haben, die mit einer Fehlfunktion von Ionenkanälen einhergehen, und zeigen, dass diese mit dem Effekt funktionell getesteter Varianten sowie mit klinischen Phänotypen bei Patienten mit neurologischen Entwicklungsstörungen korreliert sind. Unsere Ergebnisse deuten darauf hin, dass Algorithmen die, zur Unterstützung klinischer Entscheidungen die Pathogenität und Funktion von Varianten vorhersagen können, in Zukunft realisierbar sein werden.

In der vierten Studie (Iqbal und Bruenger et al., Brain, 2022) haben wir einen neuartigen Ansatz entwickelt, der evolutionäre und populationsbasierte genomische Bewertungskriterien kombiniert, um essenzielle 3D Regionen (Essential3D-sites) in Proteinstrukturen zu identifizieren. In unserer Studie identifizierten wir 14.377 Essential3D-sites innerhalb von Proteinstrukturen, die insgesamt von 189 unterschiedlichen Genen kodiert werden, welche von Genen kodiert werden die aufgrund von Mutationen zu neurologischen Entwicklungsstörungen (NDDs) führen können. Viele dieser Gene wurden jedoch erst kürzlich mit NDDs in Verbindung gebracht und sind daher noch nicht gut untersucht. Wir stellten fest, dass diese Stellen in einer unabhängigen Kohorte von über 360.000 Patienten- und Varianten aus der generellen Bevölkerung achtmal stärker mit pathogenen Varianten angereichert waren als in der Kontrollgruppe. Die Essential3D-sites ermöglichen Einblicke in wichtige molekulare Mechanismen der Proteinfunktion, wie z.B. wichtige Protein-Protein-Interaktionen. Die in dieser Studie generierten Annotationen sind unter https://es-ndd.broadinstitute.org abrufbar um eine einfache Einbindung in die klinische Varianteninterpretation zu ermöglichen.

Zusammenfassend lässt sich sagen, dass im Rahmen der Studien meiner Doktorarbeit, genetische, klinische und funktionelle Datensätze zusammengeführt wurden auf Basis dessen wir bioinformatische Ansätze entwickelt haben, um die Bewertung der Pathogenität von Varianten zu verbessern und das Verständnis der Auswirkungen genetischer Varianten auf die Proteinfunktion zu erhöhen. Die während meiner Doktorarbeit erzielten Ergebnisse zeigen auf, dass die Integration mehrerer Datenquellen wichtig ist, um die Auswirkungen von genetischen Varianten in seltenen monogenen Krankheiten zu untersuchen. Unsere Ansätze verbessern insbesondere das Verständnis der Pathogenität und Funktion von genetischen Varianten die in Genen identifiziert wurden, welche verschiedenste schweren Krankheiten hervorrufen können und bei denen die derzeit angewandten Therapien die Krankheitslast nicht

angemessen verringern können. Unsere Ergebnisse tragen somit zu einer neuen Ära der gezielten Krankheitstherapie bei, in der personalisierte Behandlungen und eine verbesserte klinische Versorgung für Patienten zunehmend zugänglich werden. Die hier entwickelten Annotationen und Methoden können als Grundlage für weitere Studien dienen, einschließlich der Anwendung von Methoden des maschinellen Lernens, um die Pathogenität von Varianten und die funktionellen Auswirkungen von Proteinen genauer vorherzusagen.

3. Introduction

3.1 Clinical genetic testing and variant interpretation in rare monogenic diseases: current state, and its limitations

Recent advances in sequencing technology have revolutionized the diagnosis of rare monogenic diseases (Fernandez-Marmiesse, Gouveia and Couce, 2018). Next-generation sequencing (NGS) techniques, such as targeted gene panels, whole exome sequencing, and whole genome sequencing, are now the gold standard for clinical genetic testing. Gene panels are designed to screen for variants in multiple potentially disease-causing genes. At the same time, whole exome sequencing captures all protein-coding regions, which already harbor approximately 85% of disease-related variants (Majewski *et al.*, 2011). Notably, these protein-coding regions reflect only ~2% of the human genome (Rabbani *et al.*, 2012). Whereas all methods are well suited to detect single nucleotide variants (SNVs) and short insertions/deletions, whole genome sequencing is the most comprehensive screening method, covering protein-coding and noncoding regions (van El *et al.*, 2013). While whole genome sequencing is best suited to detect all types of variants including structural variants, such as large deletions/duplications and translocations, it is also the most expensive technique (Royer-Bertrand and Rivolta, 2015; Koboldt, 2020). Overall, the application of these NGS techniques enables the identification of a tremendous number of novel variants, including disease-causing variants for rare monogenic diseases.

Although the number of detected variants increased substantially, the challenge of interpreting the pathogenicity of these variants remains. Accurate variant interpretation is essential for determining whether a genetic variant is pathogenic, causing or contributing to a disease, or it is benign having neutral effects. The assessment of the pathogenicity of a variant plays a critical role in precise disease prognosis and the development of personalized treatments. In contrast, misclassification can have important undesired implications for an individual, including potentially unnecessary stress and ineffective medical procedures (e.g., surgery and frequent screenings) (Plon *et al.*, 2008; Tandy-Connor *et al.*, 2018; Sharo *et al.*, 2022). Compared to the reference genome, each human genome contains about 4-5 million single nucleotides and larger genetic variations, with >100 estimated to be ultra-rare and typically absent in large population databases (Tennessen *et al.*, 2012; Auton *et al.*, 2015). However, most identified variants are not causal for any disease (Richards *et al.*, 2015). This observation holds even among the 64 *de-novo* variants found on average in each individual and not detected in the parents' genomes (Kessler *et al.*, 2020). Furthermore, some non-disease-causing variants are located in disease-associated genes, complicating variant

interpretation. The number of genes associated with diseases is continuously expanding and even accelerated in the last decade, for example for diseases such as epilepsy (Macnee *et al.*, 2023). Moreover, genotype-phenotype associations are often poorly established, and even in better-studied genes, multiple different phenotypes may be associated with variants in the same gene (Johannesen *et al.*, 2022), further complicating variant interpretation.

3.1.1 Guideline-based classification of the genetic variant pathogenicity

The guidelines published by the American College of Medical Genetics and Genomics (ACMG) in 2015 represent the current standard for classifying variants identified in Mendelian disorders (Richards *et al.*, 2015). Previously, various sources could provide valuable information for clinical variant interpretation. This includes information on the frequency of the variant in affected individuals versus controls, functional studies, familial segregation data, genotype-phenotype correlations, computational predictions, and other sources. By integrating these different pieces of information, the ACMG guidelines provide a standardized process for variant interpretation. In particular, they classify variants into five groups: "benign", "likely benign", "uncertain significance", "likely pathogenic", and "pathogenic", using 28 evidence criteria (Figure 1). The evidence criteria are weighted from supporting evidence to very strong evidence. Through a combination of applicable criteria, a variant is then classified as (likely) pathogenic, (likely) benign, or as a variant of uncertain significance (VUS).

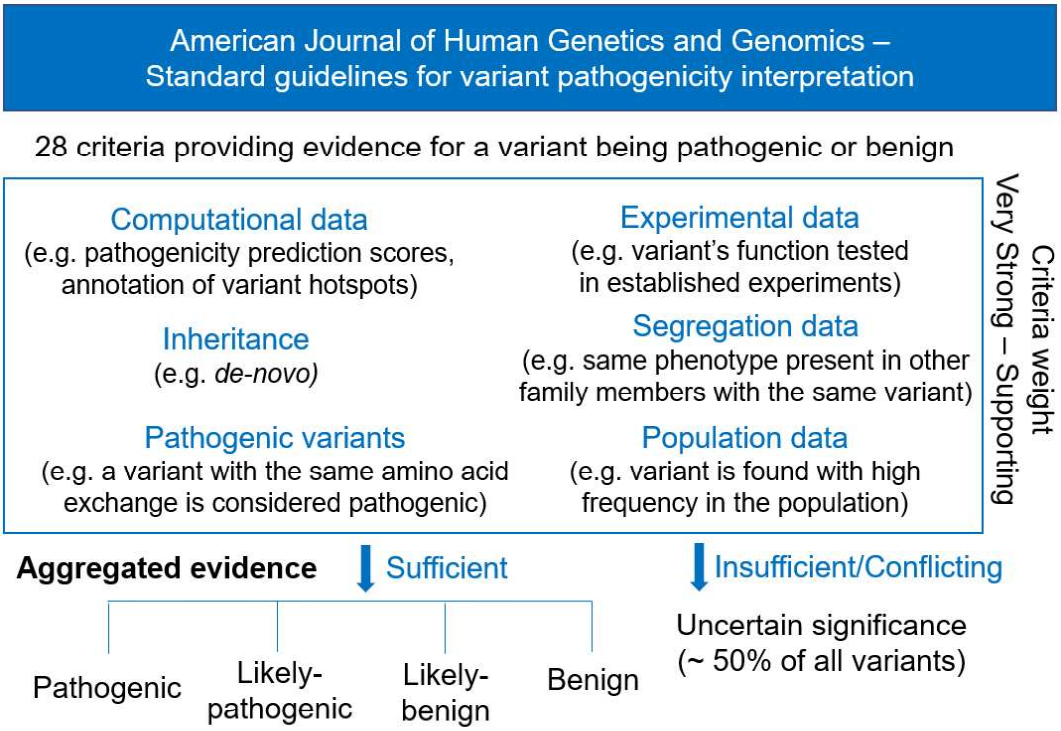


Figure 1 Standard guidelines for variant interpretation proposed by the American College of Genetics and Genomics in 2015. The framework considers evidence from several resources, including experimental and computational data. In total, 28 criteria can be assessed for each variant, which combined result in the classification of the variant. Variants are classified as (likely)-pathogenic, (likely)-benign, or uncertain significance if the evidence is uncertain or conflicting.

In the following, ACMG criteria that are key for the studies in this dissertation are introduced. A nucleotide substitution previously classified as pathogenic and resulting in the same amino acid substitution is considered to provide strong evidence of pathogenicity (PS1 – the first criterion of the ACMG guidelines that provides strong evidence for pathogenicity). Similarly, a nucleotide substitution at the same position as a substitution that was previously classified as pathogenic but leads to a different amino acid change is considered moderate evidence of pathogenicity (PM5). These criteria highlight the importance of the location of pathogenic variants as evidence for classification. In line with this, a variant located in a well-defined functional domain, such as an active site or a hotspot of disease-causing variants in the absence of known benign variation, can be considered moderate evidence for pathogenicity (PM1). Additionally, assuming a variant has been tested with a well-established *in vitro* or *in vivo* experiment, and the results indicate a damaging effect of the mutation on the molecular function of the protein, it can be considered strong evidence of pathogenicity (PS3). Computational evidence (PP3), such as algorithms predicting pathogenic and benign variants, is considered supporting evidence. Overall, in combination, the ACMG criteria provide comprehensive evidence for evaluating the pathogenicity of genetic variants (Richards *et al.*, 2015).

3.1.2 Clinical classification of variant pathogenicity for missense variants remains a major challenge.

Although these guidelines have been applied for eight years, many variants are still interpreted as VUS. For example, approximately 50% of all variants submitted to the ClinVar database (accessed December 2022) a public resource utilized by clinical laboratories, researchers, expert panels, and others for sharing interpretations of variants and associated evidence are classified as VUS (Landrum *et al.*, 2018). The classification of VUS varies among different types of genetic variants, such as missense variants, which arise from a single nucleotide substitution leading to a single amino acid exchange in the protein, protein-truncating variants, which result in a truncated protein version, or large deletions or duplications. Among them, the highest proportion of variants classified as VUS in ClinVar are missense variants (>80%) (Landrum *et al.*, 2018), demonstrating the challenges associated with their interpretation.

A VUS can result from insufficient evidence or contradictory benign and pathogenic evidence. Although theoretically 28 criteria can be applied to each variant, limited availability of information limits tier application, often leading to classification as VUS. The information needed to apply the criteria, such as clinical information on the phenotype, the inheritance of the variant, knowledge about mutational hotspots within the protein, or experimental readouts for the variant, must be more frequently available to overcome this issue. For example, deep mutational scanning (DMS) experiments present a solution to determine variant effects at scale (Livesey and Marsh, 2019) and resolve VUS (Jia *et al.*, 2020). DMS is a high-throughput experimental technique that provides functional testing for a large set of possible variants in a gene (Fowler and Fields, 2014). However, since DMS screenings are hard to establish, they are not yet available for most genes (Esposito *et al.*, 2019) and are currently limited to a few target genes and phenotypes. Alternatively, extensions or modifying existing criteria, including their weight, have been demonstrated as a promising approach to reclassify VUS (Gelb *et al.*, 2018; Kelly *et al.*, 2018; Mester *et al.*, 2018; Luo *et al.*, 2019; Frazer *et al.*, 2021; Patel *et al.*, 2021; Pejaver *et al.*, 2022). Such extensions can be widely applicable and thus present a promising direction to improve large-scale variant assessment.

3.1.3 For successfully pathogenic classified variants, the next frontier presents the classification of the variant effect on protein function

After a variant is classified as pathogenic, the next challenge for applying precision medicine therapies is understanding its functional consequences. Accurate diagnosis and effective treatment of Mendelian disorders depend on knowledge of how a variant affects protein function (Chiron *et al.*, 2000; Brunklaus *et al.*, 2012). This knowledge can also enable precision medicine therapies that specifically target the functional changes caused by the mutation (Chiron *et al.*, 2000; Jen *et al.*, 2007; Larsen *et al.*, 2015; Wolff *et al.*, 2017). Therefore, it is crucial to investigate the effects of a variant on protein function in order to provide better clinical care for patients.

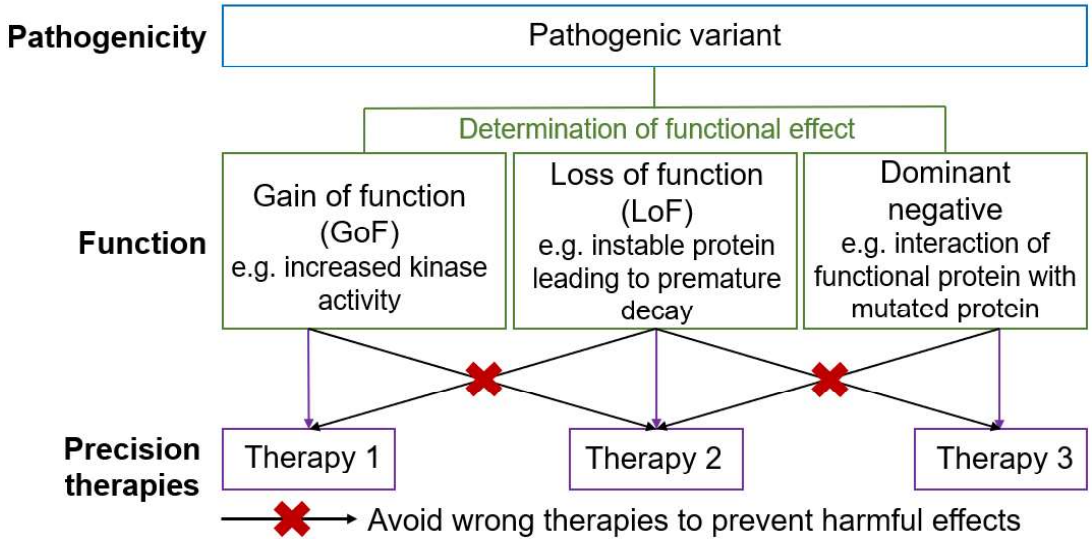


Figure 2 Pathway to enable the application of precision therapies. When a pathogenic variant has been identified, the next step to enable precision medicine is the determination of the functional effect of the variant on the protein function. Misinterpreting a variant's functional effect can lead to an inappropriate therapy with potentially harmful effects for the patient.

Changes in the molecular mechanism of proteins caused by genetic variants are typically grouped into three categories: loss of function (LoF), gain of function (GoF), and dominant-negative variants that disproportionately affect protein function even when only one allele is affected (Wilkie, 1994; Backwell and Marsh, 2022). The mode of inheritance can provide insights into the molecular mechanism underlying a disorder. Recessively inherited diseases are caused by variants in both gene alleles, typically resulting in complete loss of function, although exceptions exist (McCarthy *et al.*, 2021). In contrast, identifying the functional consequence of variants causing autosomal-dominant disorders are often more complex. Dominant disorders can be caused by different mechanisms, including loss of protein function through haploinsufficiency, where a single copy of a gene is insufficient to support normal function; dosage-sensitivity, where different levels of protein abundance can lead to

disease; and dominant-negative effects, where one mutated copy of the gene reduces overall protein function disproportionately by more than 50% (Backwell and Marsh, 2022). Alternatively, an increased protein function is also a frequent cause of dominantly inherited disorders (Backwell and Marsh, 2022).

The type of variant can provide insight into the functional effect. Commonly, variants that result in a truncated protein cause LoF due to nonsense-mediated decay, an essential RNA quality control, and gene regulatory mechanism degrading truncated mRNA products (Nickless, Bailis and You, 2017). In contrast, missense variants can cause both gain and loss of function. Experimental determination of a variant's molecular function is essential but simultaneously time-consuming and expensive. While high-throughput parallel functional analysis approaches have been developed for some genes (Livesey and Marsh, 2019), most variants have unknown functional changes, limiting the application of precision medicine therapies.

3.2 Bioinformatic approaches to overcome challenges in missense variant pathogenicity classification and function prediction.

Bioinformatic approaches have revolutionized the interpretation of genetic variants by providing valuable tools to predict their pathogenicity and gain insights into molecular disease mechanisms. *In silico* methods that derive information by analyzing conservation between multiple sequences and protein structures were among the first features integrated into pathogenicity prediction algorithms (Ramensky, Bork and Sunyaev, 2002; Ng and Henikoff, 2003). Since then, more than a hundred prediction algorithms have been developed, mostly including features based on sequence conservation or structural information (Katsonis *et al.*, 2022). Machine learning techniques, ranging from simple algorithms to advanced deep neural networks, have been utilized to combine features such as conservation scores, predicted protein stability, and the variant’s location in the protein structure (Livesey and Marsh, 2022). Although these algorithms may deliver acceptable performance, their "black box" nature often limits their ability to provide meaningful insights into the molecular pathology of genetic variants (Rudin, 2019; Iqbal *et al.*, 2020). Conversely, identifying individual residues or regions in the proteins associated with disease-causing variants can provide valuable insights into the molecular effect of pathogenic variants (Iqbal *et al.*, 2020; Laskowski *et al.*, 2020).

The following sections will outline the importance of accessing extensive variant databases and offer an overview of two bioinformatics methodologies, along with their foundational principles, utilized in this thesis.

3.2.1 Variants from large patient and population databases can identify functionally important protein regions and inform variant pathogenicity

Within the last decade, several large-scale databases have been generated to provide access to millions of genetic variants observed in both healthy and affected individuals. Among the most extensive databases, there is the Genome Aggregation Database (gnomAD) (Karczewski *et al.*, 2020), the UKBioBank (Sudlow *et al.*, 2015), DiscovEHR (Dewey *et al.*, 2016), TOPMed (Taliun *et al.*, 2021), ClinVar (Landrum *et al.*, 2018) and the Human Gene Mutation database (HGMD) (Stenson *et al.*, 2003). gnomAD provides public access to the genetic variants identified in >75,000 genomes and 125,000 exomes from human sequencing studies presenting the currently most extensive and most widely used collection of genetic variants observed in the population (Karczewski *et al.*, 2020; Gudmundsson *et al.*, 2022). The database contains only individuals without pediatric diseases and thus serves as a powerful reference for studying severe pediatric diseases (Karczewski *et al.*, 2020). In contrast, the ClinVar database (Landrum *et al.*, 2018) and the Human Gene Mutation Database (HGMD) (Stenson *et al.*, 2003) are the most extensive publicly available datasets for variants observed in patients that harbor a presumably inherited disease. Along with the variants in the disease-associated databases, an interpretation of the pathogenicity for each variant is provided. Nevertheless, as more information becomes available or new methods are developed that provide evidence for the variant pathogenicity, the interpretation of a variant can change over time. This mainly includes the reclassification of VUS (Tsai *et al.*, 2019; Glazer *et al.*, 2020), but also includes variants initially considered as likely pathogenic or pathogenic that are changed to likely benign or benign or vice versa (Mersch *et al.*, 2018). Still, a recent study demonstrated that the rate of misinterpretation of variants classified as pathogenic, although they are not the cause of the disease, decreased over time (Sharo *et al.*, 2022). Therefore, developing novel advanced approaches that provide insights into the pathogenicity of variants can contribute to reducing misclassified variants.

The absence of variants from the population can highlight important protein regions and inform variant pathogenicity. Neutral genetic variants, which do not confer any significant advantage or disadvantage to an organism's fitness, can accumulate in the population (Teixeira and Huber, 2021). In contrast, genomic regions where genetic variants

typically show negative effects on an organism's fitness tend to be depleted for variants in the populations. Population databases enable the identification of genomic regions where fewer variants are observed as expected through large mutational models (Samocha *et al.*, 2014). These regions are referred to as genetically constrained. For example, genes are considered constrained if significantly fewer variants are observed in the general population than expected (Lek *et al.*, 2016). The big number of variants aggregated in population databases, such as gnomAD (Karczewski *et al.*, 2020), facilitates the identification of genetically constraint regions within individual genes. Notably, even within a single gene, not all its parts are equally prone to disease-causing mutations. This led to the development of regional constraint scores for missense variants (Samocha *et al.*, 2017). These constrained regions within genes have been shown to frequently overlap with functionally important protein regions that they encode (Traynelis *et al.*, 2017), indicating the impact constrained annotations can have in understanding variant pathogenicity.

Pathogenic variants tend to cluster in functionally important genomic regions, while population variants in those regions are rare or absent. Thus, comparing the distribution of neutral and pathogenic variants along genomic sequences is a powerful approach for detecting disease-associated genes, exhibiting significant enrichment of pathogenic variants compared to neutral ones (Ripke *et al.*, 2014, p. 108; Epi25 Collaborative and Epi25 Collaborative, 2021). Adapted to individual genes, this approach can also be used to identify mutational hotspots within individual genes (Quinodoz *et al.*, 2022). Overlapping these hotspots of pathogenic variants with protein function annotations, such as those provided by InterPro (Paysan-Lafosse *et al.*, 2023), enables providing insights into the molecular disease mechanisms (Quinodoz *et al.*, 2022). As outlined by the ACMG guidelines, the presence of a missense variant in a mutational hotspot constitutes evidence for the variants' pathogenicity (PM1 criterion) (Richards *et al.*, 2015). Therefore, integrating identified hotspots into the interpretation framework can provide additional evidence to reclassify VUS or support the reclassification of misclassified variants.

3.2.2 Evolutionary conservation and molecular functional redundancy across paralogous genes can inform variant pathogenicity and function

The examination of human genome variation through evolutionary conservation has proven key in identifying regions of critical biological importance and has yielded valuable insights for variant interpretation (ENCODE Project Consortium, 2004; Margulies *et al.*, 2007; Davydov *et al.*, 2010). Typically, comparative sequence analysis of homologous

protein sequences is based on the alignment of multiple sequences. Variants that adversely affect biological functionality are often subject to negative selection and are removed from the population over time (Kimura, 1983). As a result, conserved sequences can serve as indications for critical biological functions encoded in the genome (Margulies *et al.*, 2007). Furthermore, sequence conservation alone or in combination with additional features is integrated into a range of *in-silico* methods to predict variant pathogenicity, as recently reviewed by Katsonis et al., 2022. Indeed, one of the currently most potent approaches to predict variant pathogenicity is solely based on co-evolutionary features (Frazer *et al.*, 2021).

Evolutionary conservation between homologous human genes serves as a proxy for variant pathogenicity. Homologous genes with high sequence similarity are considered to belong to the same gene family (Dufayard *et al.*, 2005; Yates *et al.*, 2021). A gene family contains multiple copies of a gene, known as paralogs, arising from duplication events (Koonin, 2005). They play essential roles in evolution and contribute to various biological processes, such as metabolism, signal transduction, and cellular regulation. About 75-80% of human disease-associated genes can be grouped into gene families (Dickerson and Robertson, 2012; Chen *et al.*, 2013). Protein residues conserved across paralogous genes are enriched for pathogenic variants. Hence, quantifying the conservation across genes that belong to the same gene family is a powerful predictor for variant pathogenicity and can help to identify functionally significant protein regions (Figure 3) (Lal *et al.*, 2020).

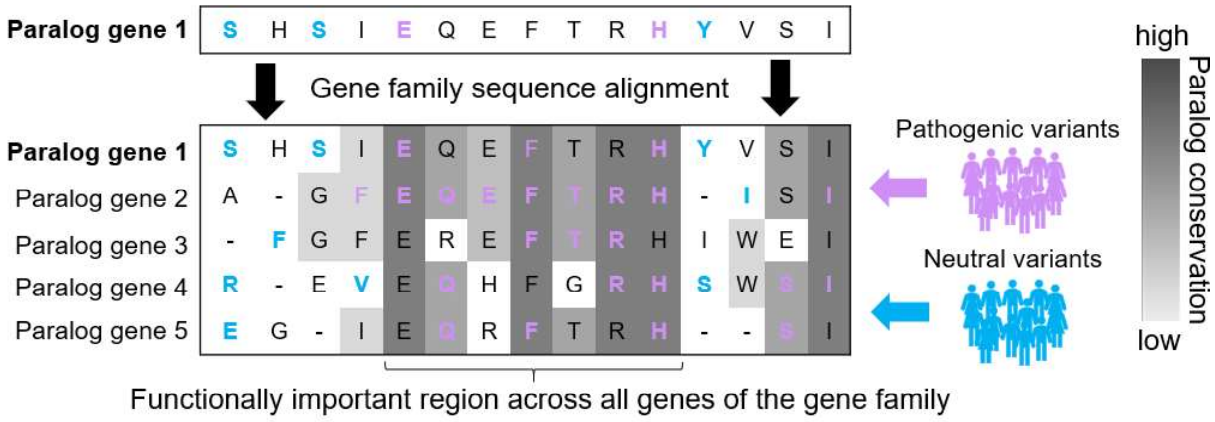


Figure 3 Identification of functionally important regions prone to a disease-causing mutation in the context of gene families. Whereas information from an individual gene may not be sufficient to identify such regions, both conservation across paralogous genes and the localization of pathogenic variants that cluster in similar gene regions provide additional evidence.

Pathogenic variants are similarly distributed across paralogous genes. In line with this, several studies identified common mutational clusters of pathogenic variants across paralogous genes (Ware *et al.*, 2012; Walsh *et al.*, 2014; Pérez-Palma *et al.*, 2020). Identification of mutational hotspots within individual genes can be challenging due to the

scarcity of neutral and pathogenic variants. However, these limitations can be overcome by transferring information, such as the presence of pathogenic variants, across paralogous genes. Aggregating and analyzing variants across multiple paralogous genes can significantly enhance the statistical power to detect common mutational hotspots and inform variant pathogenicity (Pérez-Palma *et al.*, 2020).

Paralogous genes share a similar "core" molecular function (Chen *et al.*, 2013; Lal *et al.*, 2020). Both the conservation across paralogous genes as well as the similar distribution of variants across paralogous genes suggest that the effects one variant may have on the protein function might be conserved across paralogous genes, both in terms of variant pathogenicity and function (Walsh *et al.*, 2014; Lal *et al.*, 2020; Pérez-Palma *et al.*, 2020).

3.2.3 Systematic assessment of genetic variants in a 3D-structural context can identify functionally important protein regions and inform variant pathogenicity

A hallmark of human biology is that the three-dimensional (3D) structure of proteins determines their function (Lesk and Chothia, 1980). Studying protein structures offers a wealth of information for comprehending biological phenomena, facilitating the implementation of techniques such as structure-based drug design or investigation of the functional effects of missense variants on protein function (Śledź and Caflisch, 2018; Khanna *et al.*, 2021). Protein folding, the process by which a linear chain of amino acids linked through a specific covalent bond (peptide bond) transforms into its native, functionally active, 3D conformation, constitutes a quintessential example of biological self-assembly. During folding, the protein first adopts its defining secondary structures, such as alpha-helices and beta-strands, which subsequently coalesce into a compact form known as tertiary structure (or 3D structure or native structure) (Dobson, 2003). For some proteins, an additional final step involves the integration of multiple subunits into a functional quaternary protein complex (Marsh and Teichmann, 2015). Incorporating the underlying structure of the protein is thus vital to understanding the molecular function of proteins.

Analyzing the effect of missense variants in the context of protein structures can reportedly provide insights into molecular disease mechanisms (Stefl *et al.*, 2013; David and Sternberg, 2015; Iqbal *et al.*, 2020)**.** Several studies showed that pathogenic variants tend to concentrate at specific hotspots within protein structures and demonstrated that the localization of these variants allows one to conclude the molecular disease mechanisms

(Kamburov *et al.*, 2015; Araya *et al.*, 2016; Tokheim *et al.*, 2016; Sivley *et al.*, 2018). For example, clusters of pathogenic variants at specific protein-protein interaction sites can highlight the functional importance of these particular interactions (David and Sternberg, 2015; del-Toro *et al.*, 2019). The accumulation of pathogenic variants at specific structural features, secondary structure elements, or functional units, such as ligand-binding sites, have been demonstrated to inform both a variant's pathogenicity and potential functional effects (Heyne *et al.*, 2020; Iqbal *et al.*, 2020). However, although many efforts have been made to identify disease-associated structural features (Gao, Zhou and Skolnick, 2015; Petukh, Kucukkal and Alexov, 2015; Laskowski *et al.*, 2020), the importance of the 3D characteristics associated with pathogenic variants vary among different proteins and protein classes (Iqbal *et al.*, 2020). Consequently, critical structural features must be identified individually for single proteins or protein classes.

Experimentally determining protein structures is complex and limits structure-guided variant interpretation, however, advances in protein structure prediction overcome these limitations. To date, over 50,000 human protein structures have been experimentally determined and deposited in the Protein Data Bank (PDB) (Berman *et al.*, 2000; Tunyasuvunakool *et al.*, 2021). However, only 35% of human proteins have been assigned to a structure in the PDB. Many available structures only cover a fraction of the protein, such as a single protein domain (Berman *et al.*, 2000). Advances in protein structure prediction have helped to address this gap (Won *et al.*, 2019; Pereira *et al.*, 2021; Tunyasuvunakool *et al.*, 2021), particularly with the recent release of the AlphaFold database (Varadi *et al.*, 2022). This database provides public access to pre-predicted protein structure predictions for all human proteins and other organisms. Although the quality of the predicted structures differs among the protein structures, the resource now enables comprehensive testing of structural hypotheses even for proteins without experimentally determined protein structures (Varadi *et al.*, 2022).

3.3 Selection of studied disease cohorts to develop and test novel bioinformatic approaches

The development and testing of bioinformatic approaches to inform the effects of genetic variants in rare Mendelian diseases are challenging as affected individuals are rare. Mendelian disorders are a group of genetic diseases caused by pathogenic variants in a single gene. Although, the population-wide prevalence for rare diseases is estimated as 3.5–5.9% (Nguengang Wakap *et al.*, 2020) the number of cases with a pathogenic variant

associated with a specific rare Mendelian disorder is typically only in the range of a few cases up to a few hundred cases (Orphanet, 2022). Given the small number of affected individuals and functionally tested variants, aggregating available individuals and functionally tested variants from databases, the literature, and collaborative networks is key to test hypotheses and quantify the value of novel approaches that intend to inform the pathogenicity or functional effects of variants. During my Ph.D. and as a group member of Dr. Lal's research team, I collaborated with several research groups located outside of Germany including the USA, Denmark, Australia, and Luxembourg to aggregate unique datasets and test my research hypothesis (see Goals and Objectives).

Whereas the first study in this thesis presents a novel bioinformatic approach to improve variant pathogenicity assessment applied on an exome-wide scale (see Goals and Objectives), the remaining three studies present approaches to inform variant pathogenicity and function for specific subsets of genes. All gene-specific studies involve genes where pathogenic variants are implicated in devastating diseases including lifelong cognitive and physical impairment and early mortality. As symptomatic treatment of these disorders is insufficient, precision medicine applications are needed to significantly improve the conditions of severely affected patients (Zimmern, Minassian, and Korff 2022). As these therapies require knowledge of the causal pathogenic variants and their molecular functional consequences there is a great need to understand the pathogenicity and function of genetic variants in these genes to enable precision medicine applications.

Through intensive collaboration, I aggregated distinct cohorts of individuals and functionally tested variants for each study. In the following sections, I will briefly introduce the different studied diseases and motivate why informing the pathogenicity and/or molecular function in these genes is essential for clinical care and the development of targeted therapies.

3.3.1 Challenges in the interpretation of genetic variants in ion channels

Channelopathies are a heterogeneous group of disorders resulting from ion channel malfunction, with genetic variants as the most frequent cause (Kim, 2014). Pathogenic variants in ion channels are associated with remarkable phenotypic heterogeneity. Indeed, several ion channel-encoding genes are among the most frequently associated genes for severe monogenic neurodevelopmental disorders (Heyne *et al.*, 2018). Clinical heterogenicity correlates with ion channel-encoding gene expression (Brunklaus *et al.*, 2020). However, even among individuals with variants in the same gene, there is a considerable degree of phenotypic heterogeneity, likely due to different functional effects on the protein function.

Missense variants can alter the protein's functional properties in ion channels. Electrophysiological experiments, such as patch clamp recordings, are commonly used to evaluate these functional changes (Jurkat-Rott and Lehmann-Horn, 2004). The assessed properties provide insights into changes such as increased or decreased ion permeation, dysregulation of gating elements, altered opening and closing probabilities, modified kinetics, and protein trafficking (Catterall, 2012; Addis *et al.*, 2017; Chuang and Reddy, 2018; Allen *et al.*, 2019; Rivaud, Delmar and Remme, 2020; Hansen *et al.*, 2021). Recent studies on sodium and potassium channels have highlighted the complexity of these changes, with many missense variants altering multiple properties, often with opposing effects (Thompson *et al.*, 2023).

The treatment of channelopathies depends on the impact of the variant on the protein function. Targeted therapies for ion channel dysfunction are currently being developed and explored in pre-clinical and clinical trials. These therapies are typically designed to target specific types of channel dysfunctions, such as increased or decreased channel function (Striano and Minassian, 2020; Chen and Geschwind, 2022). Therefore, a thorough understanding of the pathogenicity and functional effect of the variant is crucial for patients who participate in these trials.

Ion channels have a common "core" structure. They are membrane-spanning proteins that form ion-permeable pores, allowing the passive flow of ions across the membrane to regulate cell homeostasis and signaling in the nervous, cardiovascular, respiratory, and endocrine systems. Ion channels undergo conformational changes between an open state that allows ion flow and a closed state that blocks it (Hille, 1989). There are two main types of ion channels: voltage-gated channels and ligand-gated channels. Voltage-gated channels exhibit an increase in open probability with membrane depolarization (Hodgkin and Huxley, 1952), while ligand-gated channels open when bound to an agonist such as gamma-Aminobutyric acid (GABA), glutamate, or glycine (Hille, 1989). Despite having a common pore structure, ion channels exhibit selective ion permeability through a structural selectivity filter that distinguishes between ion molecules. Identifying key elements of the ion channels also shared across distant evolutionary channels would highlight important functional characteristics that may provide insights into both variant pathogenicity and their impact on the protein function.

3.3.2 Challenges in the interpretation of genetic variants in voltage-gated sodium channels

Genetic variants in voltage-gated sodium channels are the most frequent cause of neurodevelopmental disorders. More than 15,000 variants are listed in the disease-associated clinical database ClinVar (Landrum *et al.*, 2018). Nine out of the ten genes in the voltage-gated sodium channel gene family are associated with diseases, many of which are devastating. Even within a gene, patients with pathogenic variants are associated with different disorders (Brunklaus *et al.*, 2020). For example, pathogenic variants in one of the sodium channel genes, *SCN1A,* are associated with Dravet syndrome, a severe form of epileptic encephalopathy, as well as with a milder form of epilepsy, genetic epilepsy with febrile seizures plus, or the non-epileptic phenotype, familial hemiplegic migraine (Claes *et al.*, 2001, 2009; Mantegazza and Cestèle, 2018). Although some genotype-phenotype correlations have been established (Brunklaus *et al.*, 2020), the full spectrum of phenotypic diversity is not yet understood.

Both loss and gain of function have been established as disease mechanisms in sodium channels (Brunklaus *et al.*, 2020)**.** Only a small fraction (<1%) of possible missense variants in sodium channels have been functionally tested. While many patients with *SCN2/3/8A*-related epilepsy remain treatment-resistant, those caused by gain of function variants tend to respond to sodium channel blockers, whereas those with LoF variants do not. Strikingly, in patients with Dravet syndrome, which is caused by LoF variants in *SCN1A*, sodium channel blockers even led to a worsening of symptoms (Brunklaus *et al.*, 2012). As with other channelopathies, the treatment responsiveness of sodium channelopathies varies depending on the underlying functional effect (Zuberi *et al.*, 2011; Wolff *et al.*, 2017; Milman *et al.*, 2021; Johannesen *et al.*, 2022). Consequently, approaches that inform the functional effect of variants are key to improve clinical care and enable the application of targeted therapies.

3.3.3 Challenges in the interpretation of genetic variants associated with neurodevelopmental disorders

NDDs are a cluster of congenital or early-onset conditions affecting approximately 2-5% of children globally (Emerson, 2012; Parenti *et al.*, 2020). NDDs encompass a wide range of impairments, including autism spectrum disorder, intellectual disability, motor disorders, and epilepsy (Thapar, Cooper and Rutter, 2017). Studies indicate that around 42-48% of patients

suffering from severe developmental disorders have a pathogenic de novo mutation in a protein-coding gene (Kaplanis *et al.*, 2020). While more than 200 genes have been associated with NDDs, most have been identified only recently. Furthermore, modeling estimates suggest that more than 1,000 genes might be linked to NDDs but require further validation (Kaplanis *et al.*, 2020). Consequently, a majority of genes are not well studied, and the majority of missense variants in these are classified as VUS. Insights into the pathogenicity of genetic variants in these genes and their underlying disease mechanisms could significantly impact disease prognosis and treatment.

4. Goals and objectives

The rapid decline in sequencing costs has led to the identification of genetic variants on an exponential scale (Fernandez-Marmiesse, Gouveia and Couce, 2018). This progress marks an essential step towards the shift from first-generation symptomatic treatments to novel disease-modifying precision therapies that target specific genetic defects. Still, several roadblocks remain. The identified variants' pathogenicity and functional consequences must be assessed to safely enroll patients in precision medicine clinical trials or administer established precision therapies. General guidelines exist for determining variant pathogenicity (Richards *et al.*, 2015), yet most identified variants lack sufficient evidence for precise classification (Landrum *et al.*, 2018). To overcome this challenge, expanding and modifying existing methods and developing specific criteria for a subset of genes have been demonstrated as promising approaches (Gelb *et al.*, 2018; Pejaver *et al.*, 2022). However, gene-set-specific criteria are not established for most of the disease-causing genes. Furthermore, no general guidelines are established to assess a genetic variant's impact on protein function. Experimental determination of the effects is a common approach but is currently limited to a small number of genetic variants. Overall, limitations in determining the variant pathogenicity and function hinder the development and application of precision medicine therapies which remain particularly important for patients with severe diseases.

This dissertation aims to elucidate the pathogenicity of genetic variants and their underlying molecular disease mechanisms to determine the effect on protein function. To accomplish this, we developed bioinformatic approaches grounded in two complementary concepts, independently demonstrating their efficacy in informing variant pathogenicity and function: (1) knowledge transfer across conserved paralogous genes and (2) investigating variant effects within the context of protein structures.

Goal 1: Establish individual paralogous pathogenic variants as a proxy for variant pathogenicity to expand the applicability of two existing ACMG criteria.

Earlier research revealed that conservation among paralogous genes could indicate pathogenicity and that pathogenic variants tend to cluster in shared regions across these genes (Ware *et al.*, 2012; Walsh *et al.*, 2014; Lal *et al.*, 2020; Pérez-Palma *et al.*, 2020). However, the extent to which individual paralogous pathogenic variants can serve as a basis for predicting variant pathogenicity remains unexplored. The ACMG guidelines for variant pathogenicity interpretation consider established pathogenic variants at the same position as evidence of pathogenicity (Richards *et al.*, 2015). Nonetheless, this criterion is rarely

applicable, as most rare disease-associated variants have not been previously observed in other patients. In the first study of this thesis, we propose that the existing criteria could be expanded by including paralogous pathogenic variants that are located at corresponding protein positions, thereby enhancing the applicability of these guidelines and potentially improving the classification of novel variants.

Goal 2: Demonstrate conservation of the functional effects across paralogous genes in voltage-gated sodium channels (VGSCs) to inform the variant function of untested variants.

In this thesis's second study, we hypothesize that the knowledge of the functional consequences a variant has can be transferred to conserved variants in paralogous genes to ascertain their functional effects. To test our hypothesis, we required a set of paralogous genes where a few variants have been functionally tested. However, such variants are scarce. Thus, we collaborated with internationally renowned groups to obtain a substantial collection of functionally tested variants across various genes encoding VGSCs. These channels are ideal subjects for testing our hypothesis for two reasons. First, pathogenic variants in these genes can cause a range of severe heterogeneous disorders (Brunklaus *et al.*, 2020) and together are the gene family that is the most common cause of severe neurodevelopmental disorders (NDDs) (Heyne *et al.*, 2018), where existing medications frequently prove ineffective or only offer limited improvements (Brunklaus *et al.*, 2012; Wolff *et al.*, 2017). Second, pathogenic genetic variants in VGSC-encoding genes can lead to either an increase or a decrease in protein function (Ragsdale, 2008). Many treatment options are under development to address these diseases, but each necessitates accurate knowledge of the functional effects of the variants. Although several variants have been functionally tested, the overall proportion of all possible missense variants in these genes with an experimental readout is <1%. Therefore, enriching our understanding of the functional consequences on protein function for untested variants through knowledge transfer across paralogous genes could significantly impact the development and implementation of targeted treatments.

Goal 3: Identify functionally important regions conserved across ion channels to inform variant pathogenicity and variant function.

The transfer, for instance of known functional effects of variants across paralogous genes, requires that a variant has been functionally investigated in at least a single gene. However, even without information for a specific variant, studying the context of a residue or a genetic variant can offer insights into both variant pathogenicity and underlying disease mechanisms. Identifying structural features enriched for pathogenic variants and simultaneously depleted for

neutral variants is a powerful approach to gain insights into the pathogenicity and the underlying disease mechanisms of genetic variants (Araya *et al.*, 2016; Tokheim *et al.*, 2016). In the third study of this thesis, we aimed to identify critical structural features across all ion channels. Whereas critical structural features have been identified for individual ion channels, such as ligand binding sites or the selectivity filter (Wells *et al.*, 2018; Encinas *et al.*, 2020), a comprehensive assessment of the most important structural features across all ion channels is not available. We hypothesize that identifying such properties systematically will not only enable insights into the function of ion channels but also potentially improve the pathogenicity assessment of genetic variants and enable the establishment of correlations between the identified properties and protein function.

Goal 4: Identify essential protein regions in a 3D-structural context across NDD-associated genes to inform variant pathogenicity and provides insights into molecular disease mechanisms.

Finally, in this thesis's fourth study, we proposed an approach to identify functionally important protein sites across NDD-associated genes. Whereas functionally relevant protein sites have previously been identified through clustering of pathogenic variants (Tokheim *et al.*, 2016), evolutionary conservation (Lal *et al.*, 2020), or the identification of missense constraint sites (Traynelis *et al.*, 2017) we hypothesized that considering a consensus of multiple levels of evidence, combining pathogenic variant clustering together with conservation across paralogous genes and missense variant constraints in a structural context of proteins enables the identification of essential sites in each protein. We tested this hypothesis across NDD-associated diseases. Within NDD genes, clinical genetic testing can often identify a pathogenic variant. Moreover, many genes were only recently associated with NDDs (Kaplanis *et al.*, 2020) and are thus often poorly studied. Studying essential sites across NDD-associated genes could thus enhance the understanding of molecular disease mechanisms, even in recently disease-associated genes, to potentially improve in silico drug target development by providing insights into the molecular disease mechanisms. Finally, knowledge of essential protein sites could improve clinical variant interpretation as they provide evidence for variant pathogenicity following the ACMG guidelines (PM1 criterium) (Richards *et al.*, 2015) since missense variants within essential protein sites are likely disease-causing.

In addition to the four projects incorporated in this thesis, I applied my bioinformatic skills and the developed methods to more than ten additional projects that are already published or submitted (see List of publications).

5. Preface

I conceptualized the studies, wrote and implemented the bioinformatic code, analyzed, generated, and interpreted the results, and wrote the manuscripts where I am listed as the single first author or first shared first-author. Within the two other shared-first authorships incorporated in this manuscript, I wrote the bioinformatic code to analyze the data and contributed significantly to interpret the results. Furthermore, I significantly contributed to the study design to the writing of the manuscript. Besides my four first-author projects, I also contributed to eleven other studies during my Ph.D. training. In this section, I list only four of my five first authorship manuscripts (for a full list of all first ships, see List of publications). In all projects where I am listed as a coauthor, I contributed significantly to the bioinformatic analysis of the study and manuscript writing. Below is a detailed description of my role in the four projects included in this dissertation.

Extension of PS1/PM5 ACMG criteria through paralogous pathogenic variants – First shared first-author

In the first project, I contributed majorly as a shared-first author. I started with harmonizing the annotation of the >4 million pathogenic and control variants aggregated from ClinVar (Landrum *et al.*, 2018) and HGMD (Stenson *et al.*, 2003), and the gnomAD database (Karczewski *et al.*, 2020). For the phenotype-specific analysis, I collected the genetic variants and clinical data from patients with variants in *SCN1A*, *SCN2A*, *SCN3A*, and *SCN8A*, curated the variants, and harmonized the phenotypes. I further developed the bioinformatic code to 1) test if and to which extent individual pathogenic variants in paralogous genes that are located at analogous protein positions can serve as a proxy for pathogenicity, 2) compare the performance of our novel approach to established methods that are also based on information gathered from paralogous genes and 3) identify phenotype correlations based on their biophysical properties across paralogous genes and integrate this information into the extended version of two ACMG criteria which we propose in this project to improve pathogenicity prediction based on individual paralogous variants. I statistically analyzed, visualized, and interpreted the results and wrote the manuscript. In the corresponding manuscript, we demonstrated that individual pathogenic variants can serve as evidence for the pathogenicity of genetic variants. Integrated into the ACMG guidelines, individual pathogenic variants can extend two existing evidence criteria of the current guidelines for variant pathogenicity interpretation.

Gene variant effects across sodium channelopathies predict function and guide precision therapy – Shared first author

In the second project to which I majorly contributed as a shared-first author, I collected the population variants from the gnomAD database, which served as controls in this study. I developed the bioinformatic code to 1) compare the variant effects across conserved paralogous voltage-gated sodium channel disorders, 2) calculate the enrichments of pathogenic vs population variants as well as loss of function (LoF) variants versus gain of function (GoF) variants in different regions of the protein and 3) compare the concordance of the experimental functional readouts we collected in this study to the predicted function (Heyne *et al.*, 2020). I calculated the presented statistics in this paper, generated the protein structure visualizations and visualizations of the statistics, and supported the development of the associated webtool to disseminate our results (http://SCN-viewer.broadinstitute.org). Finally, I contributed significantly to the writing of the manuscript. In the corresponding manuscript, we demonstrated that functional effects are highly conserved across conserved paralogous genes. Thus, the experimentally determined functional effect of variants in one SCN gene can predict the functional effects of variants across other VGSC genes where experimental data is not available.

Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes – First author

In this third project, I contributed majorly as a single first author. I aggregated and annotated patient and population variants from three databases (ClinVar, HGMD, gnomAD) and collaborated with other researchers to gather well-curated clinical and functional data. I annotated and computed all the protein features and pore properties for each ion channel and wrote bioinformatic code to screen these features for the enrichment of pathogenic versus population variants across all ion channels. Additionally, I developed a structural framework to demonstrate the enrichment of patient versus population variants based on the variant's proximity to the pore and membrane axis. Furthermore, I analyzed clinical and functional data for VGSC and NMDA receptor-encoding genes using the structural framework. I was responsible for conceptualizing the study, interpreting the results, and writing the manuscript. All results were visualized by self-written code. Our study showed that certain pore residue properties and proximity to the pore axis were significantly enriched for pathogenic variants compared to population variants across all ion channels. The proximity to the pore can not only serve as a proxy for variant pathogenicity but was also associated with clinical phenotypes and functional in vitro assays.

Delineation of functionally essential protein regions for 242 neurodevelopmental genes – Shared first author

In this fourth project, I contributed majorly as a shared first author. I screened the literature to identify neurodevelopmental disorder (NDD) associated genes. Additionally, I compiled and annotated all the protein structures used in the study. To generate and validate our novel annotations, I collected patient and control variants from large databases such as UKBioBank, DiscoVEHR, and denovo-DB, and collaborated with experts to gather data from exome-wide sequencing studies of individuals with NDDs. I developed the bioinformatic code to generate the Essential3D sites and conducted the validation and application of the annotations. Furthermore, I visualized all the findings in the publication and played a significant role in interpreting the results and writing the manuscript. Through our consensus approach, we identified functionally essential sites in 242 NDD-associated genes. Our results were validated in multiple datasets proving their value for application to unknown new variants. Finally, our annotations provided essential insights into key molecular disease mechanisms.

6. Publications

6.1 Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria

Tobias Brünger[+], Alina Ivaniuk[+], Eduardo Pérez-Palma, Katrine M Johannesen, Ingo Helbig, Sameer Zuberi, Michael Nothnagel, Rikke S Møller, Patrick May, Andreas Brunklaus, Dennis Lal. (in preparation for submission to Genetics in Medicine)

[+]These authors contributed equally to the work

Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria

Tobias Brünger[1+], Alina Ivaniuk[2+], Eduardo Pérez-Palma[3], Katrine M Johannesen[4,5], Ingo Helbig [6,7,8,9], Sameer Zuberi[10,11], Michael Nothnagel[1], Rikke S Møller[4,5], Patrick May[12], Andreas Brunklaus[10,11], Dennis Lal[1,6,13,14*]

[1]Cologne Center for Genomics (CCG), University of Cologne, Cologne, 50931, Germany

[2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44106, USA.

[3]Universidad del Desarrollo, Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana. Santiago, Chile.

[4]Danish Epilepsy Centre, Filadelfia, Dianalund, Denmark.

[5]Department for Regional Health Research, University of Southern Denmark, Odense, Denmark.

[6]Division of Neurology, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

[7]The Epilepsy NeuroGenetics Initiative (ENGIN), Children's Hospital of Philadelphia, Philadelphia, PA, USA.

[8]Department of Biomedical and Health Informatics (DBHi), Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA.

[9]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA.

[10]The Paediatric Neurosciences Research Group, Royal Hospital for Children, Glasgow, UK.

[11]Institute of Health and Wellbeing, University of Glasgow, UK.

[12]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette,

Luxembourg

[13]Stanley Center of Psychiatric Research, Broad Institute of Harvard and MIT, 75 Ames Street, Cambridge, 02142, MA, USA

[14]Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44106, USA.

[+]Authors contributed equally to this work

*Corresponding author. Email: lald@ccf.org

Abstract

Interpreting missense variants in rare monogenic disorders is challenging due to the large number of identified genetic variants that can be both pathogenic or benign, even in disease-associated genes. Current guidelines for variant interpretation feature criteria that consider the presence of previously established pathogenic variants that are located in the same gene and at the same position as evidence of pathogenicity. These criteria can often not be applied because of the small number of established pathogenic variants. Here, we investigated whether variants in disease-associated genes are more likely to be pathogenic than benign if a pathogenic variant in a paralogous gene in the same family is located at a corresponding position of the protein. We mapped over four million pathogenic and control variants from the publicly available ClinVar and gnomAD databases onto 9,990 paralogous genes, annotated gene family alignments, and identified 588 gene families with more than one disease-associated gene. In contrast to previous studies that showed enrichment of pathogenic versus control variants at protein regions were pathogenic variants clustered across gene families, we demonstrate here that residues, when an individual pathogenic variant with the same amino acid exchange was detected at the corresponding alignment position in a paralogous gene, are 6.3-fold enriched for pathogenic variants compared to control variants (P = 1.2e-241). We estimate that including the information about pathogenic variants in the whole gene family in the variant interpretation can increase the application of the PS1/PM5 criteria more than four fold across the exome. For genes associated with multiple different rare disorders due to differences in the molecular effect of pathogenic variants, our gene family approach could identify disorders with potentially similar disease mechanisms across genes of the same gene family. We hypothesize that leveraging pathogenic variants in paralog genes with a similar disease mechanism and correlated phenotypes could further improve pathogenicity assessment. To prove this hypothesis, we aggregated and curated 886 missense variants across nine voltage-gated sodium channel-encoding genes from 1,346 individuals associated with 22 distinct disorders. Consistently with our gene family-wide results, we observed a 10-fold enrichment of patient variants compared to control variants when considering paralogous pathogenic variants as proxies for variant pathogenicity. Next, by including the phenotype information, we could identify phenotype pairs across paralogous genes associated with a similar variant distribution across the 3D protein structure and similar biophysical properties of the amino acid exchange. Limiting the analysis to only paralogous patient variants with correlated phenotypes resulted in an enrichment of patient variants compared to control variants up to 75-fold (P = 1.1e-13). In summary, our results demonstrate a powerful extension of existing PS1/PM5 ACMG criteria using paralogous pathogenic variants. Integrating phenotype information from clinically well-curated patients can potentially

further improve the extended criteria in gene families with heterogenous phenotypes on individual gene level.

Introduction

Large gene panel, whole exome, and whole genome sequencing have led to the identification, at an exponential rate, of novel variants at an exponential rate[1]. Up to 80% of the variants that are disease-causing are located within protein-coding regions of the gene[2], with missense variants being particularly challenging to interpret due to the variety of different molecular mechanisms through which they can cause the disease. Furthermore, several disease-associated genes are pleiotropic, which further complicates variant interpretation[3–5]. Despite these challenges, variant classification is necessary for diagnosing rare and genetically heterogeneous disorders, such as neurodevelopmental disorders[6,7], and for the development of personalized medicine.

To standardize variant interpretation, the American College of Medical Genetics and Genomics (ACMG) published recommendations for evaluating the pathogenicity of variants[8]. However, >75% of missense variants reported in the ClinVar database[9] (accessed January 2023) are classified as variants of uncertain significance, either due to conflicting interpretations or to the absence of sufficient evidence for or against variant pathogenicity. The recommendations feature criteria requiring the presence of an established pathogenic variant with the same amino acid exchange (PS1) or a different amino acid exchange (PM5) at the same position in the same gene that can provide strong to moderate evidence for pathogenicity[8]. However, since the vast majority of rare monogenic disorders are caused by singletons that have not yet been observed in other patients, the application of these evidence criteria is limited.

About 80% of diseases associated with Mendelian disorders are caused by paralogous genes[10]. These paralogous genes can be grouped into gene-families as defined by the Human gene nomenclature consortium (HGNC)[11] which can be further divided into sub-groups with >80% aligned residues[12]. Genes within a gene family arise from gene duplication events of common ancestral genes and can share >90% sequence identity at functional essential protein domains[13]. We and others could demonstrate that functional domains that are conserved within a gene family are enriched for pathogenic variants and depleted for population variants[14–16]. In molecular studies, it has been shown that the protein function at these protein domains is mechanistically conserved within a gene family and therefore, the exchange of a single residue has similar implications across gene family members[16–18]. Therefore, the current PS1/PM5 criteria could be informed not only by prior reported pathogenic variants at the same position in the gene of interest, but also by observed pathogenic variants at conserved positions in related genes of the same family. Evolutionary conserved protein regions across paralogous genes are enriched for pathogenic variants and depleted for population variants. These regions can be labeled as ‘pathogenic variant enriched regions (PERs)’[14]. We could previously show

that new missense variants within PERs are more likely to be pathogenic compared to non-PER regions in the same gene[12,14,16,18]. Although the PER-annotation can identify disease-sensitive regions across protein sequences, the approach has the disadvantage that it has low sensitivity since the identification of a PER is statistical power intensive.

To the best of our knowledge, it has not been tested across the protein-coding exome whether individual missense variants from paralogous genes can serve as a proxy for variant pathogenicity and thus extend the application of the PS1/PM5 criteria, which currently rely on the presence of pathogenic variants in the same gene at the same position. In this study, we aim to determine across 588 gene families with high sequence similarity, whether the presence of an individual pathogenic variant in a gene family member at a corresponding protein position increases the likelihood that a novel genetic variant at a conserved analogous position in the target gene is pathogenic. Furthermore, we explore for one large gene family that encodes voltage-gated sodium channels (VGSC), whether the extension of PS1/PM5 to pathogenic variants of paralogous genes can be further improved by incorporating information from correlated phenotypes. Using a clinically well-curated cohort of 1,346 patients we identify correlated phenotypes based on the biophysical properties of the associated disease-causing variants across VGSCs. We demonstrate that incorporating the phenotype correlations can improve the selection of variants incorporated in the extended version of the PS1/PM5 ACMG criteria and consequently improve the variant pathogenicity assessment.

Methods

Annotation of missense variants from public repositories

Missense variants from patients

Missense variants associated with the disease were collected from the ClinVar database[9] (ClinVar, release October 2019) and The Human Gene Mutation Database[19] (HGMD®) Professional release 2019.2. The ClinVar missense variants were obtained in a tabular format from the FTP site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/) and only those variants exclusively classified as "Pathogenic" and/or "Likely Pathogenic" were considered to ensure high stringency. The HGMD dataset was filtered for "missense variants," "High Confidence" calls (hgmd_confidence = "HIGH" flag), and "Disease causing" state (hgmd_variantType = "DM" flag). All annotations refer to the human reference genome version GRCh37.p13/hg19, and variants belonging to noncanonical transcripts as defined by Ensembl were excluded[20]. Since ClinVar and HGMD are not mutually exclusive, we used the union of both

resources and removed duplicate entries by comparing HGVS annotations. We further refer to the combined set of variants classified as likely-pathogenic or pathogenic as 'pathogenic variants'.

Similarly, we gather an updated version of the variants from ClinVar (release December 2022) and HGMD (Professional release 2023.1), processed them as described before, and extracted all novel pathogenic variants not observed in the previous dataset to obtain an independent set of variants.

Missense variants from the population

Missense variants present in the Genome Aggregation Database[21] (gnomAD, public release 2.0.2) were obtained in the Variant Call Format[22] (VCFs). We extracted the high-quality missense variants by filtering the VCF files for the "CSQ" field and "PASS" flag. The annotations were based on the human reference genome version GRCh37.p13/hg19. We extracted the entries annotated to a canonical gene transcript. The aggregated population variants serve as control variants in our study and are further referred to as controls.

Similarly, we gather an updated version of the variants from gnomAD (public release 2.1.1, processed them as described before, and extracted all novel variants not observed in the previous set of gnomAD variants to obtain an independent set of control variants.

Annotation of missense variants and associated phenotypes for the voltage-gated sodium channels

Brain-related phenotypes

We aggregated published patient missense variants observed in voltage-gated sodium channel genes (VGSC) genes from the literature. All patient variants for *SCN1A* were obtained from Brunklaus et al.,2022, and Brunklaus et al., 2022[23,24]. Variants for *SCN2A* were obtained from Wolff et al, 2017[25] and Crawford et al., 2021[26]. Variants for *SCN3A* were obtained from Zaman et al., 2018[27]. All *SCN8A* variants are taken from Johannesen et al, 2021[4]. Affected individuals were recruited through a network of collaborating clinicians, as well as GeneMatcher[28], by means of a standardized phenotyping sheet to assess clinical characteristics cognition), EEG, neuroimaging, and retrospective data on antiepileptic treatment.

Non-brain phenotypes

SCN5A variants were obtained from the studies conducted by Milman et al., 2021[3], and Walsh et al., 2021[29]. Data from *SCN4A*, *SCN9A*, *SCN10A*, and *SCN11A* variants were collected from various publications cited in Supplementary Table 3. Variants in the VGSC genes that were not missense-constrained were filtered for the maximum population frequency (MAF). We inferred the MAF

thresholds by using the approach described by Whiffin et al., 2017[30], via the authors' app (https://www.cardiodb.org/allelefrequencyapp), based on the phenotype's estimated prevalence, inheritance mode, and penetrance. We categorized *SCN4A* variants related to myotonia congenita and paramyotonia congenita and *SCN9A* variants related to primary erythromelalgia and paroxysmal episodic pain disorder into single categories (Relaxation Impairment Disorders and Paroxysmal Pain Disorders, respectively) based on their shared molecular pathology and pathophysiology after applying the MAF filter.

We mapped all variants to their Ensemble canonical transcript[24] (*SCN1A*: ENST00000303395, *SCN2A*: ENST00000283256, *SCN3A*: ENST00000283254, *SCN4A*: ENST00000435607, *SCN5A*: ENST00000423572, *SCN8A*: ENST00000283254, *SCN9A*: ENST00000409672, *SCN10A*: ENST00000449082, *SCN11A*: ENST00000302328). Only phenotypes associated with variants at >5 different protein positions were considered. Original and harmonized phenotype annotations for each phenotype are listed in Supplementary Table 3.

Gene family definition

We obtained the gene families from Pérez-Palma et al. 2019[14], as originally described in Lal et al., 2020[12].

Definition of paralogous variants

We conducted a multiple protein sequence alignment using the MUSCLE[31] software to identify variants located at analogous positions across paralogous genes, following the method outlined in Pérez-Palma et al., 2019[14]. Next, we mapped all patient and control variants onto this alignment. We classified two variants on different paralogous genes as paralogous variants if they satisfied two conditions. Firstly, they should be located at the same alignment index position within the multiple sequence alignment. Secondly, the reference amino acid should be conserved between those two paralogous genes (the reference amino acid should be the same).

Statistical analysis

If not otherwise stated, we used a two-by-two contingency table to quantify the burden of pathogenic over control variants across different sets of residues at the same position, using the odds ratio (OR), augmented by the 95% confidence interval (CI), as an estimate for the enrichment. We used a two-sided Fisher's exact test to test for the association while adjusting for multiple testing using Bonferroni correction. Meta-analyses were performed using the metafor package[32] (version 3.8-1) in R. All statistical tests were performed using R v.4.2.1.

Calculation of the enrichment of patient versus control in the presence of paralogous pathogenic variants

To investigate whether individual pathogenic paralogous variants can serve as a proxy for pathogenicity for variants in a target gene, we followed the approach depicted in Supplementary Figure 1. Our calculations are limited to gene families that harbor pathogenic variants in at least two genes. We counted the number of patient and control variants for each protein position in the target gene, for cases where 1) one or multiple paralogous pathogenic variants were observed and 2) when no paralogous pathogenic variants were observed. To calculate the enrichment of pathogenic versus control variants, we performed a Fisher's Exact test comparing the number of pathogenic versus control variants in the target gene in the presence of paralogous pathogenic. Furthermore, we distinguished between paralogous variants with the same or a different amino acid exchange as the variant in the target gene. We then combined the results for each gene family in a meta-analysis to obtain gene family-specific enrichments of pathogenic versus control variants in the presence of a paralogous pathogenic variant (Supplementary Table 1). We limited the meta-analysis to genes where more than ten pathogenic and control variants were observed. To estimate the average enrichment across all gene families included in our study we performed a meta-analysis across all genes where more than ten pathogenic and control variants were observed. Meta-analyses were performed with the metafor package in R.

To compare our results to an established gene-family-based approach on exome wide scale[14] we gather an independent set of variants (see Annotation of missense variants from public repositories) not previously used in their or our approach and repeated the calculation outlined above.

Identification of phenotype pairs associated with similar variant properties

In this study, we characterize the properties of each missense variant by considering 1) the distribution of the variants in the 3D protein structure and 2) the biophysical properties of the amino acid exchange.

1) Identification of phenotypes associated with variants with similar 3D variant distribution

Initially, we evaluated the spatial distribution of sodium channel protein structures for variants associated with each phenotype. We tallied the number of patients reported for each variant in every phenotype. Since not all sodium channels had available protein structures, we mapped the patient variants and their corresponding patient count on the Nav1.2 protein structure (PDB-ID: 6j8e) using a multiple protein sequence alignment performed across all VGSC encoding genes using Muscle[31]. We only considered patient variants that could be mapped to the protein structure for downstream analysis. For every residue in the Nav1.2 protein structure, we counted the number of patients with a variant in the residue or its local 3D neighborhood using a 12-angstrom radius cutoff, as described in Iqbal et al., 2022[33]. The number of patients per residue was evaluated independently for each phenotype. To compare the distribution of patient counts among different phenotypes, we computed Kendall's rank correlation. We measured the pairwise Euclidean distances between each phenotype pair by taking the correlation coefficients as input to quantify the similarity of the variant distribution for each pair of phenotypes. The Euclidean distance presents a measure of similar variant distribution in 3D between two phenotypes.

2) Identification of phenotypes associated with variants with similar biophysical properties of the amino acid exchange

Next, we evaluated the properties of the mutated amino acids and the alterations in the local 3D biophysical environment in comparison to the non-mutated environment for the variants associated with different phenotypes. We examined four properties that quantify changes in the polarity, hydropathy, hydrophobicity, and charge of the local environment, as well as the Grantham[34] score, a measure of the severity of the amino acid substitution. We calculated the average value of each property across all variants associated with one phenotype. Based on these average values that quantify the severity of the amino acid exchange, we computed the Euclidean distance between each pair of phenotypes. The Euclidean distance is a measure of how similar the severity of amino acid exchanges of variants that are associated with different phenotypes. A small distance between two phenotypes indicates that these phenotypes are associated with variants with amino acid exchanges of similar severity.

Combination of amino acid exchange properties and 3D variant distribution

Next, we combine the phenotype similarity that is based on the amino acid exchange properties of the associated variants and their distribution in 3D. First, we normalize the individual Euclidean distances between each pair of phenotypes which were calculated for both approaches between 0 and 1. Second, we calculated the mean of the normalized Euclidean distances. The average Euclidean distance provides a measure of the similarity between the variants associated with two phenotypes. A small distance between a pair of phenotypes indicates that both phenotypes are associated with variants with similar properties. Finally, by ranking these distances we obtain for each phenotype five phenotypes associated with the most similar variant properties and five phenotypes associated with most different variant properties.

Integrating variant similarity among phenotypes to paralogous variant-based pathogenicity assessment

We next explored whether integrating the phenotype information input into our paralogous pathogenic variant approach improves the pathogenicity assessment of a variant. Specifically, for each gene and phenotype, we calculated the enrichment of patient variants compared to control variants in the presence of a paralogous patient variant using the same approach as outlined for the exome-wide approach, but with small modifications. Instead of considering all paralogous patient variants as a proxy, we restricted our analysis to two groups of variants: 1) paralogous patient variants observed in other VGSC genes that are associated with one of the five phenotypes that were associated with the most similar variant properties (for definition see Methods - Identification of phenotypes pairs associated with similar variant properties) compared to the phenotype associated with the variant in the target gene, and 2) paralogous patient variants linked to one of the five phenotypes with the most different variant properties compared to the phenotype associated with the variant in the target gene.

To validate the phenotype-dependent enrichment of patient variants versus control variants in an independent dataset we extracted the variants of the most common phenotypes in each sodium channel with >40 different variants (*SCN1A*: Dravet Syndrome, *SCN2A*: Early onset developmental epileptic encephalopathy (DEE), *SCN5A*: Brugada Syndrome, *SCN8A*: DEE). We divided these cohorts as well as the population variants that served as controls into four sets of variants, each containing 25% of the patient and control variants. We then combined three of the four sets (representing 75% of variants for each phenotype) with our remaining patient cohort containing all variants linked to other phenotypes. Following the approach outlined in the previous section we then identified phenotypes associated with similar variant properties. We then, using the independent test cohort, calculated the enrichment of patient variants versus control variants separately for paralogous patient variants linked

to phenotypes with similar and different variant properties. We repeated this approach three times, such that each set of variants was used three times as part of the training set three times and once as the test set. Finally, we combined the results of the four independent tests in a meta-analysis.

Results

The presence of individual paralogous pathogenic variants can infer variant pathogenicity and extend current ACMG criteria

The aim of our study was to investigate whether the presence of pathogenic variants in paralogous genes at a corresponding position can be used to extend the existing strong PS1 ('Same amino acid change as a previously established pathogenic variant regardless of nucleotide change') and moderate PM5 (Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before') variant pathogenicity classification criteria of the ACMG guidelines (Figure 1). We defined a paralogous variant, as a variant that fulfills two criteria: (1) it is located in a paralogous gene within the same gene family at the corresponding protein position, as determined by multiple sequence alignment (see Methods for details), and (2) it harbors the same reference amino acid as the target gene.

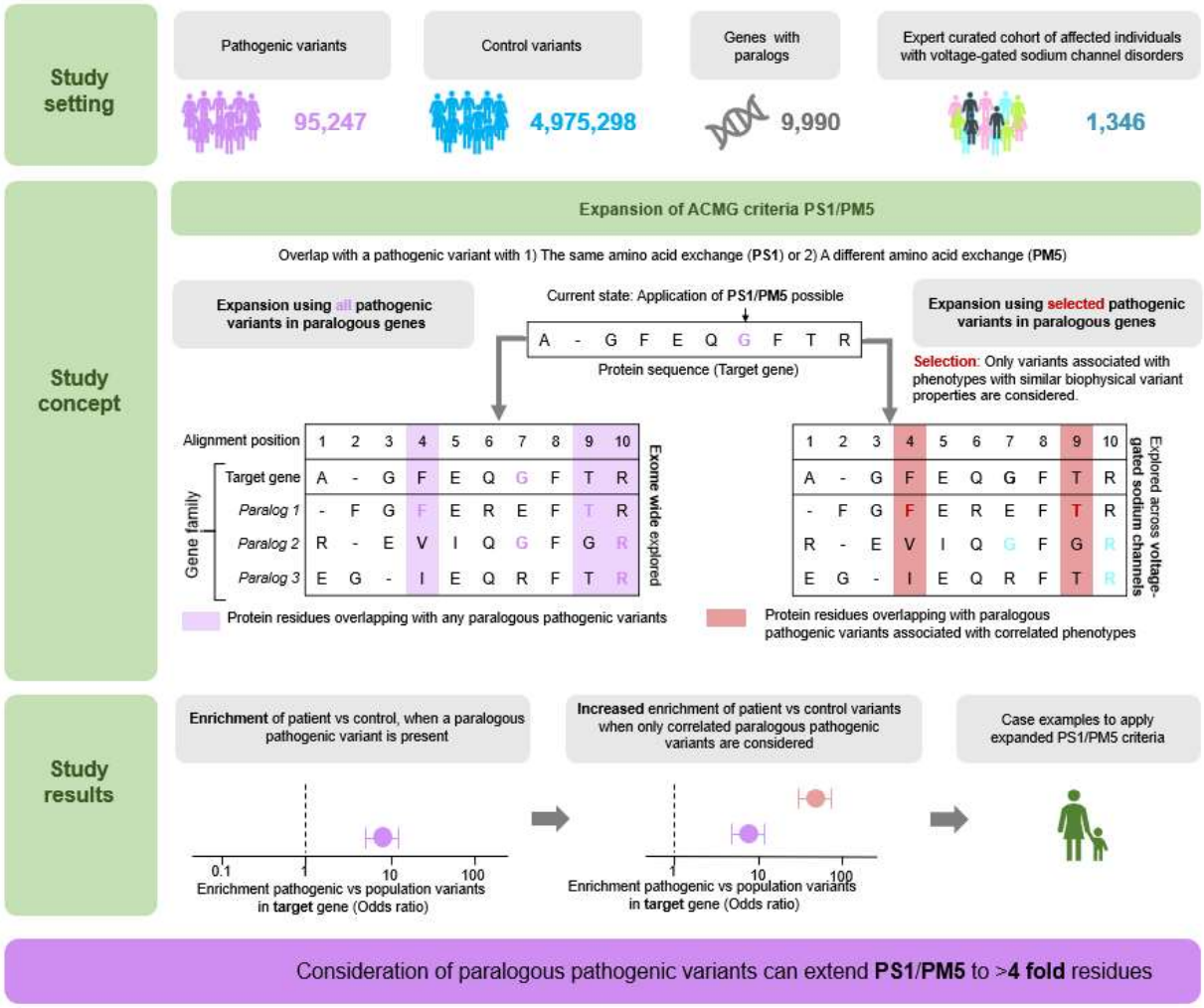


Figure 1: Graphical summary of the study.

First, we explored to which degree a paralog-level annotation of the PS1/PM5 criteria would increase the number of classifiable amino acids. We tested if paralogous pathogenic variants can be used as a proxy for determining the pathogenicity of variants in another gene (target gene) within the same gene family. We aggregated a total of 95,247 pathogenic variants from ClinVar9 and HGMD19 and 4,975,298 control variants from the gnomAD21 database and mapped them to 2,871 different gene families, consisting of 9,990 genes (Figure 1 Study setting). We identified 588 gene families, including 2,920 genes, that harbor pathogenic variants in at least two genes of each family. Within these genes, pathogenic missense variants were located at 34,518 residues, and paralogous pathogenic missense variants at 161,344 residues. Therefore, an extension of the PS1/PM5 criteria that would consider the presence of pathogenic variants in paralogous genes would increase the number of classifiable amino acids 4-fold (Figure 2).

Next, the demonstrate the validity of paralogous pathogenic variants as a proxy for variant pathogenicity in genes of the same gene family. Across all gene families, we observed a 6.3-fold enrichment of pathogenic versus control variants (P = 2.1e-241, Figure 2B) when a paralogous

pathogenic variant with the same amino acid exchange was present at a corresponding position. When we analyzed a reduced set of genes containing only those that are constrained for missense variants (Missense-z score > 3.091), the enrichment of pathogenic versus control variants was slightly increased (Missense constraint genes: OR = 6.8, P = 4.2e-45, All genes: OR = 6.3, P = 2.1e-241, Figure 2C). Notably, even for paralogous variants with a different substitution at the same alignment index position, we observed an enrichment of pathogenic versus control variants (All genes: OR = 3.6, P = 2.4e-198, Missense constraint genes: OR = 4.7, P = 7.5e-34, Figure 2B, C).

We then applied a Bayesian framework of the ACMG guidelines proposed by Tavtigian et al. in 2018, to estimate the weight of evidence for the pathogenicity of variants that overlap with paralogous pathogenic variants. The framework proposes cutoffs to classify evidence weights, with OR > 18.7 considered strong, OR > 4.3 as moderate, and OR > 2.08 as supporting evidence for variant pathogenicity[35]. We observed a 6.3-fold enrichment of pathogenic versus control variants when a pathogenic variant in a paralogous gene that harbors the same amino acid exchange was present in a corresponding protein position. Consequently, this criterium can provide moderate evidence for variant pathogenicity. However, since the enrichment varied widely across different gene families we calculated the evidence weights separately for each family (see Supplementary Table 2). Individually for each gene family, we found that 109 gene families (42 of which were statistically significant after Bonferroni correction, Supplementary Table 1) showed significant enrichment for pathogenic variants if a paralogous pathogenic variant is present at a corresponding protein position with the same amino acid exchange. Following the Bayesian framework this translates to 39 gene families with strong evidence, 54 with moderate evidence, and 14 with supporting evidence for variant pathogenicity when a paralogous pathogenic variant with the same amino acid exchange was observed (Figure 2D). We found fewer gene families where the presence of a paralogous pathogenic variant with a different amino acid exchange provides strong (N = 26 gene families) or moderate evidence (N = 42 gene families) for pathogenicity, but more gene families where paralogous pathogenic variants with a different amino acid exchange can provide supporting evidence (N = 31).

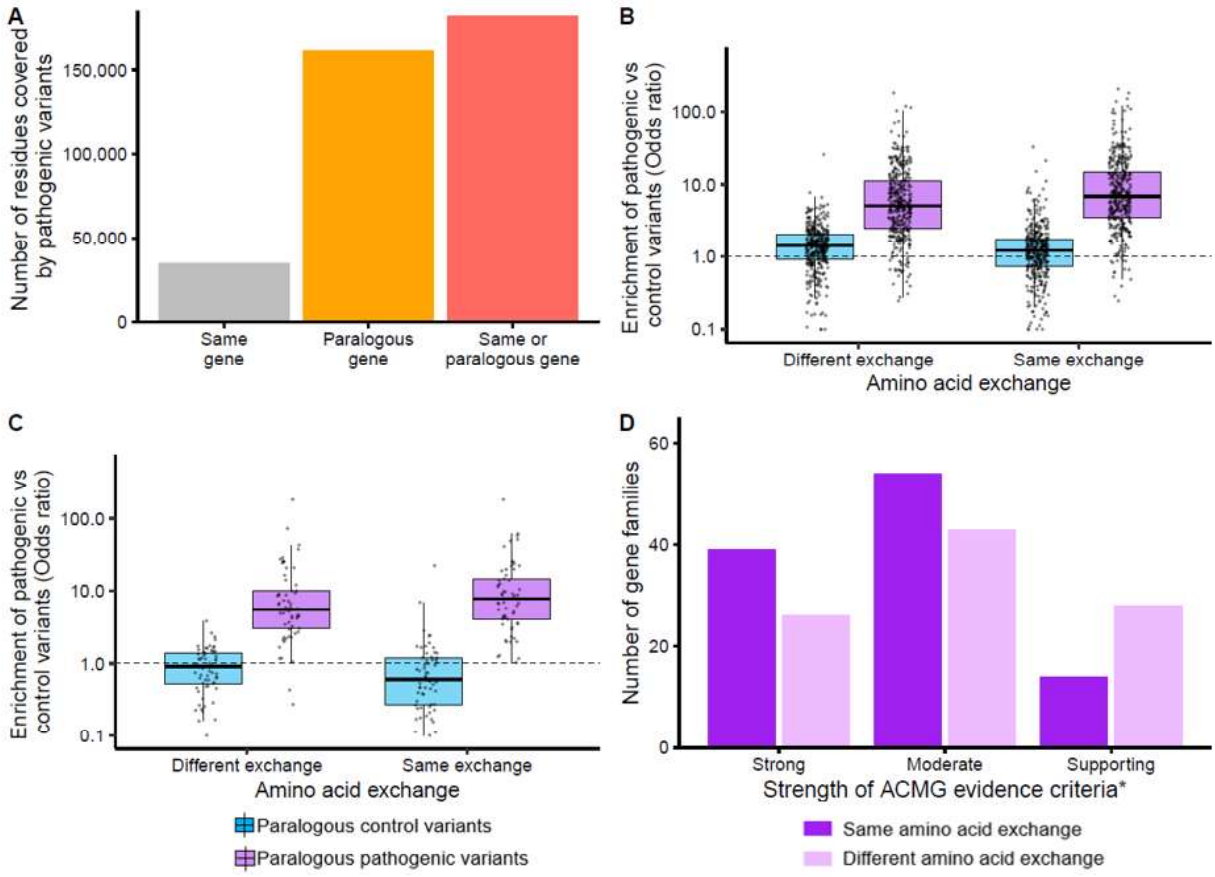


Figure 2: Individual paralogous pathogenic variants can serve as a proxy for variant pathogenicity. A) Number of protein residues with a pathogenic variant in the same gene and/or in a paralogous gene at a corresponding protein position. B) The boxplot displays the enrichment or depletion of pathogenic (ClinVar/HGMD) versus control variants (gnomAD) in the presence of paralogous pathogenic or control variants at corresponding protein positions. Each dot represents the fold enrichment (OR) observed across a gene family. The odds ratio was separately calculated for paralogous control and pathogenic variants with the same or a different amino acid exchange compared to the variant in the target gene. Only gene families with at least 10 patient variants mapped are visualized. C) As in (B), but limited to missense constraint genes (Missense-z score > 3.09). D) Bar plot showing the number of gene families where pathogenic variants with the same (purple) or different amino acid exchange (pink) provide supporting-strong evidence for variant pathogenicity. *Following the Bayesian framework proposed by Tavtigian et al., 2018.

Comparison to other gene family-based approaches

We have previously demonstrated that paralog conservation alone can distinguish pathogenic and control variants to some extent[12]. Thus, we investigated whether the observed enrichment of patient versus control variants in our approach is solely driven by paralog conservation. To do so, we repeated our analysis within groups of residues with similar paralog conservation. We found a 2.7-7.2-fold enrichment for pathogenic versus control variants within each group (Figure 3 A, B), indicating that the presence of pathogenic paralogous variants provides additional information on top of paralog

conservation. Notably, we observed the highest enrichment of patient versus control variants for residues with minor paralog conservation (Figure 3 A, B, Parazscore <0).

Furthermore, we compare our approach with our previously introduced approach to identify pathogenic enriched regions (PERs) across gene families[14]. The previous approach identified clusters of pathogenic variants that were demonstrated to serve as a proxy for variant pathogenicity. In contrast, we consider in our approach individual pathogenic variants, in line with PS1/PM5, observed in paralogous genes as a proxy for variant pathogenicity. When we considered the same set of pathogenic and control variants we found that PERs are found in 41,314 residues, whereas pathogenic paralogous variants at corresponding protein positions are found in more than 4 times more residues (Number of residues: 161,344, Figure 3C), enabling an application of our approach to many additional residues which are not informed by PERs. Additionally, we calculated the enrichment of pathogenic versus control variants in an independent dataset comprised of novel pathogenic variants and control variants (see Methods for details) and found a similar enrichment for pathogenic versus control variants using our approach and PERs (OR_{PER} = 3.6, $OR_{Same\ exchange}$ = 3.3, $OR_{Different\ exchange}$ = 2.9, Figure 3D).

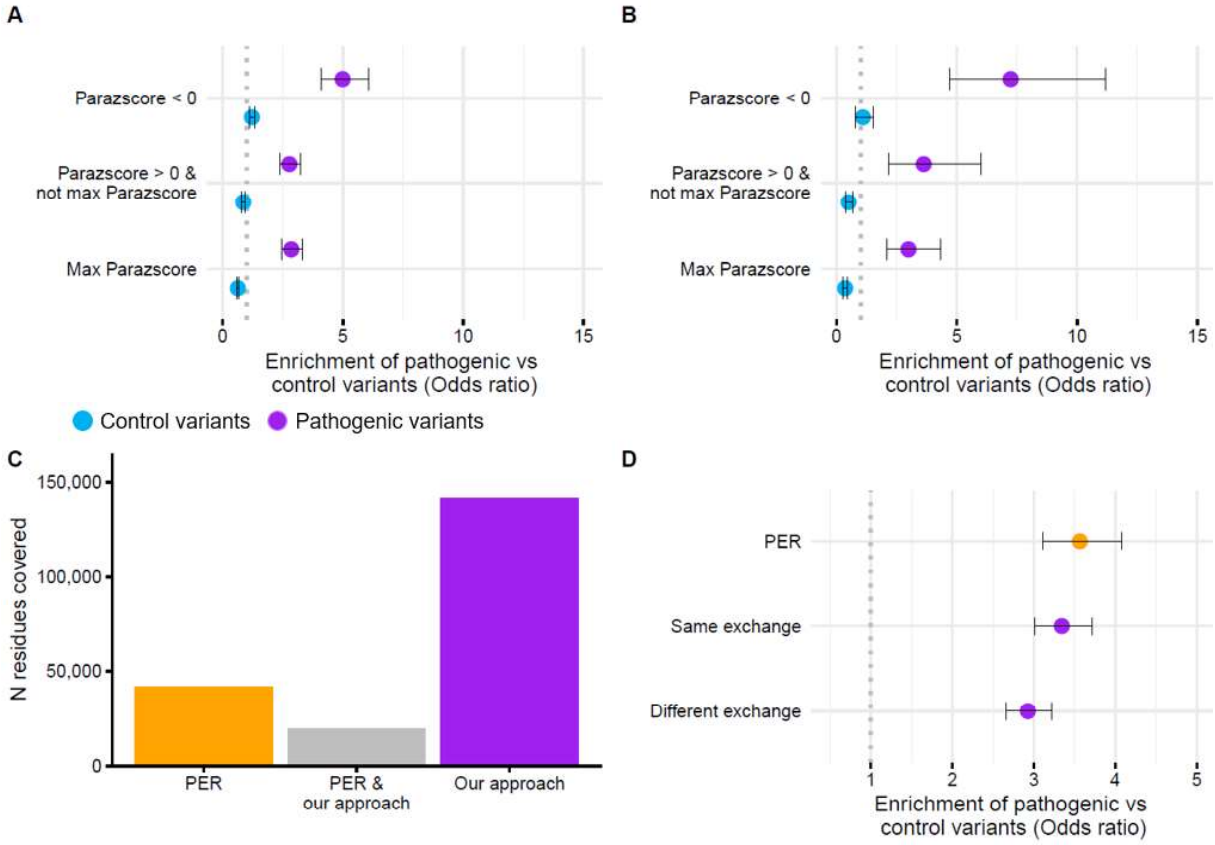


Figure 3: Comparison to established gene family-based methods. A) The forest plot illustrates the enrichment of patient versus control variants for residues with similar paralog conservation levels, as defined in Lal et al., 2020[12]. The meta-analysis summarizes the gene-wise enrichments. B) Similar to (A), but for missense constraint genes only. C) The bar plot shows the number (N) of residues across all genes where the established approach (Pathogenic Enriched Region, PER; Perez-Palma et al., 2019[14]) and our approach can be applied. D) The forest

plot compares the enrichment of patient versus control variants for the PER approach and our approach in an independent dataset (see Methods for details). The odds ratio (OR) and 95% confidence interval (CI) are indicated for each approach.

Identification of phenotype-pairs associated with similar variant properties

For genes associated with multiple different rare disorders due to differences in the molecular effect of pathogenic variants, our gene family approach could identify disorders with potentially similar disease mechanisms across genes of the same gene family. To test this hypothesis, we collected a unique dataset of 1,346 patients, encompassing 22 different phenotypes and harboring 886 distinct missense variants in genes encoding voltage-gated sodium channels (VGSCs) (Supplementary Table 3). First, we validated our previous findings and observed a 4.5-10.8-fold enrichment of patient versus control variants across all VGSC genes when a paralogous patient variant was observed at a corresponding protein position (Figure 4A, $OR_{Same\ exchange}$ = 10.8, P = 2.4e-45; $OR_{Different\ exchange}$ = 4.5, P = 4.7e-19).

Next, we sought to identify which phenotypes across these VGSCs are caused by variants associated with similar variant properties. We annotated for each variant different properties (Figure 4B, see Methods for details) and identified pairs of phenotypes that are caused by variants with similar properties (Figure 4C). For instance, *SCN1A* variants associated with Dravet syndrome had similar variant properties to variants in *SCN2A* associated with autism or Brugada syndrome in *SCN5A* (Figure 4C, Supplementary Table 4).

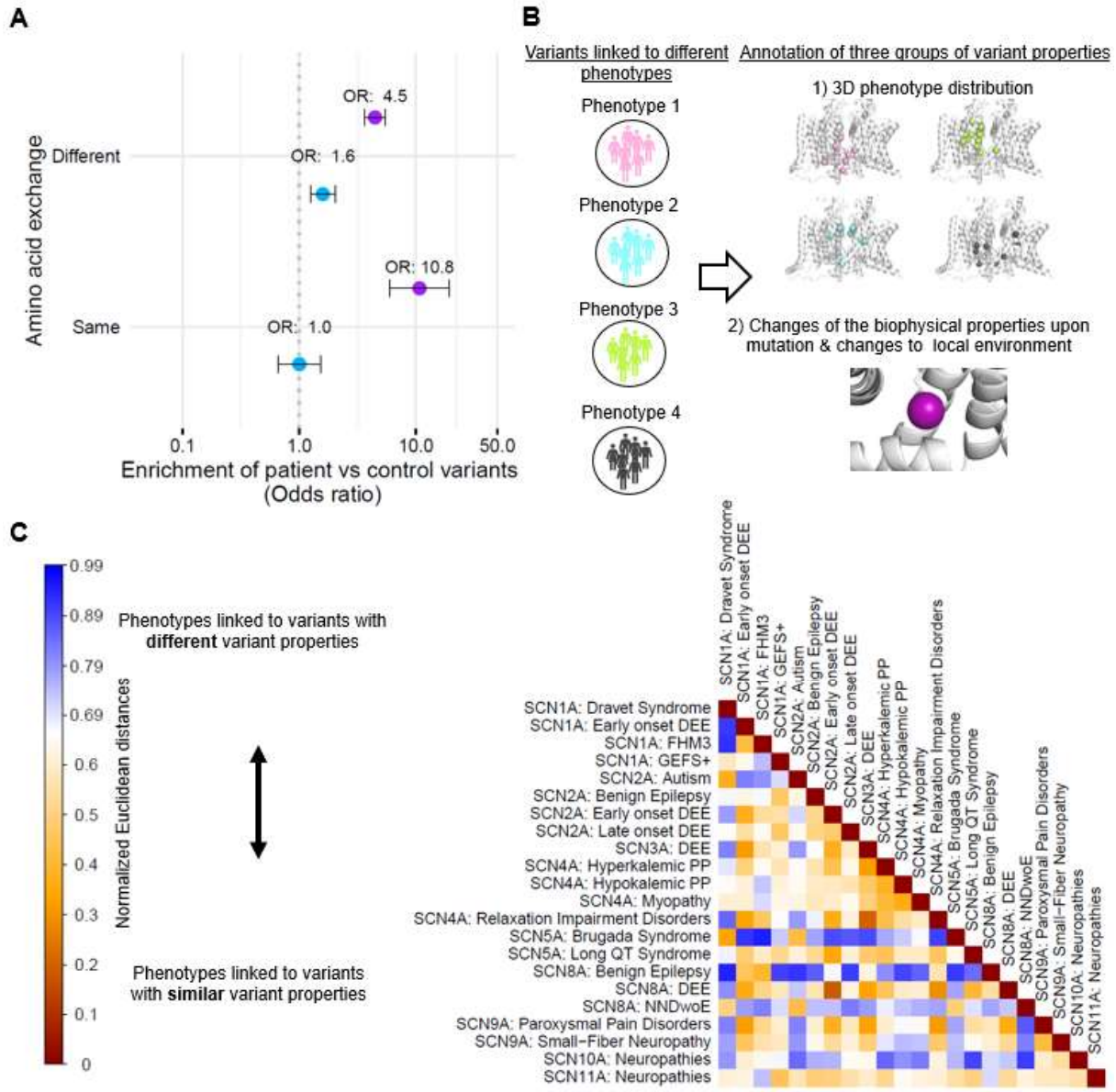


Figure 4: Paralogous patient variants as a proxy for pathogenicity and revealing phenotype-pairs associated with similar variant properties. A) The forest plot shows a meta-analysis of the enrichment of patient variants compared to control variants when paralogous patient variants are observed. The analysis was performed separately considering paralogous variants with the same or different amino acid exchange as a proxy. B) A cartoon describing the annotation of the variant properties to variants associated with the different phenotypes. The properties of a variant are defined as 1) the distribution of the variants in the 3D protein structure and 2) the biophysical properties of the amino acid exchange (see Methods for details). C) Heatmap visualizing phenotype pairs associated with similar or different variant properties. Phenotype-pairs associated with variants with similar properties are colored in red. Phenotype-pairs associated with different variant properties are colored in blue. Abbreviations: DEE – developmental epileptic encephalopathy, FHM3 – familial hemiplegic migraine type 3, PP: periodic paralysis, NDDwoE: neurodevelopmental disorders without epilepsy.

Integration of phenotype-pairs into the paralogous variant-based criteria improves their evidence for pathogenicity

We then investigated whether limiting our approach to paralogous pathogenic variants associated with similar variant properties could improve the evidence that individual paralogous pathogenic variants can provide for variant pathogenicity. We observed that taking only the subset of paralogous patient variants linked to phenotypes with similar variant properties as a proxy for pathogenicity resulted in an increased enrichment of patient versus control variants in the target gene, compared to using paralogous patient variants with phenotypes that are associated with different variant properties (Supplementary Figure 2). However, the variants to identify phenotype similarities and those used to test the enrichment of variants from patient to control variants were not independent. Thus, we validated and repeated the assessment and randomly split our dataset to gather an independent cohort (see Methods for details). Again we observed a 2-11-fold increased enrichment of patient variants versus control variants when considering paralogous patient variants linked to phenotypes with similar variant properties as a proxy for variant pathogenicity, compared to variants linked to phenotypes with different variant properties (*SCN8A* developmental epileptic encephalopathy: Paralogous variants linked to phenotypes that were associated with similar variant properties: OR = 75.8, P = 1.1e-13; Paralogous variants linked to phenotypes that were associated with similar variant properties: OR = 6.4, P = 4.9e-4, Figure 5).

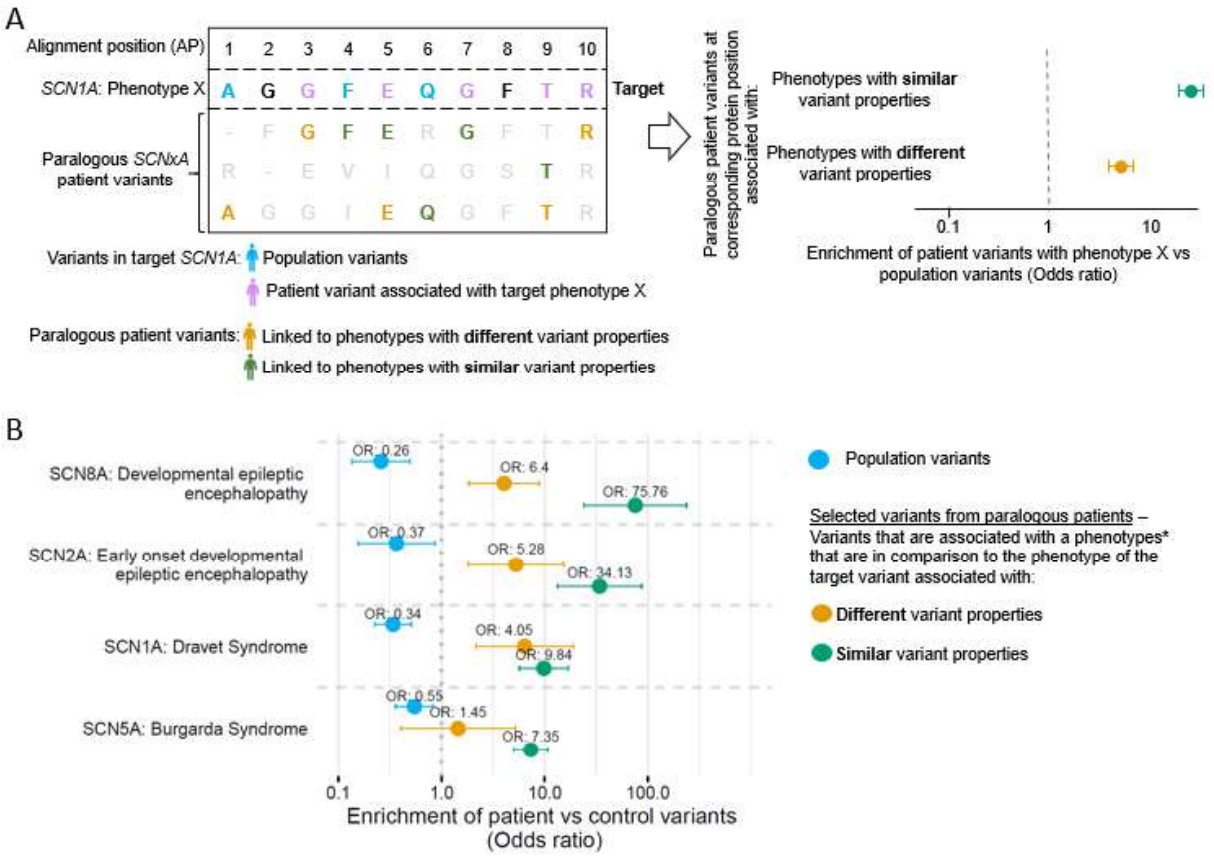


Figure 5: Incorporation of phenotype correlations improves the selection of paralogous variants for the paralog-based pathogenicity assessment. A) Cartoon describing the incorporation of phenotype similarities to select paralogous patient variants as a proxy for pathogenicity. **B)** The forest plot illustrates the enrichment of patient versus control variants for four phenotypes, where the paralogous variants that served as proxies were selected according to the phenotype similarity of the target variant. The plot distinguishes between cases where only paralogous patient variants associated with phenotypes with similar variant properties (green) were selected and those where only paralogous patient variants associated with phenotypes with different variant properties (indicated in yellow) were selected. *For the calculation we selected each of the five phenotypes associated with paralogous variants that showed the most similar/most different variant properties.

Use case examples

To demonstrate how genetic counselors, clinicians, and other professionals working with patients with genetic disorders could use our approach to compensate for a lack of other information to conclusively classify a variant, we give two examples of clinical variant assessment using the extended paralog PM5/PS1 criteria (Figure 6).

Case 1 is a 24-year-old woman assessed for recurrent episodes of syncope and diagnosed with Brugada syndrome upon electroclinical evaluation. Genetic testing revealed a heterozygous variant in *SCN5A* NM_000335.5: c.1252G>A (p.N834D). Since the patient was orphaned, no family history was available and no relatives could be tested. The variant is absent in the gnomAD database (moderate criterion

PM2), computational evidence supports its pathogenicity (CADD[36] = 26.9, PolyPhen-2[37] = 0.82, EVE[38] = 0.74 – supporting criterion PP3), and the patient's phenotype is specific (supporting criterion PP4). Using applicable ACMG criteria, the variant is classified as VUS and leaves the diagnosis uncertain. However, we identified a paralogous pathogenic variant in *SCN2A* at the same alignment position p.(N876T). As paralogous pathogenic variants in VGSCs provide moderate evidence for pathogenicity (Supplementary Table 2) we applied the extended moderate criterion PM5 and reclassified the previous VUS as likely pathogenic applying in total two moderate and two supporting criteria.

The second case is a 7-year-old male individual with moderate intellectual disability, developmental delay, and multifocal seizures starting at the age of 7 months. The genetic panel showed a heterozygous missense variant in *SCN8A* NM_001330260.2: c.2942G>T, p.(S981I). The patient's parents were not available for testing; therefore, the inheritance of the variant is unknown. The available evidence is not sufficient to classify the variant as (likely)-benign or (likely)-pathogenic, leaving the variant as VUS, with evidence for pathogenicity from the PM2 criterion (being absent in the population databases) and the two supporting criteria PP2 (occurring in a gene with a low rate of benign variation) and PP3 (computational evidence predicts the variant as pathogenic: CADD = 32, PolyPhen=1, and EVE=0.53). However, we observed a likely pathogenic variant in a case with an *SCN2A* variant at p.S987I that is presented with early-onset DEE, at the same alignment position. The phenotypes of our case and the individual with the *SCN2A* variant are associated with variants that have similar variant properties (Supplementary Table 4). Consequently, the paralogous variant can be considered as strong evidence for pathogenicity (PS1) enabling a reclassification of the *SCN8A* variant to likely pathogenic as having one strong and two supporting criteria.

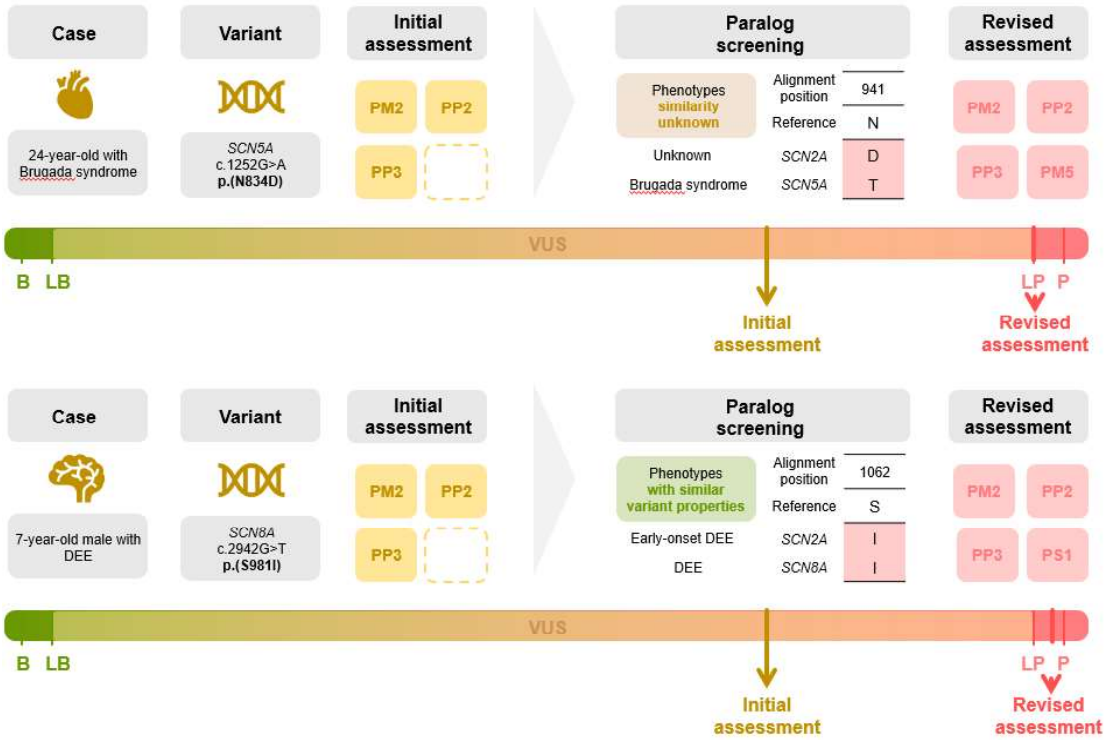


Figure 6: Case examples to demonstrate the utility of the extended PS1/PM5 criteria.

Discussion

We demonstrate a novel approach for clinical variant interpretation that utilizes individual paralogous patient variants within a gene family as a proxy for variant pathogenicity. Our analysis reveals that missense variants at amino acid positions where a pathogenic variant has been observed in a paralogous gene at a corresponding position are 6.3 times more likely to be pathogenic rather than benign. In a case study, we focus on the voltage-gated sodium channel (VGSC) and show that assessing the phenotype correlations across different VGSC genes can improve pathogenicity assessment based on individual paralogous pathogenic variants. The VGSC gene family is one of the most frequent genetic causes of neurodevelopmental disorders[39], and it exhibits high pleiotropy[7], making it an ideal target to explore variants associated with similar properties across different VGSC genes.

Despite efforts to standardize criteria for pathogenicity assignment[8] and many improvements in variant interpretation, about 80% of missense variants in the ClinVar database[9] (December 2022) have conflicting reports of pathogenicity, no interpretation, or are annotated as variants of unknown significance (VUS). We demonstrated that the PS1 criterion of the ACMG guidelines, which considers knowledge about an established pathogenic variant with the same amino acid change, as a strong evidence criterion for variant pathogenicity, can be expanded to include paralogous pathogenic variants, with, however, reduced strength. The 6.3-fold enrichment of patient versus control variants which we observed across all gene families, can be considered moderate evidence for pathogenicity

according to the Bayesian framework proposed by Tavtigitan et al, 2018[35]. We observed that the enrichment of patient versus control variants differed among different genes and gene families. Although the average odds ratio fulfills the criteria for moderate strength, we also identified 129 genes with an enrichment above 18.7, which can be considered as strong criterion for variant interpretation. Similarly, the PM5 criterion, which identifies a novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before, could be expanded to include paralogous genes following our results.

The analysis of clusters of pathogenic variants across paralogous genes along protein sequences had previously identified regions with significant enrichment for patient variants[14]. However, despite these findings, more than 70% of paralogous patient variants are located outside these patient-enriched clusters and therefore cannot be used to inform variant pathogenicity. In our study, we demonstrate that individual pathogenic variants can serve as a reliable proxy for variant pathogenicity and provide a catalog for each gene and gene family to quantify the strength of such variants. Pathogenic variants are most commonly found in protein regions that are conserved across all paralogous genes of a gene family[12]. Our results show that including individual paralogous pathogenic variants in variant interpretation, in addition to paralog conservation, can further refine variant interpretation. Additionally, even within non-paralog conserved regions, which are typically depleted for patient variants, the presence of a paralogous pathogenic variant highlights variants that may be susceptible to disease-causing mutations and are missed otherwise.

While the phenotypes associated with pathogenic variants in VGSC genes differ widely in their clinical presentation among the different genes[4,23,25,27], we identified phenotypes across these paralogous genes that may be clinically different but are caused by the same paralogous variants or by variants with similar biophysical properties. Previous evidence demonstrated the conservation of functional effects of variants at analogous positions across VGSC genes[7,17]. In line with this observation, we observed that paralogous patient variants with phenotypes linked to similar variant properties and potentially having a similar molecular effect serve as better proxies for variant pathogenicity than paralogous patient variants with different variant properties. In cases where a patient variant overlaps with a paralogous pathogenic variant that is associated with a phenotype linked to different variant properties, these functional effects might be less conserved across the VGSC genes.

Our approach comes with a few limitations. One limitation is that it does not cover gene-specific domains that are not conserved in any other paralogous genes, as pathogenic variants from paralogous genes cannot overlap in these regions. However, previous studies have shown that the majority of functionally important regions are conserved across paralogs[12,16,18], and *de novo* domains are rare and could be annotated separately. To identify variant similarities among phenotypes, we grouped

phenotypes into distinct categories. However, it is important to note that phenotypes may overlap, and as a consequence, phenotype categories may change in the future. Nonetheless, we were able to collect patient data from leading clinicians in the diagnosis of VGSC-related disorders, ensuring the highest possible phenotype quality to date. Finally, many gene families lack a sufficiently large collection of identified pathogenic variants to allow statistical testing to prove the validity of applying paralogous pathogenic variants as a proxy for pathogenicity. On an individual gene or gene family level, there are gene families where our approach lacks sufficient power to demonstrate the applicability of our approach to that gene family.

In conclusion, our study demonstrates that integrating paralogous pathogenic variants can expand existing variant interpretation guidelines. As reference databases continue to grow and include well-classified pathogenic variants, the applicability of our framework will increase further in the future and enable quantification of the evidence strength for more gene families. Additionally, the utility of integrating phenotype correlation across VGSC into the interpretation framework suggests that this extension might be also applicable to other gene families where individual gene-level pleiotropy is known. These findings suggest that our approach has significant potential to enhance variant interpretation and aid in the diagnosis of pathogenic variants in clinical settings.

Funding

Funding for this work was provided by the German Federal Ministry for Education and Research (BMBF, Treat-ION, 01GM1907D) to D.L., T.B., and P.M., the Fonds Nationale de la Recherche in Luxembourg (Research Unit FOR-2715, FNR grant NTER/DFG/21/16394868 MechEPI2) to P.M., the Agencia Nacional de Investigación y Desarrollo (ANID, PAI77200124) of Chile to E.P., the Familie SCN2A foundation 2020 Action Potential Grant to E.P., the Dravet Syndrome Foundation (grant number, 272016) to D.L, and the NIH NINDS (Channelopathy-Associated Epilepsy Research Center, 5-U54-NS108874) to D.L.

Competing interests

The authors report no competing interests.

References

1. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

2. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* **106**, 19096–19101 (2009).

3. Milman, A. *et al.* Genotype-Phenotype Correlation of SCN5A Genotype in Patients With Brugada Syndrome and Arrhythmic Events: Insights From the SABRUS in 392 Probands. *Circ Genom Precis Med* **14**, e003222 (2021).

4. Johannesen, K. M. *et al.* Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. *Brain* **145**, 2991–3009 (2022).

5. Kamada, F. *et al.* A novel KCNQ4 one-base deletion in a large pedigree with hearing loss: implication for the genotype-phenotype correlation. *J Hum Genet* **51**, 455–460 (2006).

6. McRae, J. F. *et al.* Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).

7. Brunklaus, A. *et al.* Biological concepts in human sodium channel epilepsies and their relevance in clinical practice. *Epilepsia* **61**, 387–399 (2020).

8. 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).

9. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

10. Dickerson, J. E. & Robertson, D. L. On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication. *Mol Biol Evol* **29**, 61–69 (2012).

11. Yates, B., Gray, K. A., Jones, T. E. M. & Bruford, E. A. Updates to HCOP: the HGNC comparison of orthology predictions tool. *Briefings in Bioinformatics* (2021) doi:10.1093/bib/bbab155.

12. Lal, D. *et al.* Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders. *Genome Med* **12**, 28 (2020).

13. 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).

14. Pérez-Palma, E. *et al.* Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res* **30**, 62–71 (2020).

15. Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G. & Gilissen, C. Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum Mutat* **38**, 1454–1463 (2017).

16. Ware, J. S., Walsh, R., Cunningham, F., Birney, E. & Cook, S. A. Paralogous annotation of disease-causing variants in long QT syndrome genes. *Hum Mutat* **33**, 1188–1191 (2012).

17. Brunklaus, A. *et al.* Gene variant effects across sodium channelopathies predict function and guide precision therapy. *Brain* awac006 (2022) doi:10.1093/brain/awac006.

18. Walsh, R., Peters, N. S., Cook, S. A. & Ware, J. S. Paralogue annotation identifies novel pathogenic variants in patients with Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia. *J Med Genet* **51**, 35–44 (2014).

19. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).

20. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754–D761 (2018).

21. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

22. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

23. Brunklaus, A. *et al.* The gain of function SCN1A disorder spectrum: novel epilepsy phenotypes and therapeutic implications. *Brain* awac210 (2022) doi:10.1093/brain/awac210.

24. Brunklaus, A. *et al.* Development and Validation of a Prediction Model for Early Diagnosis of SCN1A-Related Epilepsies. *Neurology* **98**, e1163–e1174 (2022).

25. Wolff, M. *et al.* Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. *Brain* **140**, 1316–1336 (2017).

26. Crawford, K. *et al.* Computational analysis of 10,860 phenotypic annotations in individuals with SCN2A-related disorders. *Genet Med* **23**, 1263–1272 (2021).

27. Zaman, T. *et al.* SCN3A-related neurodevelopmental disorder: A spectrum of epilepsy and brain malformation. *Ann Neurol* **88**, 348–362 (2020).

28. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum Mutat* **36**, 928–930 (2015).

29. Walsh, R. *et al.* Enhancing rare variant interpretation in inherited arrhythmias through quantitative analysis of consortium disease cohorts and population controls. *Genet Med* **23**, 47–58 (2021).

30. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* **19**, 1151–1158 (2017).

31. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004).

32. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software* **36**, 1–48 (2010).

33. Iqbal, S. *et al.* Delineation of functionally essential protein regions for 242 neurodevelopmental disorders. *Brain* awac381 (2022) doi:10.1093/brain/awac381.

34. Grantham, R. Amino Acid Difference Formula to Help Explain Protein Evolution. *Science* **185**, 862–864 (1974).

35. Tavtigian, S. V. *et al.* Modeling the ACMG/AMP Variant Classification Guidelines as a Bayesian Classification Framework. *Genet Med* **20**, 1054–1060 (2018).

36. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research* **47**, D886–D894 (2019).

37. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).

38. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).

39. Heyne, H. O. *et al.* De novo variants in neurodevelopmental disorders with epilepsy. *Nat. Genet.* **50**, 1048–1053 (2018).

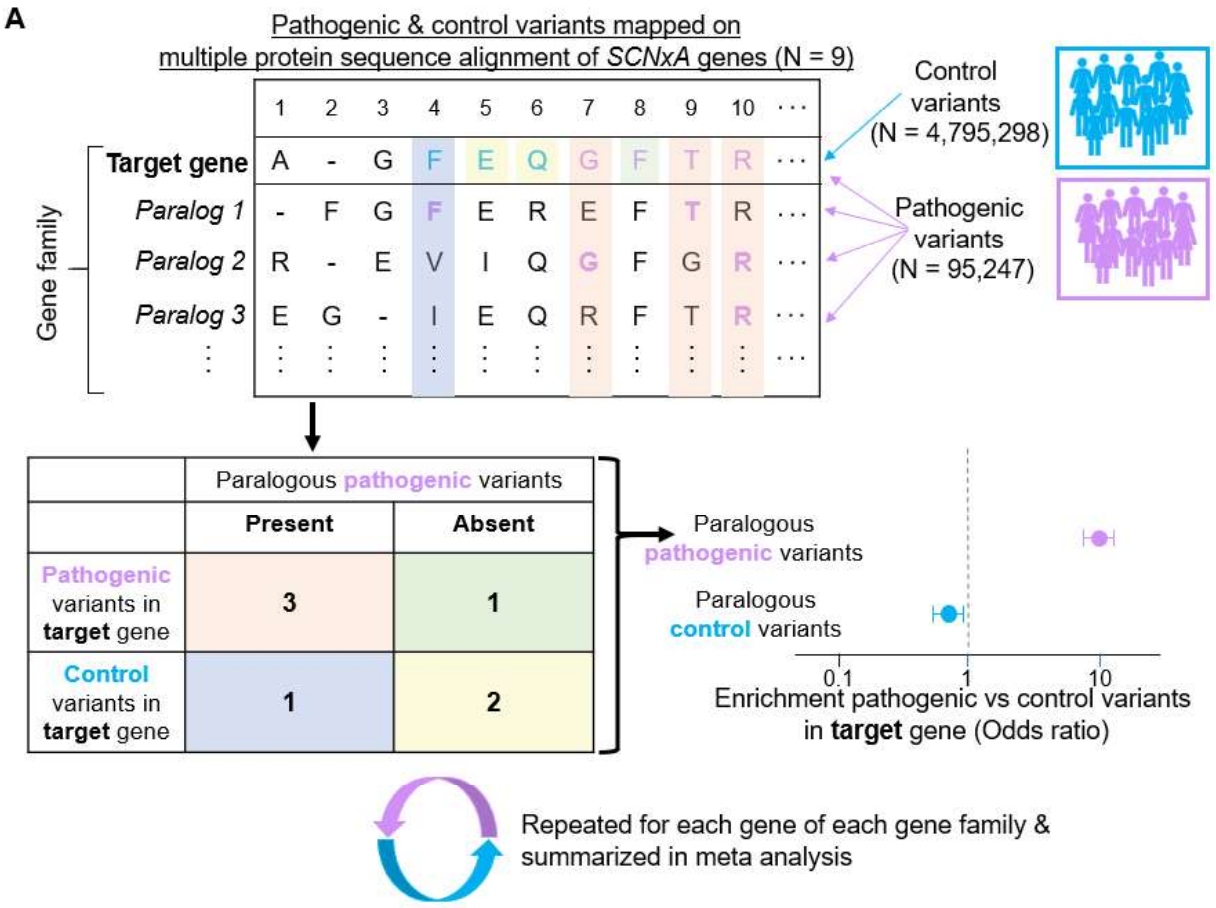
Supplementary Material

Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria

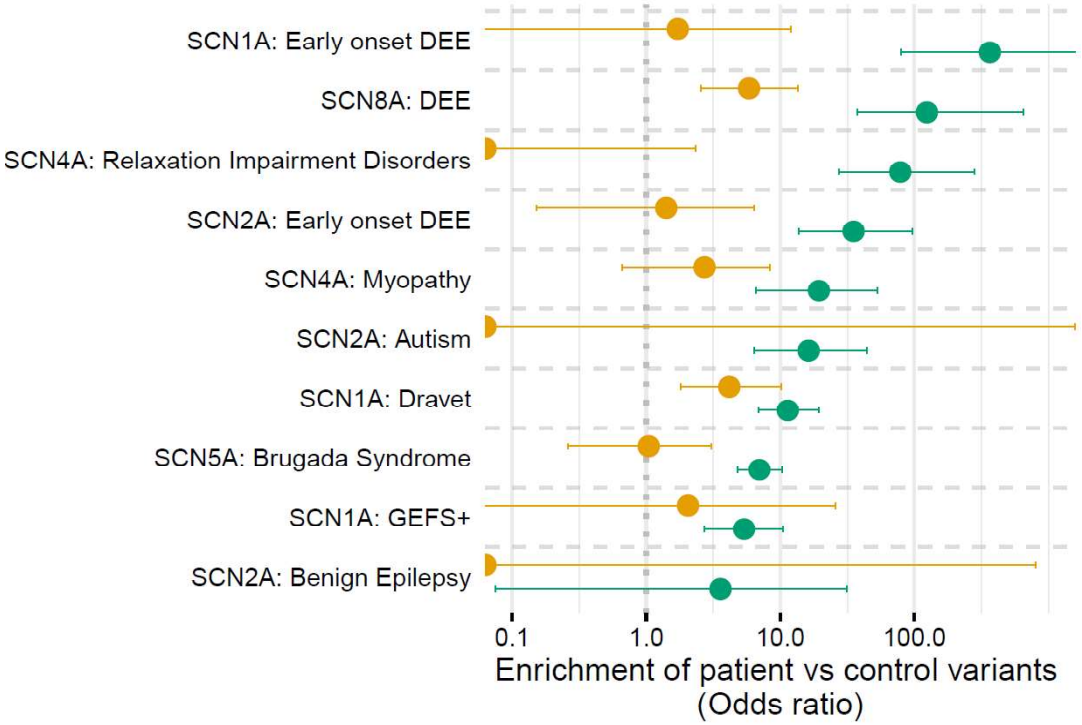
Tobias Brünger[+], Alina Ivaniuk[+], Eduardo Pérez-Palma, Katrine M Johannesen, Ingo Helbig, Sameer Zuberi, Michael Nothnagel, Rikke S Møller, Patrick May, Andreas Brunklaus, Dennis Lal. (in preparation for submission to Genetics in Medicine)

[+]These authors contributed equally to the work

The supplementary material is listed below and accessible at https://github.com/TobiasBruenger/ACMG_extension. Due to its large size Supplementary Table 1 and Supplementary Table 3 are only available online at https://github.com/TobiasBruenger/ACMG_extension.



Supplementary Figure 1: Cartoon to visualize the calculation of the enrichment of pathogenic versus control in the presence of paralogous pathogenic variants. Across all pathogenic variants in a target gene we counted 1) the number of pathogenic variants in the target gene that with at least one conserved pathogenic variant in a paralogous gene at a corresponding protein position (orange), 2) the number of pathogenic variants in the target gene that do not overlap with a pathogenic variant in another gene (green), 3) the number of control variants in the target gene with a conserved pathogenic variant in any paralogous gene at a corresponding protein position (blue), and 4) the number of control variants in the target gene that do not overlap with any paralogous pathogenic variant (yellow). We then calculated the enrichment using a 2x2 Fisher's Exact test. The same procedure was repeated for each gene of a gene family that was associated with pathogenic variants. Gene-wise results are summarized in a meta-analysis. Finally, as for the control group, the same procedure was repeated taking conserved paralogous control variants as prior instead of paralogous pathogenic variants.



Supplementary Figure 2: Incorporation of phenotype correlations improves the selection of paralogous variants for the paralog-based pathogenicity assessment. The forest plot illustrates the enrichment of affected individuals versus control variants for all phenotypes represented by more than 20 individuals in our cohort of affected individuals. In contrast to the previous approach we now select different groups of the paralogous variants of affected individuals. The selection of the variants was based on the phenotype similarity to the target phenotype. Here we considered two groups of variants 1) paralogous variants associated with phenotypes linked to similar variant properties as the variant in the target gene were considered (variants associated with the five most similar phenotypes were considered, results highlighted in green) or 2) paralogous variants associated with phenotypes linked to different variant properties as the variant in the target gene were considered (variants associated with five most different phenotypes were considered, results highlighted in orange) (see Methods for details).

Supplementary Table 2: Weights of the extended ACMG criteria (extended PS1/PM5) separately calculated for each gene family. The weights are assigned based on the enrichment of pathogenic vs control variants that we calculated if a corresponding pathogenic variant is observed in a paralogous gene. The cutoffs for the different weights are obtained from the Bayesian framework as proposed by Tavtigitan et al., 2018.

Genes of each gene family	Amino acid exchange of paralogous pathogenic variant	Weight of ACMG criteria
RYR2;RYR3;RYR1	Different exchange	Supporting
LHCGR;FSHR;TSHR	Different exchange	Supporting
POMT1;POMT2	Different exchange	Supporting
DES;NEFL;VIM;INA;PRPH;GFAP	Different exchange	Supporting
RDH5;SDR9C7;HSD11B2;KDSR	Different exchange	Moderate
MAF;MAFB	Different exchange	Strong
SLC19A2;SLC19A3	Different exchange	Supporting
SCO1;SCO2	Different exchange	Strong
F13A1;TGM1;EPB42;TGM5;TGM6;TGM2	Different exchange	Moderate
SLC12A2;SLC12A1;SLC12A3	Different exchange	Supporting
GRIN2B;GRIN2A;GRIN2C;GRIN2D	Different exchange	Strong
PAX8;PAX4;PAX5;PAX2;PAX6;PAX9	Different exchange	Moderate
HOXD13;HOXD10;HOXA10;HOXA13; HOXC13;HOXB13	Different exchange	Moderate
GRID2;GRIA1;GRIK2;GRIA4;GRIK4;GRIA3	Different exchange	Strong
FAM20C;FAM20A	Different exchange	Moderate
NR2F1;RXRB;RXRA;NR2E3;NR2F2	Different exchange	Strong
WRN;BLM	Different exchange	Moderate
GATA2;GATA3	Different exchange	Moderate
GABRA2;GABRA6;GABRA1;GABRG2;GABRA5; GABRE;GABRA3	Different exchange	Moderate
EPAS1;NPAS2;SIM1	Different exchange	Moderate
HEXB;HEXA	Different exchange	Supporting
CRYGD;CRYGC;CRYGA;CRYBA2;CRYGS;CRYBA1; CRYBB3;CRYBB2;CRYBB1;CRYBA4	Different exchange	Supporting
KRT25;KRT10;KRT12;KRT13;KRT14;KRT16;KRT17	Different exchange	Supporting
RAF1;BRAF;ARAF	Different exchange	Moderate
PIEZO1;PIEZO2	Different exchange	Strong
VANGL1;VANGL2	Different exchange	Strong
SOX11;SOX4;SOX17;SOX8;SOX9;SOX10	Different exchange	Strong
G6PC;G6PC3	Different exchange	Moderate
FAT1;FAT2	Different exchange	Strong
SLC22A4;SLC22A5;SLC22A12	Different exchange	Supporting
CDC42;RAC1;RAC3;RAC2	Different exchange	Moderate
CHST3;CHST6	Different exchange	Supporting
CNNM4;CNNM2	Different exchange	Strong
PIK3R1;PIK3R2	Different exchange	Strong
CFTR;ABCC2;ABCC6	Different exchange	Supporting
MCCC1;PCCA	Different exchange	Moderate

CHD4;CHD3	Different exchange	Strong
ACTN2;ACTN1;ACTN4	Different exchange	Supporting
ALOX12B;ALOXE3	Different exchange	Moderate
GNAI3;GNAT2;GNAT1;GNAI2;GNA14;GNAQ; GNAO1;GNA11	Different exchange	Moderate
ECE1;ECEL1;MME;PHEX	Different exchange	Supporting
NPR2;GUCY2C;GUCY2D	Different exchange	Moderate
ALDH7A1;ALDH1B1;ALDH1A2;ALDH1A3	Different exchange	Supporting
CACNA1E;CACNA1A	Different exchange	Moderate
CACNA1H;CACNA1G	Different exchange	Moderate
ATP6V0A4;TCIRG1;ATP6V0A2	Different exchange	Moderate
ATP7B;ATP7A	Different exchange	Moderate
CLCNKA;CLCNKB;CLCN7;CLCN4;CLCN5	Different exchange	Supporting
RP1L1;RP1	Different exchange	Moderate
TYRP1;TYR	Different exchange	Supporting
SMARCA2;SMARCA4	Different exchange	Strong
TRPV4;TRPV3	Different exchange	Strong
SGCE;SGCA	Different exchange	Moderate
IGF1R;INSR	Different exchange	Moderate
RUNX2;RUNX1	Different exchange	Strong
LAMA3;LAMA5	Different exchange	Strong
SLC52A2;SLC52A3	Different exchange	Supporting
LRBA;NBEA	Different exchange	Strong
SLC6A9;SLC6A1;SLC6A3;SLC6A5;SLC6A2; SLC6A8	Different exchange	Supporting
SERPINI1;SERPINE1;SERPINA6;SERPINA1;SERPINA3; SERPIND1;SERPINA7	Different exchange	Moderate
STAT1;STAT4;STAT2;STAT5B;STAT3	Different exchange	Supporting
SLC26A1;SLC26A2;SLC26A8;SLC26A5; SLC26A4;SLC26A3	Different exchange	Supporting
TUBB2A;TUBB2B;TUBB;TUBB4B;TUBB8; TUBB3;TUBB6;TUBB4A;TUBB1	Different exchange	Moderate
CYP27A1;CYP11B1;CYP11B2;CYP27B1; CYP11A1;CYP24A1	Different exchange	Supporting
CAMK2A;CAMK2B;CAMK2G	Different exchange	Strong
KCNQ5;KCNQ3;KCNQ2	Different exchange	Moderate
DNM1;DNM1L;DNM2	Different exchange	Moderate
TBX19;TBX18;TBX20;TBX5;TBX6;TBX4; TBX1;TBX22	Different exchange	Moderate
CDK13;CDK12	Different exchange	Strong
ITGB6;ITGB3;ITGB2	Different exchange	Moderate
GABRD;GABRB1;GLRB;GLRA1;GABRB2; GABRB3;GLRA2	Different exchange	Moderate
WNT10A;WNT8A;WNT10B;WNT9B	Different exchange	Moderate
SLC34A2;SLC34A1;SLC34A3	Different exchange	Moderate
ATP6V1B1;ATP6V1B2	Different exchange	Strong

CREBBP;EP300	Different exchange	Moderate
ARID1A;ARID1B	Different exchange	Strong
ZIC1;ZIC2;ZIC3	Different exchange	Strong
CRAT;CHAT;CPT1A;CPT1C	Different exchange	Moderate
GJA5;GJA8;GJC2;GJA1;GJA3;GJC1	Different exchange	Supporting
COL6A1;COL6A2	Different exchange	Strong
KCNA2;KCNA4;KCNA1;KCNA5	Different exchange	Supporting
PSEN2;PSEN1	Different exchange	Supporting
FLNB;FLNC;FLNA	Different exchange	Moderate
KCNJ10;KCNJ11;KCNJ1;KCNJ5;KCNJ8;KCNJ18; KCNJ16;KCNJ2;KCNJ6	Different exchange	Moderate
COL11A1;COL3A1;COL5A2; COL4A4;COL4A3; COL11A2;COL1A2;COL27A1;COL5A1; COL2A1;COL4A1;COL4A2; COL1A1;COL4A6;COL4A5	Different exchange	Moderate
ERBB4;EGFR;ERBB3;ERBB2	Different exchange	Moderate
PDGFRA;KIT;CSF1R;PDGFRB;FLT3	Different exchange	Moderate
DMBX1;OTX2;CRX	Different exchange	Moderate
CARD11;CARD14;CARD10	Different exchange	Moderate
RHD;RHAG	Different exchange	Strong
DNAH6;DNAH7;DNAH1;DNAH5;DNAH11; DNAH10;DYNC1H1;DNAH2;DNAH9	Different exchange	Supporting
F5;F8	Different exchange	Moderate
VPS13A;VPS13C	Different exchange	Strong
NEUROD1;NEUROD2	Different exchange	Strong
KCNH1;CNGA3;KCNH7;CNGA1;HCN1;KCNH2; CNGB3;KCNH5;HCN2	Different exchange	Supporting
NOTCH2;NOTCH1;NOTCH3	Different exchange	Moderate
FBN2;FBN1;FBN3	Different exchange	Supporting
RYR2;RYR3;RYR1	Same exchange	Moderate
NR4A2;NR3C2;NR3C1;PGR;AR	Same exchange	Moderate
LHCGR;FSHR;TSHR	Same exchange	Moderate
POMT1;POMT2	Same exchange	Moderate
DES;NEFL;VIM;INA;PRPH;GFAP	Same exchange	Moderate
SLC17A3;SLC17A5;SLC17A8;SLC17A9	Same exchange	Moderate
MAF;MAFB	Same exchange	Strong
F13A1;TGM1;EPB42;TGM5;TGM6;TGM2	Same exchange	Supporting
GRIN2B;GRIN2A;GRIN2C;GRIN2D	Same exchange	Strong
PAX8;PAX4;PAX5;PAX2;PAX6;PAX9	Same exchange	Moderate
TRPM3;TRPM6;TRPM1;TRPM2	Same exchange	Strong
ROBO2;ROBO1;ROBO3	Same exchange	Moderate
FAM20C;FAM20A	Same exchange	Strong
TTPA;RLBP1	Same exchange	Strong
CYP1B1;CYP21A2;CYP17A1	Same exchange	Supporting
WRN;BLM	Same exchange	Moderate

GATA2;GATA3	Same exchange	Strong
DGUOK;TK2	Same exchange	Supporting
EPAS1;NPAS2;SIM1	Same exchange	Strong
HEXB;HEXA	Same exchange	Moderate
CYP4V2;TBXAS1;CYP4F22	Same exchange	Moderate
KRT25;KRT10;KRT12;KRT13;KRT14; KRT16; KRT17	Same exchange	Strong
RAF1;BRAF;ARAF	Same exchange	Moderate
LIPH;LPL;PNLIP;LIPC;LIPG;LIPI	Same exchange	Moderate
PIEZO1;PIEZO2	Same exchange	Strong
VARS2;VARS	Same exchange	Moderate
SOX11;SOX4;SOX17;SOX8;SOX9;SOX10	Same exchange	Strong
PIK3CD;PIK3CB;PIK3CA	Same exchange	Strong
ABCB6;ABCB7	Same exchange	Moderate
PTH2R;PTH1R;GLP1R;GHRHR;GCGR	Same exchange	Moderate
G6PC;G6PC3	Same exchange	Moderate
FAT1;FAT2	Same exchange	Strong
CELSR2;CELSR3;CELSR1	Same exchange	Moderate
TENM3;TENM4;TENM1	Same exchange	Strong
ABCC8;ABCC9	Same exchange	Moderate
CNNM4;CNNM2	Same exchange	Strong
PIK3R1;PIK3R2	Same exchange	Strong
EFEMP1;EFEMP2;FBLN5	Same exchange	Moderate
CFTR;ABCC2;ABCC6	Same exchange	Supporting
CHD4;CHD3	Same exchange	Strong
CHD7;CHD8	Same exchange	Moderate
CHRNB2;CHRNA1;CHRND;CHRNG;CHRNA2;CHRNE; CHRNB1;CHRNA4	Same exchange	Moderate
SPTA1;SPTAN1;SPTBN2;SPTB;SPTBN4	Same exchange	Moderate
SLC4A4;SLC4A1	Same exchange	Moderate
SAMD9;SAMD9L	Same exchange	Strong
GNAI3;GNAT2;GNAT1;GNAI2;GNA14;GNAQ; GNAO1;GNA11	Same exchange	Moderate
PROC;F7;F10;PROZ;F9	Same exchange	Supporting
ECE1;ECEL1;MME;PHEX	Same exchange	Moderate
NPR2;GUCY2C;GUCY2D	Same exchange	Moderate
CACNA1E;CACNA1A	Same exchange	Strong
CACNA1H;CACNA1G	Same exchange	Strong
SCN3A;SCN2A;SCN1A;SCN9A;SCN5A;SCN10A; SCN11A;SCN8A;SCN4A	Same exchange	Moderate
AQP1;AQP2;AQP5;MIP;AQP4	Same exchange	Moderate
ATP6V0A4;TCIRG1;ATP6V0A2	Same exchange	Moderate
ATP7B;ATP7A	Same exchange	Moderate
RP1L1;RP1	Same exchange	Strong
TYRP1;TYR	Same exchange	Supporting

MAP2K1;MAP2K2	Same exchange	Strong
SGCE;SGCA	Same exchange	Moderate
PDE11A;PDE6B;PDE6A;PDE10A;PDE6C;PDE2A	Same exchange	Moderate
RUNX2;RUNX1	Same exchange	Strong
MYH6;MYH7;MYH11;MYH10;MYH8;MYH1; MYH2;MYH3;MYH14;MYH7B;MYH9	Same exchange	Supporting
GALNT2;GALNT14;GALNT3;GALNT12	Same exchange	Strong
ARSB;GALNS;ARSG;ARSA;ARSE;STS	Same exchange	Supporting
LAMA3;LAMA5	Same exchange	Strong
SLC52A2;SLC52A3	Same exchange	Moderate
LRBA;NBEA	Same exchange	Strong
SLC6A9;SLC6A1;SLC6A3;SLC6A5; SLC6A2;SLC6A8	Same exchange	Moderate
STAT1;STAT4;STAT2;STAT5B;STAT3	Same exchange	Moderate
SLC26A1;SLC26A2;SLC26A8;SLC26A5; SLC26A4;SLC26A3	Same exchange	Supporting
TUBB2A;TUBB2B;TUBB;TUBB4B;TUBB8; TUBB3;TUBB6;TUBB4A;TUBB1	Same exchange	Moderate
CYP27A1;CYP11B1;CYP11B2;CYP27B1; CYP11A1;CYP24A1	Same exchange	Moderate
KCNQ5;KCNQ3;KCNQ2	Same exchange	Strong
TBX19;TBX18;TBX20;TBX5;TBX6;TBX4; TBX1;TBX22	Same exchange	Supporting
ITGB6;ITGB3;ITGB2	Same exchange	Moderate
RPGRIP1;RPGRIP1L	Same exchange	Strong
GABRD;GABRB1;GLRB;GLRA1;GABRB2; GABRB3;GLRA2	Same exchange	Strong
ABCA4;ABCA1	Same exchange	Moderate
KRT81;KRT86;KRT83;KRT6B;KRT6C;KRT6A;KRT5; KRT71;KRT74;KRT2;KRT1;KRT3;KRT4;KRT8	Same exchange	Strong
CREBBP;EP300	Same exchange	Strong
HSPB3;HSPB1;CRYAB;HSPB8;HSPB6;CRYAA	Same exchange	Moderate
ARID1A;ARID1B	Same exchange	Strong
SOS1;SOS2	Same exchange	Strong
NRAS;MRAS;RALA;HRAS;RRAS2;KRAS;RRAS	Same exchange	Strong
DNMT3A;DNMT3B	Same exchange	Strong
IHH;SHH;DHH	Same exchange	Moderate
DSG1;DSG4;DSG2	Same exchange	Strong
GLI2;GLI3	Same exchange	Strong
ZIC1;ZIC2;ZIC3	Same exchange	Moderate
GJA5;GJA8;GJC2;GJA1;GJA3;GJC1	Same exchange	Moderate
COL6A1;COL6A2	Same exchange	Strong
KCNA2;KCNA4;KCNA1;KCNA5	Same exchange	Moderate
PSEN2;PSEN1	Same exchange	Moderate
FLNB;FLNC;FLNA	Same exchange	Moderate

KCNJ10;KCNJ11;KCNJ1;KCNJ5;KCNJ8;KCNJ18; KCNJ16;KCNJ2;KCNJ6	Same exchange	Supporting
COL11A1;COL3A1;COL5A2;COL4A4; COL4A3;COL11A2;COL1A2;COL27A1; COL5A1;COL2A1;COL4A1;COL4A2; COL1A1;COL4A6;COL4A5	Same exchange	Moderate
FGFR3;FGFR4;FGFR1;FGFR2	Same exchange	Supporting
PDGFRA;KIT;CSF1R;PDGFRB;FLT3	Same exchange	Strong
GJB4;GJB3;GJC3;GJB2;GJB6;GJB1	Same exchange	Supporting
DMBX1;OTX2;CRX	Same exchange	Moderate
DNAH6;DNAH7;DNAH1;DNAH5;DNAH11;DNAH10; DYNC1H1;DNAH2;DNAH9	Same exchange	Moderate
F5;F8	Same exchange	Moderate
VPS13A;VPS13C	Same exchange	Strong
HSD3B2;HSD3B7;NSDHL	Same exchange	Moderate
KCNH1;CNGA3;KCNH7;CNGA1;HCN1;KCNH2; CNGB3;KCNH5;HCN2	Same exchange	Supporting
NOTCH2;NOTCH1;NOTCH3	Same exchange	Moderate
FBN2;FBN1;FBN3	Same exchange	Moderate

Supplementary Table 4: Listed are, for every phenotype, those five phenotypes that were associated with the most similar variant properties as variants associated with the target phenotype. Similarly, those five phenotypes are listed that were associated with the most different variant properties as variants associated with the target phenotype.

Target phenotype	Phenotypes associated with paralogous genes with the most similar variant properties (N = 5)	Phenotypes associated with paralogous genes with most different variant properties (N = 5)
SCN1A: Dravet Syndrome	SCN5A: Brugada Syndrome; SCN2A: Autism; SCN8A: NNDwoE; SCN4A: Myopathy; SCN11A: Neuropathies	SCN8A: DEE; SCN3A: DEE; SCN9A: Paroxysmal Pain Disorders; SCN4A: Relaxation Impairment Disorders; SCN8A: Benign Epilepsy
SCN1A: Early onset DEE	SCN3A: DEE; SCN9A: Paroxysmal Pain Disorders; SCN4A: Relaxation Impairment Disorders; SCN8A: DEE; SCN2A: Early onset DEE	SCN11A: Neuropathies; SCN2A: Benign Epilepsy; SCN8A: NNDwoE; SCN2A: Autism; SCN5A: Brugada Syndrome
SCN1A: FHM3	SCN8A: Benign Epilepsy; SCN4A: Relaxation Impairment Disorders; SCN8A: DEE; SCN9A: Paroxysmal Pain Disorders; SCN2A: Early onset DEE	SCN4A: Hypokalemic PP; SCN11A: Neuropathies; SCN2A: Autism; SCN8A: NNDwoE; SCN5A: Brugada Syndrome
SCN1A: GEFS+	SCN11A: Neuropathies; SCN2A: Late onset DEE; SCN2A: Benign Epilepsy; SCN5A: Long QT Syndrome; SCN2A: Early onset DEE	SCN10A: Neuropathies; SCN2A: Autism; SCN8A: NNDwoE; SCN5A: Brugada Syndrome; SCN8A: Benign Epilepsy
SCN2A: Autism	SCN1A: Dravet Syndrome; SCN5A: Brugada Syndrome; SCN8A: NNDwoE; SCN11A: Neuropathies; SCN5A: Long QT Syndrome	SCN9A: Paroxysmal Pain Disorders; SCN10A: Neuropathies; SCN1A: FHM3; SCN1A: Early onset DEE; SCN8A: Benign Epilepsy
SCN2A: Benign Epilepsy	SCN11A: Neuropathies; SCN1A: GEFS+; SCN5A: Long QT Syndrome; SCN4A: Hyperkalemic PP; SCN8A: DEE	SCN5A: Brugada Syndrome; SCN1A: Early onset DEE; SCN1A: FHM3; SCN1A: Dravet Syndrome; SCN8A: Benign Epilepsy
SCN2A: Early onset DEE	SCN8A: DEE; SCN5A: Long QT Syndrome; SCN4A: Relaxation Impairment Disorders; SCN3A: DEE; SCN9A: Paroxysmal Pain Disorders	SCN8A: Benign Epilepsy; SCN8A: NNDwoE; SCN10A: Neuropathies; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN2A: Late-onset DEE	SCN11A: Neuropathies; SCN1A: GEFS+; SCN4A: Hypokalemic PP; SCN9A: Paroxysmal Pain Disorders; SCN3A: DEE	SCN8A: NNDwoE; SCN1A: FHM3; SCN5A: Brugada Syndrome; SCN8A: Benign Epilepsy; SCN1A: Dravet Syndrome

SCN3A: DEE	SCN4A: Relaxation Impairment Disorders; SCN4A: Hyperkalemic PP; SCN9A: Paroxysmal Pain Disorders; SCN8A: DEE; SCN1A: Early onset DEE	SCN10A: Neuropathies; SCN2A: Autism; SCN8A: NNDwoE; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN4A: Hyperkalemic PP	SCN3A: DEE; SCN2A: Early onset DEE; SCN5A: Long QT Syndrome; SCN8A: DEE; SCN2A: Benign Epilepsy	SCN5A: Brugada Syndrome; SCN8A: Benign Epilepsy; SCN1A: FHM3; SCN1A: Dravet Syndrome; SCN10A: Neuropathies
SCN4A: Hypokalemic PP	SCN2A: Late onset DEE; SCN3A: DEE; SCN2A: Benign Epilepsy; SCN5A: Long QT Syndrome; SCN2A: Early onset DEE	SCN9A: Small-Fiber Neuropathy; SCN1A: Dravet Syndrome; SCN10A: Neuropathies; SCN1A: FHM3; SCN8A: Benign Epilepsy
SCN4A: Myopathy	SCN3A: DEE; SCN2A: Late onset DEE; SCN2A: Benign Epilepsy; SCN5A: Long QT Syndrome; SCN2A: Early onset DEE	SCN9A: Small-Fiber Neuropathy; SCN8A: NNDwoE; SCN10A: Neuropathies; SCN1A: FHM3; SCN8A: Benign Epilepsy
SCN4A: Relaxation Impairment Disorders	SCN3A: DEE; SCN8A: DEE; SCN9A: Paroxysmal Pain Disorders; SCN2A: Early onset DEE; SCN1A: Early onset DEE	SCN10A: Neuropathies; SCN2A: Autism; SCN8A: NNDwoE; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN5A: Brugada Syndrome	SCN1A: Dravet Syndrome; SCN2A: Autism; SCN8A: NNDwoE; SCN11A: Neuropathies; SCN1A: GEFS+	SCN3A: DEE; SCN4A: Relaxation Impairment Disorders; SCN1A: Early onset DEE; SCN8A: Benign Epilepsy; SCN1A: FHM3
SCN5A: Long QT Syndrome	SCN2A: Early onset DEE; SCN8A: DEE; SCN4A: Hyperkalemic PP; SCN11A: Neuropathies; SCN2A: Benign Epilepsy	SCN2A: Autism; SCN1A: FHM3; SCN8A: Benign Epilepsy; SCN10A: Neuropathies; SCN1A: Dravet Syndrome
SCN8A: Benign Epilepsy	SCN1A: FHM3; SCN1A: Early onset DEE; SCN4A: Relaxation Impairment Disorders; SCN2A: Early onset DEE; SCN9A: Paroxysmal Pain Disorders	SCN1A: GEFS+; SCN4A: Hypokalemic PP; SCN2A: Autism; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN8A: DEE	SCN2A: Early onset DEE; SCN4A: Relaxation Impairment Disorders; SCN3A: DEE; SCN5A: Long QT Syndrome; SCN9A: Paroxysmal Pain Disorders	SCN4A: Myopathy; SCN10A: Neuropathies; SCN2A: Autism; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN8A: NNDwoE	SCN2A: Autism; SCN1A: Dravet Syndrome; SCN5A: Brugada Syndrome; SCN11A: Neuropathies; SCN5A: Long QT Syndrome	SCN4A: Relaxation Impairment Disorders; SCN9A: Paroxysmal Pain Disorders; SCN1A: Early onset DEE; SCN10A: Neuropathies; SCN1A: FHM3

SCN9A: Paroxysmal Pain Disorders	SCN4A: Relaxation Impairment Disorders; SCN1A: Early onset DEE; SCN3A: DEE; SCN2A: Early onset DEE; SCN8A: DEE	SCN4A: Hypokalemic PP; SCN2A: Autism; SCN8A: NNDwoE; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN9A: Small-Fiber Neuropathy	SCN2A: Late onset DEE; SCN1A: Early onset DEE; SCN4A: Relaxation Impairment Disorders; SCN10A: Neuropathies; SCN3A: DEE	SCN4A: Hypokalemic PP; SCN8A: NNDwoE; SCN2A: Autism; SCN5A: Brugada Syndrome; SCN1A: Dravet Syndrome
SCN10A: Neuropathies	SCN9A: Small-Fiber Neuropathy; SCN2A: Late onset DEE; SCN1A: GEFS+; SCN9A: Paroxysmal Pain Disorders; SCN11A: Neuropathies	SCN4A: Hyperkalemic PP; SCN2A: Autism; SCN5A: Brugada Syndrome; SCN8A: NNDwoE; SCN1A: Dravet Syndrome
SCN11A: Neuropathies	SCN1A: GEFS+; SCN2A: Benign Epilepsy; SCN5A: Long QT Syndrome; SCN2A: Late onset DEE; SCN2A: Early onset DEE	SCN10A: Neuropathies; SCN1A: Early onset DEE; SCN1A: Dravet Syndrome; SCN8A: Benign Epilepsy; SCN1A: FHM3

6.2 Gene variant effects across sodium channelopathies predict function and guide precision therapy

Andreas Brunklaus[+], Tony Feng[+], **Tobias Brünger[+]**, Eduardo Perez-Palma, Henrike Heyne, Emma Matthews, Christopher Semsarian, Joseph D Symonds, Sameer M Zuberi, Dennis Lal, Stephanie Schorge. **Brain**. DOI:10.1093/brain/awac006. (published)

[+]These authors contributed equally to the work

https://doi.org/10.1093/brain/awac006 BRAIN 2022: 145; 4275–4286 | 4275

BRAIN
ORIGINAL ARTICLE

Gene variant effects across sodium channelopathies predict function and guide precision therapy

Andreas Brunklaus,[1,2,†] Tony Feng,[1,2,†] Tobias Brünger,[3,†] Eduardo Perez-Palma,[4] Henrike Heyne,[5,6,7] Emma Matthews,[8,9] Christopher Semsarian,[10,11,12] Joseph D. Symonds,[1,2] Sameer M. Zuberi,[1,2] Dennis Lal[13,14] and Stephanie Schorge[15]

[†]These authors contributed equally to this work.

See Mantegazza and Cestèle (https://doi.org/10.1093/brain/awac397) for a scientific commentary on this article.

Pathogenic variants in the voltage-gated sodium channel gene family lead to early onset epilepsies, neurodevelopmental disorders, skeletal muscle channelopathies, peripheral neuropathies and cardiac arrhythmias. Disease-associated variants have diverse functional effects ranging from complete loss-of-function to marked gain-of-function. Therapeutic strategy is likely to depend on functional effect. Experimental studies offer important insights into channel function but are resource intensive and only performed in a minority of cases.

Given the evolutionarily conserved nature of the sodium channel genes, we investigated whether similarities in biophysical properties between different voltage-gated sodium channels can predict function and inform precision treatment across sodium channelopathies. We performed a systematic literature search identifying functionally assessed variants in any of the nine voltage-gated sodium channel genes until 28 April 2021. We included missense variants that had been electrophysiologically characterized in mammalian cells in whole-cell patch-clamp recordings. We performed an alignment of linear protein sequences of all sodium channel genes and correlated variants by their overall functional effect on biophysical properties.

Of 951 identified records, 437 sodium channel-variants met our inclusion criteria and were reviewed for functional properties. Of these, 141 variants were epilepsy-associated (*SCN1/2/3/8A*), 79 had a neuromuscular phenotype (*SCN4/9/10/11A*), 149 were associated with a cardiac phenotype (*SCN5/10A*) and 68 (16%) were considered benign. We detected 38 missense variant pairs with an identical disease-associated variant in a different sodium channel gene. Thirty-five out of 38 of those pairs resulted in similar functional consequences, indicating up to 92% biophysical agreement between corresponding sodium channel variants (odds ratio = 11.3; 95% confidence interval = 2.8 to 66.9; $P < 0.001$). Pathogenic missense variants were clustered in specific functional domains, whereas population variants were significantly more frequent across non-conserved domains (odds ratio = 18.6; 95% confidence interval = 10.9–34.4; $P < 0.001$). Pore-loop regions were frequently associated with loss-of-function variants, whereas inactivation sites were associated with gain-of-function (odds ratio = 42.1, 95% confidence interval = 14.5–122.4; $P < 0.001$), whilst variants occurring in voltage-sensing regions comprised a range of gain- and loss-of-function effects.

Our findings suggest that biophysical characterisation of variants in one *SCN*-gene can predict channel function across different *SCN*-genes where experimental data are not available. The collected data represent the first gain- versus loss-of-function topological map of SCN proteins indicating shared patterns of biophysical effects aiding variant analysis and guiding precision therapy. We integrated our findings into a free online webtool to facilitate functional sodium channel gene variant interpretation (http://SCN-viewer.broadinstitute.org).

Received July 5, 2021. Revised November 27, 2021. Accepted December 10, 2021. Advance access publication January 17, 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of the Guarantors of Brain.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

4276 | BRAIN 2022: 145; 4275–4286 A. Brunklaus et al.

1 The Paediatric Neurosciences Research Group, Royal Hospital for Children, Glasgow, UK
2 Institute of Health and Wellbeing, University of Glasgow, Glasgow, UK
3 Cologne Center for Genomics, University of Cologne, Cologne, Germany
4 Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana, Universidad del Desarrollo, Santiago, Chile
5 Genomic and Personalized Medicine, Digital Health Center, Hasso Plattner Institute, Potsdam, Germany
6 Hasso Plattner Institute, Mount Sinai School of Medicine, New York, NY, USA
7 Institute for Molecular Medicine Finland: FIMM, Helsinki, Finland
8 Atkinson Morley Neuromuscular Centre, St George's University Hospitals NHS Foundation Trust, London, UK
9 Molecular and Clinical Sciences Research Institute, St George's University of London, London, UK
10 Agnes Ginges Centre for Molecular Cardiology at Centenary Institute, The University of Sydney, Sydney, Australia
11 Sydney Medical School Faculty of Medicine and Health, The University of Sydney, Sydney, Australia
12 Department of Cardiology, Royal Prince Alfred Hospital, Sydney, Australia
13 Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, USA
14 Stanley Center for Psychiatric Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
15 Department of Neuroscience, Physiology and Pharmacology, UCL, London, UK

Correspondence to: Dr Andreas Brunklaus, MD
Fraser of Allander Neurosciences Unit
Office Block, Ground Floor, Zone 2
Royal Hospital for Children
1345 Govan Road
Glasgow G51 4TF, UK
E-mail: andreas.brunklaus@glasgow.ac.uk

Correspondence may also be addressed to: Professor Stephanie Schorge, PhD
Department of Neuroscience
Physiology and Pharmacology
UCL, London WC1E 6BT, UK
E-mail: s.schorge@ucl.ac.uk

Keywords: SCN1A; SCN2A; SCN4A; SCN5A; SCN8A

Introduction

Voltage-gated sodium channel genes (*SCN1–11A*) encode a homologous family of nine functionally expressed sodium channels (SCN) from Nav1.1 to Nav1.9.[1,2] They play a key role in initiating action potentials[3] and are extensively distributed throughout the nervous system. Variants in *SCN*-genes are associated with early onset epilepsies, neurodevelopmental disorders, skeletal muscle channelopathies, peripheral neuropathies and cardiac conduction defects.[4,5] All nine *SCN*-genes share common evolutionary origins and a conserved basic structure, consisting of four homologous domains (DI–IV), each containing six transmembrane segments (S1–6) with up to 85% amino acid sequence similarity between them.[6,7] With the emergence of modern sequencing techniques facilitating genetic diagnosis, *SCN*-related disorders are promising candidates for precision therapies.[8] However, predicting the impact of a variant on channel kinetics without prior functional characterization is challenging. Variants occurring within the same gene show remarkable phenotype variability depending on their location and effect on biophysical properties, while variants occurring in different *SCN* genes may result in similar phenotypes. Such genetic and clinical complexity hinders the establishment of genotype-phenotype correlations.

Genetic variants in *SCN1/2/3/8A* are responsible for a significant proportion of monogenic epilepsies and neurodevelopmental disorders. Loss-of-function (LoF) variants in *SCN1A* manifest variable phenotypes, ranging from milder presentations such as genetic epilepsy with febrile seizures plus to the severe developmental and epileptic encephalopathy Dravet syndrome, whereas gain-of-function (GoF) variants are associated with familial hemiplegic migraine type 3 (FHM3).[9–14] Variants in *SCN2/3/8A* are clinically heterogenous, causing different forms of epilepsy ranging from self-limited infantile epilepsy to developmental and epileptic encephalopathies (DEEs) including early infantile DEE (Ohtahara syndrome) and infantile spasms syndrome.[15–17] Patients with *SCN5A* variants showing GoF manifest long-QT syndrome (LQT3), while LoF variants cause Brugada syndrome.[18,19] *SCN4A* variants are responsible for a significant proportion of skeletal muscle channelopathies,[20,21] while *SCN9/10/11A* are primarily associated with peripheral neuropathies; both are predominantly caused by GoF variants and present with variable severity.[22–24] Whilst the majority of *SCN*-related disorders are dominant conditions, newer *SCN4A*-linked phenotypes, including congenital myasthenia and congenital myopathy, have been identified to be caused by LoF variants inherited in a recessive manner.[25,26]

Sodium channelopathies display varying treatment responsiveness depending on the underlying functional effect. While many patients with *SCN2/3/8A*-related epilepsy remain treatment resistant, those caused by GoF variants, tend to respond to sodium channel blockers (SCBs), whereas those with LoF variants do not.[27–30] In contrast, patients with Dravet syndrome due to LoF *SCN1A*-variants worsen with SCBs.[31,32] Similarly, SCBs suppress pathological currents in patients with LQT3 caused by GoF *SCN5A* variants,[33]

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

whereas patients with Brugada syndrome caused by LoF are often treatment-resistant, relying on device implantation for arrhythmia control.[34,35] As functional knowledge is a key determinant for optimal management, variant interpretation without requiring resource-intensive electrophysiological studies represents an unmet clinical need.

To improve our understanding of SCN-related disorders, our objective was to investigate relationships between variant type, location and biophysical properties across all SCN-subtypes. Applying evidence from functional studies we aimed to build a framework that informs clinical practice and guides precision therapy.

Materials and methods

Search strategy and selection criteria

In accordance with PRISMA guidance, we systematically searched PubMed to 28 April 2021 to identify studies published in English describing functional characteristics of missense variants using the terms 'clamp AND SCN1A'. The same search was applied for SCN2-11A. To narrow search criteria for SCN5A, we added 'Brugada' or 'QT'. In addition, we manually searched sodium channel mutation databases and bibliographies that were found through our systematic PubMed search. Any duplicate citations were removed, and all remaining studies were screened for relevance (Supplementary Fig. 1). Only missense variants electrophysiologically characterized by whole-cell patch clamp experiments were assessed for eligibility. To improve comparison of biophysical properties, only variants characterized in mammalian cells were included. We excluded variants if they had no evidence of a recognized human disease phenotype. We did not double count identical variants but did consider patch clamp data from different sources if available. Data were independently reviewed by three researchers (A.B., T.F. and S.S.).

Variant analysis

Variants were categorized as either GoF, LoF, mixed-function (mixed) or similar-to-wild-type function depending on their effect on biophysical properties (Supplementary material). We defined any biophysical change entailing an increase in the Na+ permeability as GoF, and the opposite for LoF. In some cases, variants demonstrated a paradoxical effect on channel properties, i.e. decreased peak current and increased persistent current. Where one effect was not clearly dominant, they were classified as 'mixed'.[36] Variants that exhibited wild-type-like function or lacked pronounced impact on channel function were classified as similar-to-wild-type and assumed benign. To detect analogous missense variants amongst SCN1-11A genes, we performed an alignment of linear protein sequences using the 'PER viewer' (pathogenic variant enriched regions) across genes and gene families (http://per.broadinstitute.org).[37] Variants detected in the same alignment index position were investigated for similarities in their overall functional effects. Variants occurring in positions where the reference amino acid was different across SCN1-11A in the PER alignment were classified as 'not conserved', as this suggests that they were not conserved across evolution and thus less likely to be functionally significant.

Population variants were collected for all SCN1A-11A genes where at least one pathogenic variant was previously identified from the Genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org), which provides access to germline variants from >140 000 exomes as well as 15 000 genomes from the general population. Assumed benign variants, which showed a wild-type like behaviour were added to the population variant cohort from gnomAD. Non-conserved variants that were also present in gnomAD were assumed benign. The term benign implies that these variants occur frequently without disease; however, they might still be associated with functional change.

In-silico prediction of functional variant effects

Heyne *et al.*[38] recently developed an *in-silico* prediction tool ('*funNCion*'; http://funNCion.broadinstitute.org) to estimate the functional consequences of voltage-gated sodium and calcium channels. The tool does not consider evidence from biophysical experiments but infers channel function from a large dataset of clinical disease phenotypes. We applied the '*funNCion*' tool on our cohort of functionally characterized SCN variants to establish the tool's accuracy and clinical utility compared with actual biophysical readouts.

Data analysis

A two-tailed Fisher's Exact test with Bonferroni correction was performed to assess the burden of pathogenic variants across different regions of the SCN protein to compare pathogenic LoF and GoF variants and to determine categorical differences in functional effect and differences in phenotype or variant distribution across related SCN-subtypes. Similarly, Fisher's Exact test was performed to test whether missense variant pairs with identical disease-associated variants across different SCN-genes more often have similar functional consequences than expected from random sampling. Significance for all tests was at the 5% level and analysis was performed using R version 4.0.3.

Data availability

All data used in this study are available in the manuscript, the Supplementary material and via the free online webtool http://SCN-viewer.broadinstitute.org.

Results

Literature search

From our systematic PubMed search, we found 951 records, and following elimination of 127 duplicate citations, screened 824 titles and abstracts. A total of 535 records were excluded, as they did not include a patch-clamp experiment on a missense *SCN* variant, and 289 records including 569 variants were subsequently assessed by full-text analysis. After additional manual searching and exclusion due to ineligibility, we identified 437 missense variants in the literature that were functionally characterized by whole-cell patch-clamp experiments. Of these, 369 (84%) variants were assessed as pathogenic, including 141 associated with epilepsy (*SCN1/2/3/8A*), 79 with a neuromuscular phenotype (*SCN4/9/10/11A*) and 149 with a cardiac phenotype (*SCN5/10A*). Sixty-eight (16%) variants were considered benign, as they either had properties similar to wild-type or were both not conserved and present in gnomAD (Supplementary Fig. 1 and Supplementary Table 1).

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universiteits- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

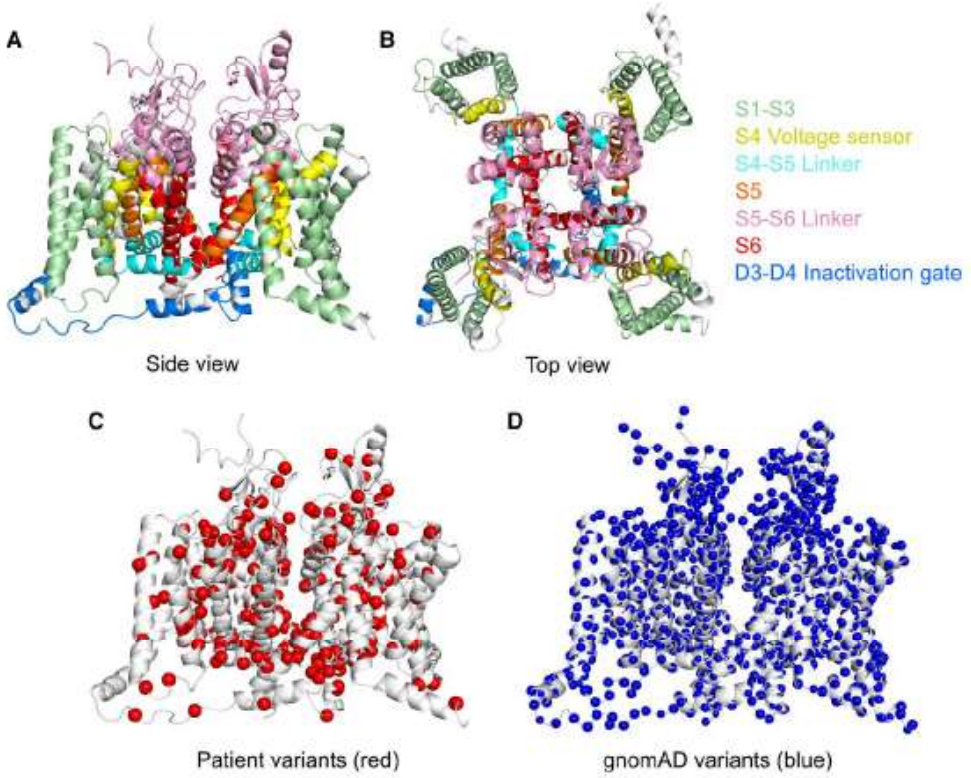


Figure 1 SCN protein structure and position of disease-causing missense versus gnomAD variants. (A) and (B) SCN protein structure from side and top view. (C) Patient variants shown in red. (D) GnomAD variants shown in blue.

Similarities in variant function and distribution across related sodium channels

To illustrate the distribution of missense variants and similarities in functional consequences across related sodium channels, we plotted the position of 369 *SCN1-11A* pathogenic variants according to their corresponding *SCN1A* alignment index position and compared their position to the distribution of 3454 gnomAD variants (Fig. 1). Enrichment analysis comparing pathogenic versus gnomAD variants demonstrated that pathogenic missense variants are clustered in specific functional domains, whereas gnomAD variants are significantly more frequent across functionally less important and not conserved cytoplasmic domains [odds ratio (OR) = 18.6; 95% confidence interval (CI) = 10.9–34.4; $P < 0.001$; Fig. 2].

The majority of pathogenic variants were located in the homologous domains D1–4 (Fig. 3). Pathogenic variants distributed across all four S5–6 pore-loop regions appeared to be predominantly LoF (91%, 58/64), whereas variants occurring in voltage-sensing regions (including S3–4, S4 and S4–5) comprised a range of GoF (47%, 51/107), LoF (38%, 41/107) and mixed function (15%, 16/107) effects. In the fast inactivation gate (DIII–IV), 83% of variants were GoF (24/29), all occurring throughout the first half of the intracellular linker. Other sites implicated in inactivation gating, including the intracellular S4–5 regions in DIII and DIV as well as DIVS6, shared this pattern of harbouring predominantly GoF variants (69%, 34/49).[39,40] Overall, pore-loop regions were frequently associated with LoF variants, whereas inactivation sites were associated with GoF (OR = 42.1, 95% CI = 14.5–122.4; $P < 0.001$). The C-terminus displayed a range of GoF (46%, 12/26), LoF (27%, 7/26) and mixed effects (27%, 7/26). Here, GoF and mixed variants appeared to cluster in the proximal region, whereas a minority of LoF variants were found distally. Overall, very few variants occurred in cytoplasmic regions (N-terminus, DI–II and DII–III linkers). A 3D illustration comparing GoF versus LoF locations across the SCN protein is detailed in Fig. 4.

Benign variants displayed a different distribution, comprising 40% of all variants identified in the N-terminus (8/20—mainly located in the initial segment), 56% in the large DI–II intracellular linker (10/18), 57% in the large DII–III intracellular linker (8/14) and 26% in the C-terminus (9/35—mainly located in the distal segment). In contrast, the fast inactivation gate was free of benign variants, and very few were found in other sites implicated in inactivation gating, including S4–5 of DIII and DIV, and DIVS6.

Corresponding variants in different SCN-genes have similar function

Among all functionally characterized *SCN1-11A* variants, we identified 38 random pairs with a corresponding analogous identical disease-associated variant in a different *SCN*-gene, including six previously reported pairs[41] (Supplementary Table 2 and Fig. 3). The missense variants in each of these pairs have similar functional consequences in 35 out of the 38 pairs (92%) regardless of the type

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

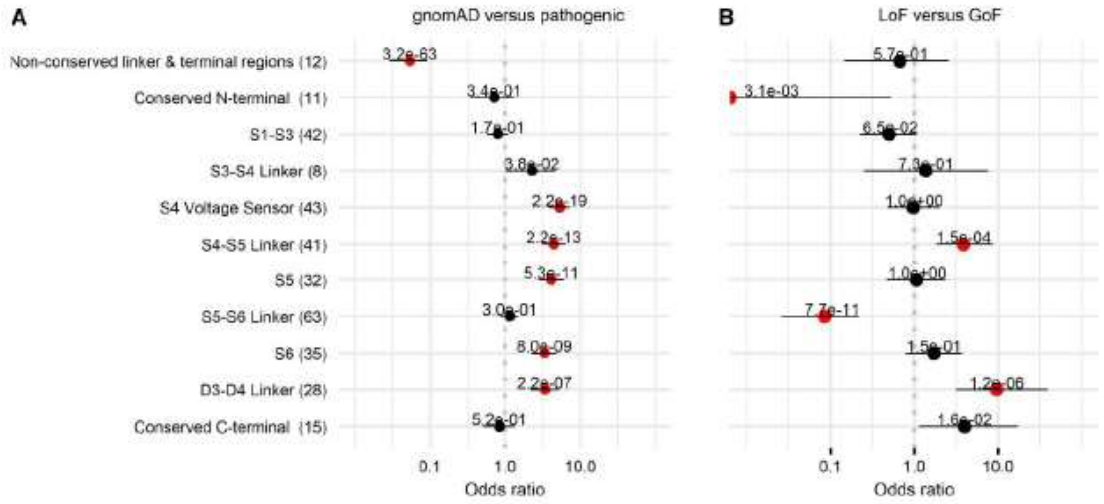


Figure 2 Enrichment analysis of gnomAD versus pathogenic variants and LoF versus GoF variants. Enrichment analysis of (A) gnomAD versus pathogenic and (B) LoF versus GoF variants across different protein parts including four homologous domains (D1–D4), each consisting of six transmembrane segments (S1–S6) and large cytoplasmic loops.

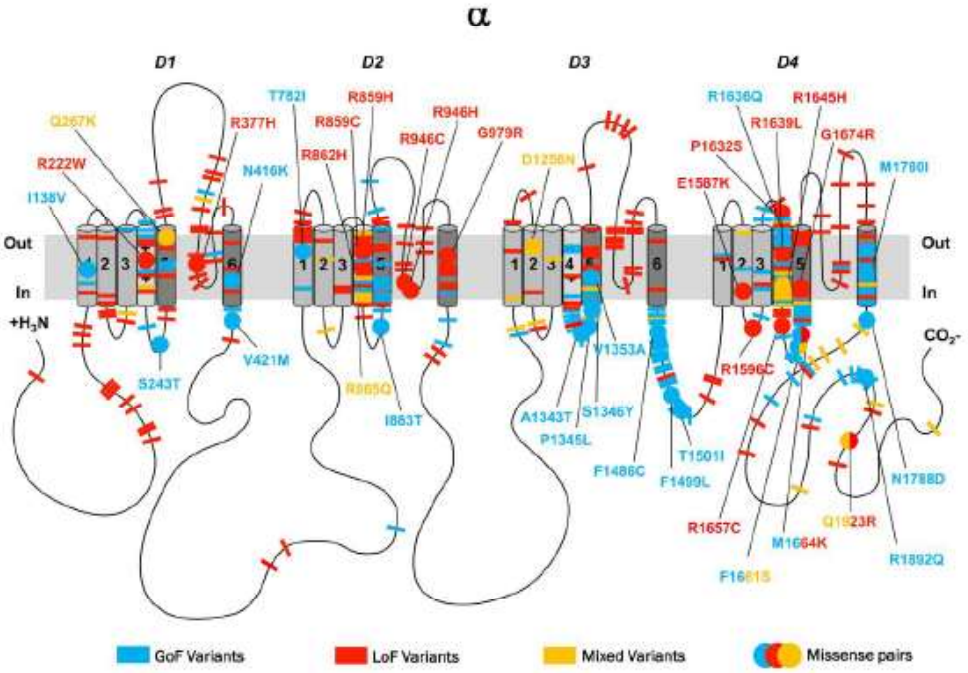


Figure 3 2D representation of pathogenic *SCN* variants with functional effects. 2D representation of the SCN protein. The alpha subunit consists of four homologous domains (D1–4) each formed of six transmembrane segments (S1–S6). Segment 4 represents the voltage sensor and S5–6 the pore region. Individual missense variants are displayed as different coloured bars. Blue denotes GoF, red LoF and yellow mixed function. Analogue missense pairs are displayed as circles with amino acid details.

of voltage-gated sodium channel affected (OR = 11.3; 95% CI = 2.8–66.9; P < 0.001). Many of these pairs were found at conserved channel locations, including three LoF pairs between *SCN1/2/5A* at pore-loop regions, seven GoF pairs between *SCN1-9A* at sites implicated in channel inactivation and eight pairs between SCN1/4/5/8A at S4 showing a mixture of GoF, LoF and mixed-function effects (Fig. 3). The only three pairs with divergent function included the *SCN1A* F1661S/*SCN4A* F1473S[42–44] variant pair and the *SCN1A* M1664K/*SCN9A* M1627K[45–48] variant pair, both located in the D4 S4–5 linker region, as well as the *SCN1A* Q1923R/*SCN5A* Q1909R[49–51] variant

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

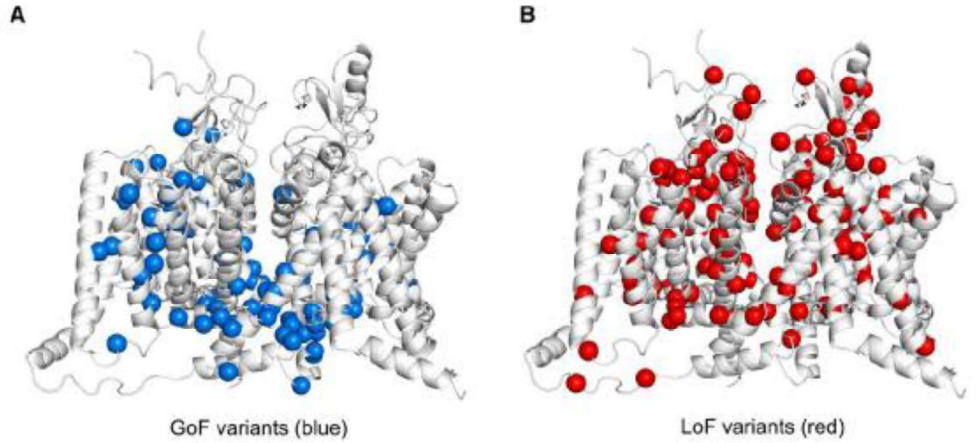


Figure 4 3D illustration comparing GoF with LoF locations across the SCN protein. (A) GoF variants are illustrated in blue. **(B)** LoF variants are illustrated in red.

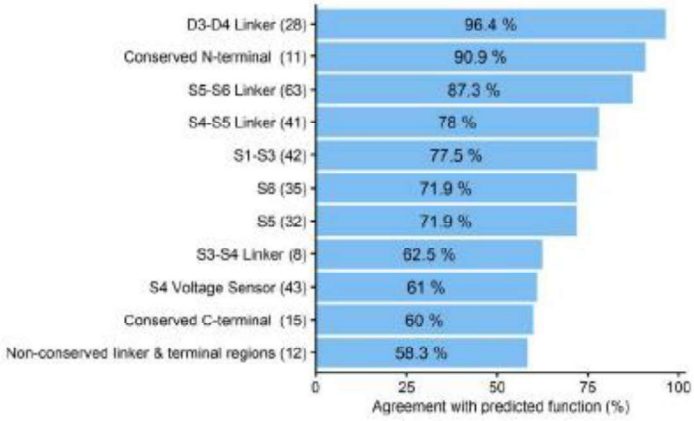


Figure 5 ***In-silico*** **prediction versus reported biophysical** ***SCN*** **variant effects. Prediction agreement [as a percentage (%)] detailed according to different protein parts including four homologous domains (D1–D4), each consisting of six transmembrane segments (S1–S6) and large cytoplasmic loops.** ***In silico*** **prediction was performed according to 'funNCion' (http://funNCion.broadinstitute.org).**

pair located in the distal C-terminus. The *SCN4A* variant was identified in a patient with paramyotonia congenita (PMC) due to GoF, the *SCN9A* variant in a patient with paroxysmal extreme pain disorder (PEPD), equally due to GoF, and the *SCN5A* variant in a patient with sudden infant death syndrome, associated with mixed function; the three corresponding *SCN1A* variants were identified in patients with genetic epilepsy with febrile seizures plus/Dravet syndrome and found to be LoF/mixed function. The *SCN1A* variants led to impaired channel trafficking and reduced cell surface expression, resulting in a reduction of peak current, not allowing for detailed biophysics to be recorded. In contrast, the trafficking of the *SCN4A*, *SCN5A* and *SCN9A* variants did not appear to be affected.

Comparison of reported biophysical effects versus predicted functional outcomes

We applied the recently-developed *in-silico* prediction tool '*funNCion*' to predict GoF versus LoF effects in the 369 biophysically characterized variants to evaluate the accuracy and usefulness of this type of tool. Compared with the gold standard whole-cell patch clamp experiment results, the *in-silico* tool achieved 78.5% agreement in the prediction of GoF properties and 75.0% agreement in the prediction of LoF properties. Agreement differed depending on location within the channel, with certain regions such as inactivation and pore loop sites achieving better agreement (78–96%) compared with others including the S4 region (<62%, Fig. 5).

Detailed SCN1-11A variant analysis

Each of the different *SCN* subtypes (*SCN1–11A*) presented with a specific distribution pattern of pathogenic versus benign variants according to channel function and location (Fig. 6). Supplementary Table 3 lists the channel specific clinical phenotypes and associated function.

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

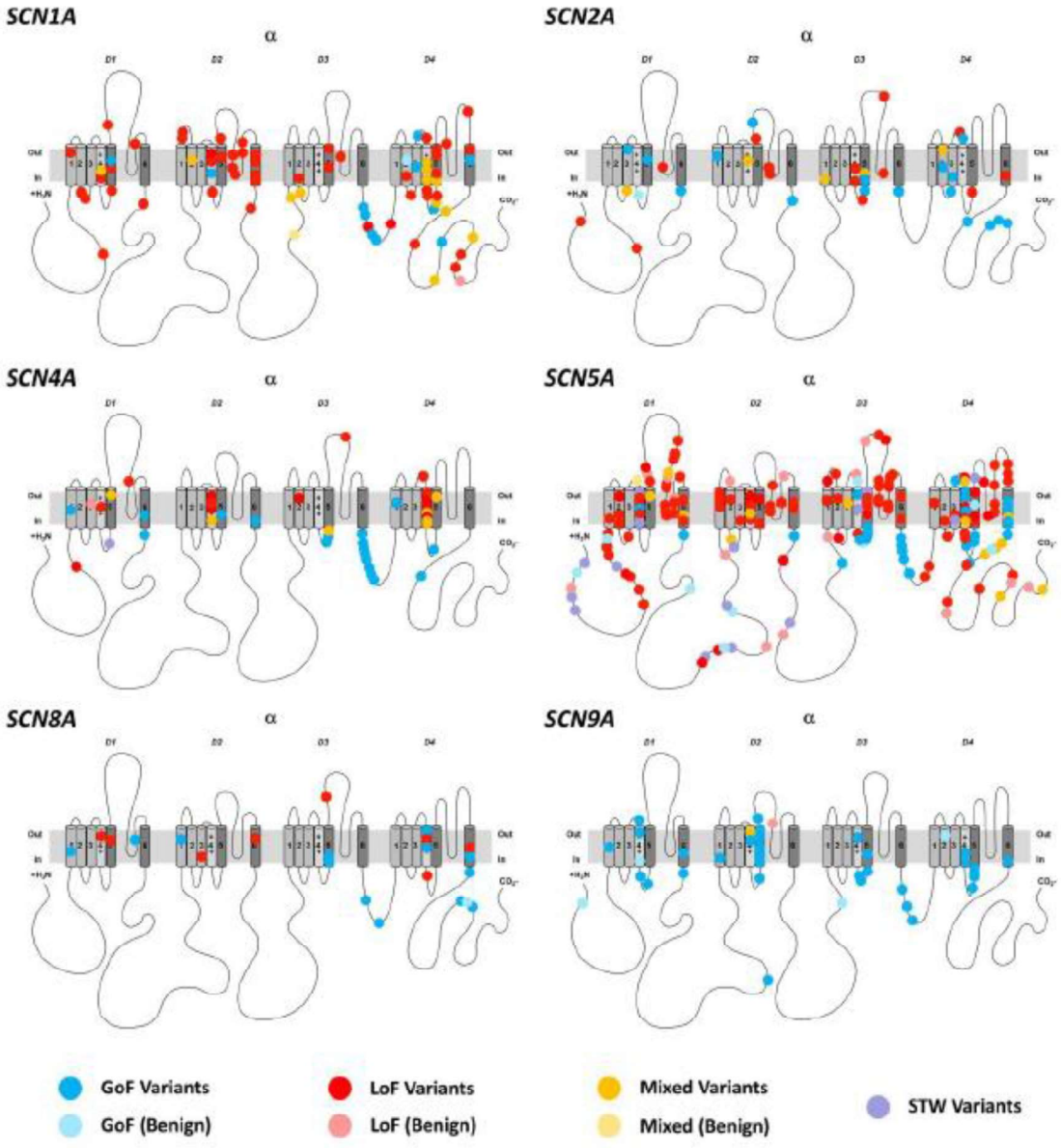


Figure 6 2D representation of *SCN* variants with functional effects according to single sodium channels. 2D representation of different *SCN* proteins. The alpha subunit consists of four homologous domains (D1–4) each formed of six transmembrane segments (S1–S6). Segment 4 represents the voltage sensor and segments S5–6 the pore region. Individual missense variants are displayed as different coloured circles. Blue denotes GoF, red LoF, yellow mixed function and purple similar-to-wild-type (STW). Benign variants are illustrated in pale colours.

Discussion

Comparing the distribution of disease-associated missense variants across all nine *SCN*-subtypes, we observed striking similarities in altered biophysical channel properties induced by missense variants at analogous position across *SCN* proteins. Almost all identified variant pairs in different *SCN* genes exhibited similar biophysical properties regardless of the *SCN*-subtype affected. Functional data of variants at analogous positions in *SCN* genes can predict variant effects in related sodium channel genes and may inform genotype-guided precision therapies in patients with neurological and cardiac sodium channelopathies.

Voltage-gated sodium channels contain a central pore composed of S5 and S6 segments from all four domains that line the inner cavity and form the intracellular exit. Intervening S5–6 pore-loops line the extracellular end of the pore, forming a large ion-selective filter.[52]

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universiteits- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

In keeping with previous work, we observed that variants occurring in the S5-6 pore-loop predominantly caused LoF effects.[5,53] This region displayed clustering of epilepsy-associated LoF SCN1A variants, whereas very few epilepsy-associated SCN2/8A variants occurred here. The relationship between variant distribution and functional consequences was also observed in SCN5A. Similar to reported case series,[54] the vast majority of Brugada syndrome-associated variants displayed LoF effects and clustered in S5–6 pore-loops. In contrast, neither LQT3-associated variants that predominantly caused GoF nor any GoF variants across other SCN-subtypes clustered in this region. This showed that, across different SCNs, variants occurring in S5–6 pore-loops often lead to LoF, causing detrimental effects on channel kinetics.

Sites implicated in inactivation gating harboured predominantly GoF variants. The short DIII–IV linker is responsible for fast inactivation. This loop folds into a hinged-lid structure that blocks the intracellular end of the pore during sustained depolarization and contains residues (IFM motif) that form a latch, holding the inactivation gate shut.[52] Interference within this region leads to impaired inactivation and hyperexcitability that is in keeping with the underlying mechanism of SCN-related disorders.[5] Respectively, SCN1A and SCN2/8A variants distributed across the DIII–IV linker were associated with FHM3 and epilepsy, while SCN4A variants were associated with myotonia. Similarly, a significant proportion of LQT3-associated SCN5A variants and painful neuropathy-associated SCN9A variants occurred across all regions implicated in inactivation gating. This suggested that variants occurring in inactivation sites frequently present GoF across different SCNs.

Voltage-sensing regions in each domain are composed of S4 and adjacent linkers. S4 serves as the voltage-sensor, containing a dense composition of highly-conserved arginine residues.[7] In response to depolarization, this segment moves across the membrane and facilitates channel activation.[52] We observed a mixture of biophysical effects in S4, constituting a cluster of painful neuropathy-associated SCN9/11A, hypokalaemic periodic paralysis-associated SCN4A and LQT3-associated SCN5A variants with a moderate distribution of epilepsy-associated SCN1/2/3/8A variants. Rather than a regional effect (as in the aforementioned S5–6 linker region), functional impact in S4 appears to be determined by individual variant changes that manifest a spectrum of disorders across all SCN-subtypes. The five conserved arginine residues R0–R4 serve as example for this observation: whilst pathogenic variants in R0 are associated with increased function, there appears to be a steady shift from nearly complete LoF in variants of R1/R2 to LoF/mixed in R3 and mainly mixed effects in R4, the deeper the progression into the S4 (Supplementary Fig. 2).

In contrast, the large non-conserved cytoplasmic regions and distal N- and C-terminal cytoplasmic regions are devoid of pathogenic variants across all SCN-subtypes.

Comparison of corresponding variants aids clinical prediction

Our alignment of all nine channel subtypes allowed a comparison of SCN-variants by their corresponding index position, revealing similar functional consequences occurring in specific regions. The clinical manifestation of the observed biophysical change depends on neuron type and neuronal network distribution. For example, while all variant pairs observed in the S5–6 pore-loop regions displayed the same LoF functional consequences, a LoF in SCN1A leads to epilepsy due to impaired expression of inhibitory interneurons,[10] whilst the corresponding LoF SCN2A and SCN5A variants result in autism spectrum disorder and Brugada syndrome, respectively, due to high expression of these genes in excitatory brain-expressed SCN2A neurons and cardiac-expressed SCN5A channels.[17,18] The same pattern was observed in sites of inactivation, where GoF variants in SCN1A cause FHM3, whilst analogous GoF variants in SCN2/8A cause epilepsy. Similarly, GoF variants at inactivation sites occurring in SCN4A cause myotonia due to hyperexcitability of the sarcolemma,[20] while corresponding variants in SCN9/10/11A that display the same GoF properties cause painful peripheral neuropathy owing to expression in peripheral neurons.[22] Importantly, these similarities in functional consequences across related SCNs apply in particular to regions that are conserved across evolution. For example, all variant pairs occurring in the S4 segment involved substitution of arginine residues which are highly-conserved across all related SCNs.[55]

In the three cases where there were discrepancies between how variants affected different channels (Supplementary Table 2), we noted that all three were cases where a variant produced LoF/mixed function (reduced or no peak current) of SCN1A but GoF or mixed effects in other channels. This discrepancy may be related to impaired channel trafficking by which SCN1A may be particularly affected. Where trafficking is suspected, pharmacological rescue is a possibility, first suggested when mexiletine partially-rescued trafficking for a mutation in SCN5A.[56] Since then, trafficking mutations have been shown for SCN1A,[51] and several of these can be rescued pharmacologically with anti-epileptic drugs such as phenytoin[43] and other chaperonins.[12,57] Indeed, SCN1A is so sensitive to trafficking that a mutation initially identified as LoF in cell lines due to trafficking becomes GoF simply by expressing in neurons.[58] Thus, we propose that an additional use of our data may arise when classifying an SCN1A variant associated with LoF in functional studies but linked to a variant known to produce GoF or mixed effects in other channels when heterologously expressed. In this exceptional case, our data may support cautious interpretation of the functional data (which in almost every other case should be the gold standard). Our data suggest that in these cases it is particularly important to consider trafficking, as SCN1A may be more sensitive to trafficking deficits when expressed in non-native cells such as HEK cells. Where complete LoF may be seen in HEK cells, it is possible that functional channels—potentially with aberrant gating—are produced in native cells. As new genetic therapies may begin to distinguish between complete LoF and aberrant function (such as gating pore leaks), it will be particularly important to rule out the possibility of a mutant allele having gating changes or trafficking issues that could alter neuronal health if expression is increased. As treatments targeting trafficking become clinically available, these mutations may be prioritized for interrogation for these treatments: if in functional studies treatments that increase trafficking produce wild-type-like currents, then these mutations would represent a potentially valuable treatment via trafficking for children.

Overlapping boundaries between presumed benign and pathogenic variants

We observed that a significant proportion of SCN4/5/10A variants were not conserved or assumed benign, compared with very few SCN1/2/8A variants. This difference may be explained by the easily recognizable clinical presentation of SCN1/2/8A-related disease with difficult to treat epilepsy and severe intellectual disability.[10] In comparison, SCN4/5/10A-related disorders appear to have less noticeable clinical features, often presenting with only one

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6500259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

paroxysmal symptom with variable severity.[59–61] Furthermore, there is significant evolutionary constraint among *SCN* genes, suggesting that whilst patients with *SCN1/2/3/8A* variants present with early onset severe dominant *de novo* disease and marked reduction in fecundity, variants associated with familial SCN disease such as *SCN4/5/9/10/11A* are better tolerated and might go unnoticed.[36,62]

The distribution pattern across *SCN*-subtypes for gnomAD/benign variants clearly contrasted that of pathogenic variants. Consistent with previous work, very few benign variants were found at sites of inactivation and pore-forming regions.[63,64] A significant proportion of variants occurring at DI–II and DII–III linkers were assumed benign. Many did not share the same amino acid as their corresponding *SCN1A* position and were present in gnomAD, demonstrating how these regions are not conserved across evolution, and of lesser functional significance.[65] The distribution of variants found in reference populations correlate strongly with the functional significance of these regions, and thus aids the interpretation of variant pathogenicity across different *SCN*-subtypes (http://per.broadinstitute.org).

Clinical implications on precision treatment and its limitations

Precision medicine aims to tailor individual treatment to reverse or modify an underlying disease pathophysiology.[66,67] Knowledge of the functional impact of a *SCN*-variant can inform clinical management and therapeutic choice. Our approach will be particularly helpful for clinicians faced with early presentation in very young children with a novel mutation of unknown function, where biophysical estimation could guide treatment. Functional characterization remains the gold standard, but can lead to many months of delay, whilst our data will allow almost instantaneous information about the likely effect of a novel variant. However, we emphasize that this should not replace experimental functional analysis, which allows the actual identification of detailed effects and mechanisms. Evidence of successful precision treatment approaches have been reported across SCN-related disorders. GoF variant carriers are more likely to respond to SCBs in *SCN2/8A*-related epilepsies,[27,68] *SCN4A*-related myotonias[69] and *SCN5A*-mediated arrhythmias/LQT3.[33,70] In contrast, use of SCBs in LoF cases is contraindicated. Inadvertently selecting incorrect treatment without knowledge of function can be detrimental to clinical symptoms.[27,32] However, success with precision treatment is not guaranteed. A recent survey of precision medicine in genetic epilepsies demonstrated that many individuals with genetic epilepsy continue to have symptoms, with >50% seizure reduction observed in just 30% of patients.[66] Limitations to precision medicine include the genetic background between patients, variation in gene expression, epigenetics and environmental factors. *In vitro* functional studies are not able to account for these modifying factors which limits their clinical utility.

We identified a number of SCN gene-specific clinical examples where knowledge of the underlying functional properties assists patient management. Early in disease presentation when phenotypes have not fully evolved, functional inference may guide early management and help predict the ultimate phenotype. In *SCN1A*-related disorders, this is particularly useful in those presenting <4 months of age as these can either be LoF, associated with Dravet syndrome, or GoF/mixed function, associated with early onset DEE.[71,72] Our approach aids interpretation of novel *SCN1A* variants in FHM phenotypes, since these would be expected to be GoF. In *SCN2A*-related disorders it guides therapy for individuals with early onset DEE presenting between 2–4 months, since either GoF or LoF variant carriers might manifest around this age. Although many patients remain pharmacoresistant, GoF variant carriers tend to experience a reduction in seizure frequency with SCBs, in particular phenytoin in high doses, whereas SCBs lead to seizure aggravation in LoF variant carriers.[27] Given that the number of reported variants and functional studies in *SCN3A* are very limited, our approach can be used to help determine pathogenicity of novel variants in this gene. In *SCN8A*-related epilepsies our approach guides therapy for those presenting between 9 months and 3 years of life that can harbour either GoF or LoF variants as recently shown in a large *SCN8A* cohort.[30] Whilst many patients still have pharmacoresistant epilepsy, GoF variant carriers respond significantly better to SCBs than to other anti-seizure medications.[30]

In addition to conventional SCBs, more selective agents, including those that target mutations with specific effects on channel function[73,74] and those that target specific types of sodium channels, are being developed.[75,76] As these approaches rely on variant position and functional impact on the sodium channel, our method could facilitate identification of variants suitable for these treatments.

New disease modifying treatments are being developed for *SCN1A*, *SCN2A* and *SCN8A*-related disorders targeting gene specific LoF and GoF effects.[77] Phase 1/2 clinical trials are ongoing/being developed for Dravet syndrome, aiming to increase *SCN1A* expression in LoF disease. This approach would likely be detrimental in GoF disease.[77] Our method allows estimation of channel function that will be useful for variant interpretation in specific GoF- or LoF-directed disease modifying therapies. However, any decision of which therapeutic approach should be taken will be informed by multiple factors including functional estimation as well as clinical presentation.

In light of these developments, being able to map specific GoF and LoF regions across sodium channel paralogues allows clinicians to make the best use of extensive functional data to help inform both drug choice in the individual patient as well as allowing targeted drug compound development across different channels.[76,78,79] Specific examples are discussed below, where LOF in *SCN1A* may be prioritized for new treatments improving trafficking, or where homology suggests a LOF might also induce toxic gating pore leaks, potentially precluding treatments aimed at increasing expression of both alleles. It has recently been shown how identifying key functional regions in one channel (Nav1.7/*SCN9A*) can be used to accelerate the design of next generation Nav modulators across other channels.[80] Likewise, the increasing recognition that loss of S4 arginines may have a pathogenic mechanism of introducing an aberrant gating pore leak in multiple channels including *SCN4A*-related muscle channelopathies[78] and *SCN2A*-related epilepsies[81] means that drugs which specifically block gating pore leaks may be applicable in cases where gating pore leak is suspected and could apply across a range of channels, even where gating pore recordings are not possible. However, clinicians should consider functional variant information in the context of the clinical presentation of their patient, including seizure types, EEG signature, previous response to specific medications and medication side effects, all of which might offer important clues towards diagnosis and management.

Gold standard patch-clamp recordings in mammalian models are time and resource intensive, precluding their use in routine practice and the ability of functional variant interpretation without prior biophysical studies represents an unmet clinical need.[82] However, our finding of up to 92% functional agreement between corresponding sodium channel variants illustrates that *SCN* paralogues can aid variant interpretation and guide precision treatment.

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universiteits- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

In comparison, agreement between the existing 'funNCion' prediction technique and gold standard patch-clamp recordings ranged between 50–96% (Fig. 5). This lack in precision of the 'funNCion' tool may in part be explained as the tool does not consider biophysical evidence but infers channel function from clinical datasets. We noticed that specific protein locations could be predicted with better accuracy than others. For example, functionally homogeneous GoF areas such as the D3–4 linker inactivation gate region and LoF areas such as the S5–6 linker pore region were predicted with high confidence. In contrast, mixed gain- and loss-of-function areas such as the S4 segments that include different arginine residues appear more challenging to predict using the 'funNCion' tool.

Compared to the current gold standard, our approach is very effective and >90% accurate. We are not aware of any other techniques that would allow such immediate SCN functional prediction. Molecular dynamics can be useful for understanding specific channel properties including the movement of ions through the pore, but not for processes like inactivation and does not allow prediction of overall function.[81] AlphaFold is a promising new approach to solving static protein structures that may be combined with molecular dynamics. However, for the present it may be envisaged that collated functional data from published electrophysiological studies will be needed to affirm predictions from molecular dynamics and AlphaFold in the future. At present, these approaches are too computationally expensive for prediction of the overall functional change caused by a variant.[82]

Based on our findings we created a freely accessible 'SCN-Viewer', allowing clinicians and scientists immediate access to published biophysical data across all voltage gated sodium channels (http://SCN-viewer.broadinstitute.org). The tool enables users to identify any *SCN*-paralogues detailing both functional and clinical characteristics associated with specific SCN1-11A variants.

In practice, when a new voltage gated sodium channel missense variant is identified and pathogenicity has been established by American College of Medical Genetics criteria, we propose the following key indicators to estimate the variants functional properties (Box 1). In the absence of gold standard functional data, these tools offer clinicians and scientists valuable insights when interpreting newly identified variants. Whilst this approach is not intended to replace clinical judgment, it will inform and complement clinical decision making based on objective and quantifiable data.

There are several limitations to this study. We limited our literature search to peer reviewed studies reported in PubMed and sodium channel databases and did not include other sources. Whilst a number of reports will have been missed, this is unlikely to affect our main findings of functional similarity among different *SCN*-genes. Our simplified approach of variant categorization into gain-, loss- and mixed function often does not fully reflect the complex biophysical properties of SCNs; however, our data illustrate that there is value in this pragmatic approach of functional variant categorization.

Box 1 Approach to functional SCN variant estimation

Key indicators of variant characteristics

(i) Identical missense variant in a paralogue sodium channel has been functionally characterized (http://SCN-viewer.broadinstitute.org): up to 92% likelihood that reported findings apply to new variant.

(ii) Use of *in silico* functional prediction tool (http://funNCion.broadinstitute.org): 58–96% likelihood that prediction applies to new variant (depending on variant location)'

(iii) Similar missense variant at equivalent position in a paralogue sodium channel has been reported in a related characteristic gain- or loss-of-function *SCN* disorder (http://per.broadinstitute.org): it is likely that the new variant is in keeping with the reported phenotypes.

Conclusion

Our findings suggest that biophysical characterization of variants in one *SCN*-gene can predict channel function across different *SCN*-genes where experimental data are not available. Shared patterns of functional effects aid variant interpretation and guide precision therapy.

Funding

D.L. and E.P. received a grant from the Dravet Syndrome Foundation (grant number, 272016) to support this work. E.P. is supported by the Agencia Nacional de Investigación y Desarrollo (ANID, PAI77200124) of Chile and the FamilieSCN2A foundation 2020 Action Potential Grant. A.B. and S.Z. received a grant from Dravet syndrome UK (grant number, 16GLW00) for the Glasgow SCN1A database. C.S. is the recipient of a National Health and Medical Research Council (NHMRC) Practitioner Fellowship (#1154992).

Competing interests

A.B. has received honoraria for presenting at educational events, advisory boards and consultancy work for Biocodex, GW Pharma, Encoded Therapeutics, Stoke Therapeutics, Nutricia and Zogenix. E.P. has received honoraria for consultancy work for the Friends of Faces foundation. S.M.Z. has received honoraria for presenting at educational events, advisory boards and consultancy work for GW Pharma, Zogenix, Biocodex, Encoded Therapeutics, Stoke Therapeutics and Nutricia. D.L. has received honoraria for advisory board work for Encoded Therapeutics. No other competing interests are reported.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Goldin AL, Barchi RL, Caldwell JH, *et al*. Nomenclature of voltage-gated sodium channels. *Neuron*. 2000;28(2):365–368.
2. Yu FH, Catterall WA. Overview of the voltage-gated sodium channel family. *Genome Biol*. 2003;4(3):207.
3. Hodgkin AL, Huxley AF. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol*. 1952;117(4):500–544.
4. Spillane J, Kullmann DM, Hanna MG. Genetic neurological channelopathies: molecular genetics and clinical phenotypes. *J Neurol Neurosurg Psychiatry*. 2016;87(1):37–48.
5. Brunklaus A, Ellis R, Reavey E, Semsarian C, Zuberi SM. Genotype phenotype associations across the voltage-gated sodium channel family. *J Med Genet*. 2014;51(10):650–658.

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

6. Catterall WA, Goldin AL, Waxman SG. International union of pharmacology. XLVII. Nomenclature and structure-function relationships of voltage-gated sodium channels. *Pharmacol Rev*. 2005;57(4):397–409.
7. Yu FH, Catterall WA. The VGL-chanome: a protein superfamily specialized for electrical signaling and ionic homeostasis. *Sci STKE*. 2004;2004(253):re15.
8. Brunklaus A. Precision medicine in sodium channelopathies - Moving beyond seizure control towards disease modification. *Eur J Paediatr Neurol*. 2020;24:7.
9. Yu FH, Mantegazza M, Westenbroek RE, *et al*. Reduced sodium current in GABAergic interneurons in a mouse model of severe myoclonic epilepsy in infancy. *Nat Neurosci*. 2006;9(9):1142–1149.
10. Mantegazza M, Broccoli V. SCN1A/NaV1.1 channelopathies: Mechanisms in expression systems, animal models, and human iPSC models. *Epilepsia*. 2019;60(S3):S25–S38.
11. Cestèle S, Scalmani P, Rusconi R, Terragni B, Franceschetti S, Mantegazza M. Self-limited hyperexcitability: Functional effect of a familial hemiplegic migraine mutation of the Nav1.1 (SCN1A) Na+ channel. *J Neurosci*. 2008;28(29):7273–7283.
12. Cestèle S, Schiavon E, Rusconi R, Franceschetti S, Mantegazza M. Nonfunctional NaV1.1 familial hemiplegic migraine mutant transformed into gain of function by partial rescue of folding defects. *Proc Natl Acad Sci USA*. 2013;110(43):17546–17551.
13. Kahlig KM, Rhodes TH, Pusch M, *et al*. Divergent sodium channel defects in familial hemiplegic migraine. *Proc Natl Acad Sci USA*. 2008;105(28):9799–9804.
14. Barbieri R, Bertelli S, Pusch M, Gavazzo P. Late sodium current blocker GS967 inhibits persistent currents induced by familial hemiplegic migraine type 3 mutations of the SCN1A gene. *J Headache Pain*. 2019;20(1):107.
15. Berkovic SF, Heron SE, Giordano L, *et al*. Benign familial neonatal-infantile seizures: characterization of a new sodium channelopathy. *Ann Neurol*. 2004;55(4):550–557.
16. Nakamura K, Kato M, Osaka H, *et al*. Clinical spectrum of SCN2A mutations expanding to Ohtahara syndrome. *Neurology*. 2013;81(11):992–998.
17. Shi X, Yasumoto S, Kurahashi H, *et al*. Clinical spectrum of SCN2A mutations. *Brain Dev*. 2012;34(7):541–545.
18. Bennett PB, Yazawa K, Makita N, George AL, Jr. Molecular mechanism for an inherited cardiac arrhythmia. *Nature*. 1995;376(6542):683–685.
19. Fonseca DJ, Vaz da Silva MJ. Cardiac channelopathies: The role of sodium channel mutations. *Rev Port Cardiol*. 2018;37(2):179–199.
20. Jurkat-Rott K, Holzherr B, Fauler M, Lehmann-Horn F. Sodium channelopathies of skeletal muscle result from gain or loss of function. *Pflugers Arch*. 2010;460(2):239–248.
21. Lee SC, Kim HS, Park YE, Choi YC, Park KH, Kim DS. Clinical diversity of SCN4A-mutation-associated skeletal muscle sodium channelopathy. *J Clin Neurol*. 2009;5(4):186–191.
22. Waxman SG. Painful Na-channelopathies: an expanding universe. *Trends Mol Med*. 2013;19(7):406–409.
23. Faber CG, Lauria G, Merkies ISJ, *et al*. Gain-of-function Nav1.8 mutations in painful neuropathy. *Proc Natl Acad Sci USA*. 2012;109(47):19444–19449.
24. Huang J, Han C, Estacion M, *et al*. Gain-of-function mutations in sodium channel Na(v)1.9 in painful neuropathy. *Brain*. 2014;137(Pt 6):1627–1642.
25. Elia N, Palmio J, Castañeda MS, *et al*. Myasthenic congenital myopathy from recessive mutations at a single residue in Na(V)1.4. *Neurology*. 2019;92(13):e1405–e1415.
26. Habbout K, Poulin H, Rivier F, *et al*. A recessive Nav1.4 mutation underlies congenital myasthenic syndrome with periodic paralysis. *Neurology*. 2016;86(2):161–169.
27. Wolff M, Johannesen KM, Hedrich UBS, *et al*. Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. *Brain*. 2017;140(5):1316–1336.
28. Sanders SJ, Campbell AJ, Cottrell JR, *et al*. Progress in understanding and treating SCN2A-mediated disorders. *Trends Neurosci*. 2018;41(7):442–456.
29. Møller RS, Johannesen KM. Precision medicine: SCN8A encephalopathy treated with sodium channel blockers. *Neurotherapeutics*. 2016;13(1):190–191.
30. Johannesen KM, Liu Y, Koko M, *et al*. Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. *Brain*. 2022;145:2991–3009.
31. Brunklaus A, Ellis R, Reavey E, Forbes GH, Zuberi SM. Prognostic, clinical and demographic features in SCN1A mutation-positive Dravet syndrome. *Brain*. 2012;135(Pt 8):2329–2336.
32. Guerrini R, Dravet C, Genton P, Belmonte A, Kaminska A, Dulac O. Lamotrigine and seizure aggravation in severe myoclonic epilepsy. *Epilepsia*. 1998;39(5):508–512.
33. Mazzanti A, Maragna R, Faragli A, *et al*. Gene-specific therapy with mexiletine reduces arrhythmic events in patients with long QT syndrome type 3. *J Am Coll Cardiol*. 2016;67(9):1053–1058.
34. Kambouris NG, Nuss HB, Johns DC, Marbán E, Tomaselli GF, Balser JR. A revised view of cardiac sodium channel "blockade" in the long-QT syndrome. *J Clin Invest*. 2000;105(8):1133–1140.
35. Priori SG, Wilde AA, Horie M, *et al*. HRS/EHRA/APHRS expert consensus statement on the diagnosis and management of patients with inherited primary arrhythmia syndromes: document endorsed by HRS, EHRA, and APHRS in May 2013 and by ACCF, AHA, PACES, and AEPC in June 2013. *Heart Rhythm*. 2013;10(12):1932–1963.
36. Brunklaus A, Schorge S, Smith AD, *et al*. SCN1A variants from bench to bedside-improved clinical prediction from functional characterization. *Hum Mutat*. 2020;41(2):363–374.
37. Pérez-Palma E, May P, Iqbal S, *et al*. Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res*. 2020;30(1):62–71.
38. 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.
39. Meisler MH, Hill SF, Yu W. Sodium channelopathies in neurodevelopmental disorders. *Nat Rev Neurosci*. 2021;22(3):152–166.
40. Mantegazza M, Cestèle S, Catterall WA. Sodium channelopathies of skeletal muscle and brain. *Physiol Rev*. 2021;101(4):1633–1689.
41. Brunklaus A, Du J, Steckler F, *et al*. Biological concepts in human sodium channel epilepsies and their relevance in clinical practice. *Epilepsia*. 2020;61(3):387–399.
42. Rhodes TH, Lossin C, Vanoye CG, Wang DW, George AL Jr. Noninactivating voltage-gated sodium channels in severe myoclonic epilepsy of infancy. *Proc Natl Acad Sci USA*. 2004;101(30):11147–11152.
43. Thompson CH, Porter JC, Kahlig KM, Daniels MA, George AL Jr. Nontruncating SCN1A mutations associated with severe myoclonic epilepsy of infancy impair cell surface expression. *J Biol Chem*. 2012;287(50):42001–42008.
44. Fleischhauer R, Mitrovic N, Deymeer F, Lehmann-Horn F, Lerche H. Effects of temperature and mexiletine on the F1473S Na+ channel mutation causing paramyotonia congenita. *Pflugers Arch*. 1998;436(5):757–765.
45. Bechi G, Rusconi R, Cestèle S, Striano P, Franceschetti S, Mantegazza M. Rescuable folding defective NaV1.1 (SCN1A) mutants in epilepsy: properties, occurrence, and novel rescuing strategy with peptides targeted to the endoplasmic reticulum. *Neurobiol Dis*. 2015;75:100–114.

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

46. Fertleman CR, Baker MD, Parker KA, *et al.* SCN9A mutations in paroxysmal extreme pain disorder: allelic variants underlie distinct channel defects and phenotypes. *Neuron*. 2006;52(5):767–774.
47. Dib-Hajj SD, Estacion M, Jarecki BW, *et al.* Paroxysmal extreme pain disorder M1627K mutation in human Nav1.7 renders DRG neurons hyperexcitable. *Mol Pain*. 2008;4:37.
48. Theile JW, Jarecki BW, Piekarz AD, Cummins TR. Nav1.7 mutations associated with paroxysmal extreme pain disorder, but not erythromelalgia, enhance Navbeta4 peptide-mediated resurgent sodium currents. *J Physiol*. 2011;589(Pt 3):597–608.
49. Winkel BG, Yuan L, Olesen MS, *et al.* The role of the sodium current complex in a nonreferred nationwide cohort of sudden infant death syndrome. *Heart Rhythm*. 2015;12(6):1241–1249.
50. Abdelsayed M, Baruteau AE, Gibbs K, *et al.* Differential calcium sensitivity in NaV 1.5 mixed syndrome mutants. *J Physiol*. 2017;595(18):6165–6186.
51. Rusconi R, Scalmani P, Cassulini RR, *et al.* Modulatory proteins can rescue a trafficking defective epileptogenic Nav1.1 Na+ channel mutant. *J Neurosci*. 2007;27(41):11037–11046.
52. Catterall WA. Structure and function of voltage-gated sodium channels at atomic resolution. *Exp Physiol*. 2014;99(1):35–51.
53. Zuberi SM, Brunklaus A, Birch R, Reavey E, Duncan J, Forbes GH. Genotype-phenotype associations in SCN1A-related epilepsies. *Neurology*. 2011;76(7):594–600.
54. Tarradas A, Selga E, Beltran-Alvarez P, *et al.* A novel missense mutation, I890T, in the pore region of cardiac sodium channel causes Brugada syndrome. *PLoS One*. 2013;8(1):e53220.
55. Liebeskind BJ, Hillis DM, Zakon HH. Evolution of sodium channels predates the origin of nervous systems in animals. *Proc Natl Acad Sci USA*. 2011;108(22):9154–9159.
56. Valdivia CR, Ackerman MJ, Tester DJ, *et al.* A novel SCN5A arrhythmia mutation, M1766L, with expression defect rescued by mexiletine. *Cardiovasc Res*. 2002;55(2):279–289.
57. Bechi G, Rusconi R, Cestele S, Striano P, Franceschetti S, Mantegazza M. Rescuable folding defective NaV1.1 (SCN1A) mutants in epilepsy: properties, occurrence, and novel rescuing strategy with peptides targeted to the endoplasmic reticulum. *Neurobiol Dis*. 2015;75:100–114.
58. Dhifallah S, Lancaster E, Merrill S, Leroudier N, Mantegazza M, Cesteèle S. Gain of function for the SCN1A/hNav1.1-L1670W mutation responsible for familial hemiplegic migraine. *Front Mol Neurosci*. 2018;11:232.
59. Horga A, Raja Rayan DL, Matthews E, *et al.* Prevalence study of genetically defined skeletal muscle channelopathies in England. *Neurology*. 2013;80(16):1472–1475.
60. Loussouarn G, Sternberg D, Nicole S, *et al.* Physiological and pathophysiological insights of Nav1.4 and Nav1.5 comparison. *Front Pharmacol*. 2015;6:314.
61. Antzelevitch C, Brugada P, Borggrefe M, *et al.* Brugada syndrome: report of the second consensus conference. *Heart Rhythm*. 2005;2(4):429–440.
62. Ku CS, Polychronakos C, Tan EK, *et al.* A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol Psychiatry*. 2013;18(2):141–153.
63. Holland KD, Bouley TM, Horn PS. Location: A surrogate for personalized treatment of sodium channelopathies. *Ann Neurol*. 2018;84(1):1–9.
64. Yoshinaga H, Sakoda S, Shibata T, *et al.* Phenotypic variability in childhood of skeletal muscle sodium channelopathies. *Pediatr Neurol*. 2015;52(5):504–508.
65. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15(7):901–913.
66. Balestrini S, Chiarello D, Gogou M, *et al.* Real-life survey of pitfalls and successes of precision medicine in genetic epilepsies. *J Neurol Neurosurg Psychiatry*. 2021;92(10):1044–1052.
67. National Research Council. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. The National Academies Press; 2011:142.
68. Gardella E, Møller RS. Phenotypic and genetic spectrum of SCN8A-related disorders, treatment options, and outcomes. *Epilepsia*. 2019;60(Suppl 3):S77–S85.
69. Suetterlin KJ, Bugiardini E, Kaski JP, *et al.* Long-term safety and efficacy of mexiletine for patients with skeletal muscle channelopathies. *JAMA Neurol*. 2015;72(12):1531–1553.
70. Mann SA, Castro ML, Ohanian M, *et al.* R222Q SCN5A mutation is associated with reversible ventricular ectopy and dilated cardiomyopathy. *J Am Coll Cardiol*. 2012;60(16):1566–1573.
71. Sadleir LG, Mountier EI, Gill D, *et al.* Not all SCN1A epileptic encephalopathies are Dravet syndrome: Early profound Thr226Met phenotype. *Neurology*. 2017;89(10):1035–1042.
72. Berecki G, Bryson A, Terhag J, *et al.* SCN1A gain of function in early infantile encephalopathy. *Ann Neurol*. 2019;85(4):514–525.
73. Anderson LL, Thompson CH, Hawkins NA, *et al.* Antiepileptic activity of preferential inhibitors of persistent sodium current. *Epilepsia*. 2014;55(8):1274–1283.
74. Männikkö R, Shenkarev ZO, Thor MG, *et al.* Spider toxin inhibits gating pore currents underlying periodic paralysis. *Proc Natl Acad Sci USA*. 2018;115(17):4495–4500.
75. Weuring WJ, Singh S, Volkers L, *et al.* NaV1.1 and NaV1.6 selective compounds reduce the behavior phenotype and epileptiform activity in a novel zebrafish model for Dravet Syndrome. *PLoS One*. 2020;15(3):e0219106.
76. Xu H, Li T, Rohou A, *et al.* Structural basis of Nav1.7 inhibition by a gating-modifier spider toxin. *Cell*. 2019;176(4):702–715.
77. Han Z, Chen C, Christiansen A, *et al.* Antisense oligonucleotides increase Scn1a expression and reduce seizures and SUDEP incidence in a mouse model of Dravet syndrome. *Sci Transl Med*. 2020;12(558):eaaz6100.
78. Matthews E, Labrum R, Sweeney MG, *et al.* Voltage sensor charge loss accounts for most cases of hypokalemic periodic paralysis. *Neurology*. 2009;72(18):1544–1547.
79. Castañeda MS, Zanoteli E, Scalco RS, *et al.* A novel ATP1A2 mutation in a patient with hypokalaemic periodic paralysis and CNS symptoms. *Brain*. 2018;141(12):3308–3318.
80. Xu H, Li T, Rohou A, *et al.* Structural basis of Nav1.7 inhibition by a gating-modifier spider toxin. *Cell*. 2019;176(4):702–715.e14.
81. Mason ER, Wu F, Patel RR, Xiao Y, Cannon SC, Cummins TR. Resurgent and gating pore currents induced by De Novo SCN2A epilepsy mutations. *eNeuro*. 2019;6(5):ENEURO.0141-19.2019.
82. Dunlop J, Bowlby M, Peri R, Vasilyev D, Arias R. High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology. *Nat Rev Drug Discov*. 2008;7(4):358–368.

Downloaded from https://academic.oup.com/brain/article/145/12/4275/6509259 by Universitaets- und Stadtbibliothek Koeln, Med. Abt. user on 29 April 2023

Supplementary Material

Gene variant effects across sodium channelopathies predict function and guide precision therapy

Andreas Brunklaus[+], Tony Feng[+], **Tobias Brünger[+]**, Eduardo Perez-Palma, Henrike Heyne, Emma Matthews, Christopher Semsarian, Joseph D Symonds, Sameer M Zuberi, Dennis Lal, Stephanie Schorge. **Brain**. DOI:10.1093/brain/awac006. (published)

[+]These authors contributed equally to the work

The supplementary material is listed below and accessible at https://doi.org/10.1093/brain/awac006. Due to its large size Supplementary Table 2 is only available online at https://doi.org/10.1093/brain/awac006.

Supplementary Methods

Functional classification of mutations:

For each mutation described we assessed the overall effect with respect to channel function in terms of net current flow during voltage steps. Channel function is not necessarily the same as overall effect on *cellular* function as elegantly shown recently in Liu et al., 2019,[1] where a gain in function for the channel, led to a paradoxical loss of function through depolarising block at the cellular level. As the vast majority of channels are characterised in non-excitable mammalian cell lines in voltage clamp only, we have used this baseline to allow comparisons between published studies. Assessing the net effect where a mutation can have multiple effects on channels, we have looked at different parameters in the following approach:

1. Peak currents: we first checked whether a mutation reduced or increased the peak current substantially. Where data allow (e.g. for FHM3 mutations in Cestele et al., 2013)[2] we have also asked whether these effects on peak current are dependent on expression system in a way that means they may not be physiologically relevant in the endogenous cells. If there are data suggesting the effects on peak current are dependent on system, we have noted this as 'mixed' in the first instance.

2. Voltage dependence of inactivation: as the channels must be able to activate in order to pass current, we have asked whether there were shifts in the voltage dependence of inactivation that suggest a substantial portion of channels would be stuck in inactivated states in resting cells at ~-70 or -80 mV. In a very few cases we have suggested that were a large shift in voltage dependence of inactivation would effectively make channels non-functional, we have indicated this is likely an overall LoF effect, even if the peak current is increased and voltage dependence of activation (see #3) was also shifted (we gave inactivation precedence as the channels would be unable to open in resting cells if inactivated at resting potentials).

3. Voltage dependence of activation: given functional channels (i.e. peak current is not lost) that are not stuck in inactivated states at rest (i.e. voltage dependence of inactivation not shifted so all channels are inactive at -80 mV), we asked how much stimulation is required to open channels – that is to say, is the voltage dependence of activation shifted significantly? Here a shift to more

negative voltage dependences of activation would be a gain of function, but again – only if there are sufficient currents, and a significant population of non-inactivated channels.

4. Persistent currents: These can be increased, sometimes even when peak currents are reduced. Increases in persistent currents are included as gain of function, unless peak currents or voltage dependence of inactivation are shifted so much that channels are unlikely to be open at all, in which case, given the profound effects of sodium leaks on cellular activity we have still called cases with large increases in persistent currents 'mixed'.

5. Gating pore currents: Few groups using neuronal channels have been able to interrogate whether in the cases of S4 arginines these include gating pore leak, we have not rigorously separated gating pore leaks (but see R853Q in SCN2A for one example where these currents have been seen in neuronal channels; Mason et al., 2019)[3]. Where gating pore leaks are described, these are a gain of function (and one that is highly likely to be pathogenic in any cell). Further technically challenging work on neuronal channels will be needed to confirm the conservation of these gating pore currents across channel subtypes.

6. There are many additional features that can come into play, and where the above 5 criteria do not give an overriding answer, we have looked at these. For example, many groups also look at recovery from inactivation, but in our survey, this was rarely changed enough to be the dominating feature of a channel's functional change. However, where there were large impacts on recovery from inactivation and these were in contrast to other effects on the channel, we have given the results as mixed.

In practice this functional hierarchy, albeit simplistic and incomplete, typically agrees with the author assessments of the channels, with rare exceptions. We emphasise that this functional assessment is from the perspective of the channel, not the cell or the organism (indeed as these are mutations associated with diseases, they would all be loss of function from the organism's perspective).

Additional studies, including dynamic clamp, expression in excitable cells and interrogating current clamp properties, modelling, and the golden standard of knock-in studies in mice, all will bring additional insights to the consequences of the mutations on the cells, networks and behaviours, but for the purposes of comparing the effects of mutations across different channels we have relied on the reductionist, most commonly used experimental approach. It remains to be

seen whether functional effects that alter, for example channel trafficking, are conserved in different types of sodium channels expressed in different cellular backgrounds. In addition, some effects, for example where the different threshold of activation of SCN9A has specific effects on cellular activity (in this case in DRG neurons, Dib-Hajj et al., 2012)[4] which are highly unlikely to be conserved in (for example) SCN4A in muscle cells. In these cases, the reductionist effect on voltage dependence of activation may be conserved, but the cellular effect divergent.

References

1. Liu Y, Schubert J, Sonnenberg L, et al. Neuronal mechanisms of mutations in SCN8A causing epilepsy or intellectual disability. *Brain*. 2019;142(2):376-390.

2. Cestèle S, Schiavon E, Rusconi R, Franceschetti S, Mantegazza M. Nonfunctional NaV1.1 familial hemiplegic migraine mutant transformed into gain of function by partial rescue of folding defects. *Proc Natl Acad Sci USA*. 2013;110(43):17546-17551.

3. Mason ER, Wu F, Patel RR, Xiao Y, Cannon SC, Cummins TR. Resurgent and Gating Pore Currents Induced by De Novo SCN2A Epilepsy Mutations. *eNeuro*. 2019;6(5)

4. Dib-Hajj SD, Yang Y, Black JA, Waxman SG. The NaV1.7 sodium channel: from molecule to man. *Nature Reviews Neuroscience*. 2013;14(1):49-62.

Supplementary Table 2: Corresponding variants, phenotypes, and function across different sodium channels

Pair	*SCN1A* Position	Gene/ Variant	Function	Phenotype	Reference*	Corresponding Gene/Variant	Function	Phenotype	Reference*
1	I138V	***SCN4A***; I141V D1 S1	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	Sodium channel myotonia	Petitprez (2008)[23]	***SCN9A***; I136V D1 S1	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	IEM	Cheng (2008)[24]
2	R222W	***SCN4A***; R225W D1 S4	LoF; WCC: Y, ↓CD, →$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	Congenital Myopathy	Zaharieva (2016)[14]	***SCN5A***; R225W D1 S4	LoF; WCC: Y, ↓↓CD, →$V_{1/2\ Act.}$, →$V_{1/2\ Fl}$	Severe conduction disease	Bezzina (2003)[45]
3	S243T	***SCN9A***; S241T D1 S4-5	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	IEM	Lampert (2006)[58]	***SCN10A***; S242T D1 S4-5	GoF; WCC: Y, ←$V_{1/2\ Act.}$, ←$V_{1/2\ Fl}$	PPN, PDN; carbamazepine responder	Han (2018)[60]
4	Q267K	***SCN4A***; Q270K D1 S5	GoF/Mixed; WCC: Y, →$V_{1/2\ Act.}$, →$V_{1/2\ Fl}$	PMC	Carle (2009)[66]	***SCN5A***; Q270K D1 S5	GoF/Mixed; WCC: Y, ↓CD, ↑I_{NaP}, →$V_{1/2\ Act.}$, →$V_{1/2\ Fl}$	LQT3/BrS overlap syndrome; ECG: fetal tachycardia/ fibrillation TdP, LQT	Calloe (2011)[67]
5	R377H	***SCN2A***; R379H D1 S5-6	LoF; WCC: None	ASD	Ben-Shalom (2017)[5]	***SCN5A***; R367H D1 S5-6	LoF; WCC: None	BrS; SCD; ECG: ST elevation	Hong (2004)[83]
6	N416K	***SCN4A***; N440K D1 S6	GoF; WCC: Y, ↑I_{NaP}, no change$V_{1/2\ Act.}$, →$V_{1/2\ Fl}$	PMC	Lossin (2012)[92]	***SCN9A***; N395K D1 S6	GoF; WCC: Y, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	IEM	Sheets (2007)[41]
						SCN5A; N406K D1 S6	GoF/Mixed; WCC: Y, ↓CD, ↑I_{NaP}, no change$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$,	LQT3; ECG: TdP, LQT, polymorphic VT, mexiletine responder	Hu (2018)[93] Kato (2014)[94]
7	V421M	***SCN4A***; V445M D1 S6	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.}$, ←$V_{1/2\ Fl}$	PMC	Wang (1999)[97] Huang (2020)[98]	***SCN5A***; V411M D1 S6	GoF; WCC: Y, ↑CD, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	LQT3; ECG: neonatal LQT with 2:1 block	Horne (2011)[100] Zhou (2015)[101]
						SCN9A; V400M D1 S6	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.}$, →$V_{1/2\ Fl}$	IEM; carbamazepine responder	Fischer (2009)[102]
8	T782I	***SCN2A***; T773I D2 S1	GoF; WCC: Y, no changeCD, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	DEE, Sz onset 1 day	Lauxmann (2018)[33]	***SCN8A***; T767I D2 S1	GoF; WCC: Y, ↓CD, ↑I_{NaP}, ←$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	DEE, Sz onset 2 weeks	Pan (2020)[118] Estacion (2014)[119]
9	R859H	***SCN1A***; R859H D2 S4	LoF/mixed; WCC: Y, ↑I_{NaP}, no changeCD, ←$V_{1/2\ Act.}$, ←$V_{1/2\ Fl}$	GEFS+	Volkers (2011)[134]	***SCN4A***; R669H D2 S4	LoF/mixed; WCC: Y, ↓CD, no change$V_{1/2\ Act.}$, ←$V_{1/2\ Fl}$	HypoPP	Kuzmenkin (2002)[135]
10	R859C	***SCN1A***; R859C D2 S4	LoF; WCC: Y, ↓CD, no change$V_{1/2\ Act.}$, no change$V_{1/2\ Fl}$	EPI	Bechi (2015)[124]	***SCN5A***; R808C D2 S4	LoF; WCC: Y, ↓CD, no change$V_{1/2\ Act.}$, ←$V_{1/2\ Fl}$	BrS	Glazer (2020)[1]

11	R862H	***SCN4A***; R672H D2 S4	LoF; WCC: Y, ↓CD; $\rightarrow V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fi}$	Hypo-PP	Jurkatt-Rott (2000)[140] Kuzmenkin (2002)[135]	***SCN5A***; R811H D2 S4	LoF; WCC: Y, ↓CD; no change$V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fi}$	BrS; family history of sudden death,	Calloe (2013)[114]
12	R865Q	***SCN4A***; R675Q D2 S4	Mixed; WCC: Y, no change CD, no change$V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$, $\leftarrow V_{1/2\ SI}$	Potassium sensitive normoPP	Wu (2014)[149]	***SCN5A***; R814Q D2 S4	Mixed; WCC: Y, no change CD, no change$V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$	LQT3; BrS	Glazer (2020)[1]
13	I883T	***SCN3A***; I875T D2 S4-5	GoF; WCC: Y, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	EPI/PMG	Zaman (2018)[155]	***SCN9A***; I848T D2 S4-5	GoF; WCC: Y, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$	IEM	Cummins (2004)[157] Namer (2015)[158] Theile (2011)[159]
14	R946C	***SCN1A***; R946C; D2 S5-6	LoF; WCC: None	DS	Volkers (2011)[134]	***SCN2A***; R937C; D2 S5-6	LoF; WCC: None	ASD	Begemann (2019)[180]
15	R946H	***SCN1A***; R946H; D2 S5-6	LoF; WCC: None	DS	Liao (2010)[182] Volkers (2011)[134]	***SCN2A***; R937H; D2 S5-6	LoF; WCC: None	ASD	Ben-Shalom (2017)[5]
16	G979R	***SCN1A***; G979R; D2 S6	LoF; WCC: None	DS	Sugawara (2003)[184] Rhodes (2005)[130]	***SCN8A***; G964R; D2 S6	LoF; WCC: None	NDD without EPI	Wagnon (2017)[185]
17	D1256N	***SCN4A***; D1069N D3 S2	LoF/Mixed; WCC: Y, no change CD, $\rightarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	Congenital myopathy	Zaharieva (2016)[14]	***SCN5A***; D1243N D3 S2	LoF/Mixed; WCC: Y, no change CD, $\rightarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	BrS	Glazer (2020)[1]
18	A1343T	***SCN4A***; A1156T D3 S4-5	GoF; WCC: Y, no change$V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	PMC with prominent myalgia	Palmio (2017)[236]	***SCN5A***; A1330T D3 S4-5	GoF; WCC: Y, no change$V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	LQT3; SCD	Smits (2005)[238]
19	P1345L	***SCN3A***; P1333L D3 S4-5	GoF; WCC: Y, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$	EPI	Zaman (2018)[155]	***SCN4A***; P1158L D3 S4-5	GoF; WCC: Y, no change$V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	Sodium channel myotonia	Desaphy (2016)[239]
						SCN5A; P1332L D3 S4-5	GoF/mixed; WCC: Y, $\leftarrow V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fi}$	LQT3, ECG: TdP, LQT, mexiletine responder	Ruan (2007)[190]
						SCN9A; P1308L D3 S4-5	GoF; WCC: Y, ↓CD, $\leftarrow V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$	IEM	Cheng (2010)[226]
20	S1346Y	***SCN2A***; S1336Y D3 S4-5	GoF/Mixed ($Na_V1.2N$); WCC: Y, ↓CD, $\leftarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	DEE	Thompson (2020)[54]	***SCN5A***; S1333Y D3 S4-5	GoF; WCC: Y, no change CD, $\leftarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	SIDS; LQT3	Huang (2009)[243]
21	V1353A	***SCN9A***; V1316A D3 S4-5	GoF; WCC: Y, $\leftarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	IEM	Wu (2013)[252], Estacion (2013)[253]	***SCN11A***; V1184A D3 S5	GoF; WCC: Y, ↑CD, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, no change$V_{1/2\ Fi}$	PPN, CAP	Leipold (2015)[255]
22	F1486C	***SCN4A***; F1298C D3-D4 linker	GoF/Mixed; WCC: Y, $\rightarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	Sodium channel myotonia	Farinato (2019)[272]	***SCN5A***; F1473C D3-D4 linker	GoF; WCC: Y, $\uparrow I_{NaP}$, no change$V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fi}$	LQT3; ECG: TdP, LQT with 2:1 block, mexiletine responder	Bankston (2007)[276]

23	F1499L	***SCN1A***; F1499L D3-D4 linker	GoF; WCC: Y, $\uparrow I_{NaP}$, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	FHM	Barbieri (2019)[283]	***SCN5A***; F1486L D3-D4 linker	GoF; WCC: Y, $\uparrow I_{NaP}$, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	LQT3	Wang (2007)[36]
24	T1501I	***SCN3A***; T1486I D3-D4 linker	GoF; WCC: Y, $\uparrow I_{NaP}$, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	EPI/PMG	Zaman (2020)[147]	***SCN9A***; T1464I D3-D4 linker	GoF; WCC: Y, $\uparrow I_{NaP}$, $\rightarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	PEPD, responsive to SCB	Fertleman (2006)[227] Theile (2011)[159]
25	E1587K	***SCN1A***; E1587K D4 S2	LoF; WCC: None	EPI	Kluckova (2020)[10]	***SCN5A***; E1574K D4 S2	LoF; WCC: Y, $\downarrow\downarrow$CD, $\rightarrow V_{1/2\ Act.}$	BrS	Glazer (2020)[1]
26	R1596C	***SCN1A***; R1596C D4 S2-3	LoF; WCC: None	EPI	Kluckova (2020)[10]	***SCN5A***; R1583C D4 S2-3	LoF; WCC: Y, $\downarrow$CD, no change I_{NaP}, no change $V_{1/2\ Act.}$, no change $V_{1/2\ Fl}$	BrS	Glazer (2020)[1]
27	P1632S	***SCN1A***; P1632S; D4 S3-4	LoF; WCC: Y, $\leftarrow V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fl}$	DS	Rhodes (2005)[130]	***SCN2A***; P1622S; D4 S3-4	LoF; WCC: Y, $\leftarrow V_{1/2\ Fl}$	ASD and Sz onset 21 months	Wolff (2017)[96]
28	R1636Q	***SCN3A***: R1621Q D4 S4	GoF; WCC: Y, no change CD, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, no change $V_{1/2\ Fl}$	EPI/PMG	Zaman (2020)[147]	***SCN8A***; R1617Q D4 S4	GoF; WCC: Y, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	DEE, Sz onset 3 months	Wagnon (2015)[314]
						SCN5A; R1623Q D4 S4	GoF; WCC: Y, $\uparrow I_{NaP}$, $\leftarrow V_{1/2\ Act.}$	LQT3; ECG: TdP. LQT. mexiletine responder	Kambouris (1998)[310] Tsurugi (2009)[311]
29	R1639L	***SCN4A***; R1451L D4 S4	LoF; WCC: Y, $\downarrow$CD, no change $V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fl}$	Complex phenotype including myotonia and paralysis (both potassium sensitive and hypoPP	Poulin (2018)[318]	***SCN8A***; R1620L D4 S4	LoF; WCC: Y, $\downarrow\downarrow$CD, no change $V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fl}$	ASD	Liu (2019)[288]
30	R1645H	***SCN4A***; R1457H D4 S4	LoF; WCC: Y, no change $V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fl}$	CMS (in patient homozygous for R1457H variant)	Arnold (2015)[326]	***SCN5A***; R1632H D4 S4	LoF; WCC: Y, no change CD, no change $V_{1/2\ Act.}$, $\leftarrow V_{1/2\ Fl}$	SSS; ECG: bradycardia, absent atrial depolarizations, prolonged QRS, 1° heart block	Benson (2003)[327]
31	R1657C	***SCN1A***; R1657C; D4 S4-5	LoF; WCC: Y, $\downarrow$CD, $\rightarrow V_{1/2\ Act.}$, no change $V_{1/2\ Fl}$	GEFS+	Lossin (2003)[186]	***SCN5A***; R1644C D4 S4-5	LoF; WCC: Y, $\rightarrow V_{1/2\ Act.}$, no change $V_{1/2\ Fl}$	BrS; ECG: ST elevation, echo: CM changes	Frustaci (2005)[87]
						SCN8A; R1638C; D4 S4-5	LoF; WCC: Y, $\rightarrow V_{1/2\ Act.}$, no change $V_{1/2\ Fl}$	NDD without epilepsy	Wengert (2019)[65]
32	F1661S	***SCN1A***; F1661S D4 S4-5	Mixed; 50% reduction in trafficking WCC: Y, $\downarrow$CD, $\uparrow I_{NaP}$, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	DS	Rhodes (2004)[172] Thompson (2012)[187]	***SCN4A***; F1473S D4 S4-5	GoF; WCC: Y, CD not reported, $\uparrow I_{NaP}$, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	PMC	Fleischhauer (1998)[336]
33	M1664K	***SCN1A***; M1664K D4 S4-5	LoF; 90% reduction in peak current and trafficking not allowing for detailed *SCN1A* biophysics	GEFS+/DS	Bechi (2015)[124]	***SCN9A***; M1627K D4 S4-5	GoF; WCC: Y, no change CD, no change $V_{1/2\ Act.}$, $\rightarrow V_{1/2\ Fl}$	PEPD	Fertleman (2006)[227] Dib-Hajj (2008)[338] Theile (2011)[159]

34	G1674R	***SCN1A***; G1674R D4 S5	LoF; WCC: None ↓↓CD	EPI	Rhodes (2004)[172] Thompson (2012)[87]	***SCN5A***; G1661R D4 S5	LoF; WCC: Y (barely) ↓↓CD	BrS	Glazer (2020)[1]
35	M1780I	***SCN3A***; M1765I D4 S6	GoF; WCC: Y, ↑I_{NaP}, ←$V_{1/2\ Act.,\ no\ change}V_{1/2\ FI}$	EPI/PMG	Zaman (2020)[147]	***SCN8A***; M1760I D4 S6	GoF; WCC: Y, ←$V_{1/2\ Act.,\ no\ change}V_{1/2\ FI}$	EPI	Liu (2019)[288]
36	N1788D	***SCN5A***; N1774D C-Term	GoF; WCC: Y, ↑CD, ↑I_{NaP}, ←$V_{1/2\ Act.,\ no\ change}V_{1/2\ FI}$	LQT3; ECG: TdP, LQT with 2:1 block, mexiletine responder	Kato (2014)[94]	***SCN8A***; N1768D C-Term	GoF; WCC: Y, ↑I_{NaP}, $_{no\ change}V_{1/2\ Act.,}$ →$V_{1/2\ FI}$	DEE, Sz onset 6 months	Veeramah (2012)[366] Patel (2016)[247] Baker (2018)[367]
37	R1892Q	***SCN2A***; R1882Q; C-Term	GoF; WCC: Y, ↑CD, ↑I_{NaP}, ←$V_{1/2\ Act.,}$ →$V_{1/2\ FI}$	DEE, Sz onset 1 day	Berecki (2018)[141] Mason (2019)[142] Wolff (2017)[96]	***SCN8A***; R1872Q; C-Term	GoF; WCC: Y, ↑CD, ←$V_{1/2\ Act.,}$ →$V_{1/2\ FI}$	DEE, Sz onset 4 months	Wagnon (2015)[314] Aktin (2018)[393]
38	Q1923R	***SCN1A***; Q1923R C-Term	LoF; WCC: None	DS	Nissenkorn (2019)[29]	***SCN5A***; Q1909R C-Term	Mixed; ←$V_{1/2\ Act.,\ no\ change}V_{1/2\ FI}$ decrease in peak current by 50%.	SIDS (not a known cardiac patient)	Winkel (2015)[103] Abdelsayed (2017)[369]

Corresponding variant = identical variant among different *SCN* at the same position/location in the SCN protein (the corresponding sequence numbers are not identical as the amino acid sequence between *SCN* variants differs slightly).

Rows marked in grey denote variant pairs with divergent functional properties.

*References relate to those detailed in Supplementary Table 1.

Phenotypical features: ASD = autism spectrum disorder, BrS = Brugada syndrome, CAP = cold aggravated pain, CM changes = cardiomyopathic changes, CMS = congenital myasthenic syndrome, DEE = developmental and epileptic encephalopathy, DS = Dravet syndrome, ECG = electrocardiogram, echo = echocardiogram, EPI = epilepsy, FHM3 = familial hemiplegic migraine type 3, GEFS+ = genetic epilepsy with febrile seizures plus, Hyper-PP = hyperkalaemic periodic paralysis, Hypo-PP = hypokalaemic periodic paralysis, IEM = inherited erythromelalgia, LQT3 = long QT3 syndrome, $Na_v1.2N$ = neonatal proteoform, NDD = neurodevelopmental disorder, PAM = potassium-aggravated myotonia, PDN = painful diabetic neuropathy, PEPD = paroxysmal extreme pain disorder, PMC = paramyotonia congenita, PMG = polymicrogyria, PPN = painful peripheral neuropathy, SCB = sodium channel blocker, SCD = sudden cardiac death, SIDS = sudden infant death syndrome, SSS = sick sinus syndrome, SUD = sudden unexplained death, Sz = seizure, TdP = torsade de pointes, VT = ventricular tachycardia

Electrophysiological key features: Arrows (→) are used for electrophysiological parameters. The direction of the arrows indicates hyperpolarizing (←) or depolarizing shifts (→), as well as an increase (↑) or decrease (↓) of parameters, (↓↓ = >50% decrease)

Electrophysiological abbreviations: GoF: gain-of-function, LoF: loss-of-function, WCC: whole cell current (Y = measurable, N = not measurable), Act: activation, CD: current density, FI: fast inactivation, SI: slow inactivation, I_{NaP}: persistent sodium current, $V_{1/2Act.}$: half-activation of steady-state activation curve, $V_{1/2FI}$: half-inactivation of steady-state fast inactivation curve.

Supplementary Table 3: Detailed *SCN1-11A* Analysis

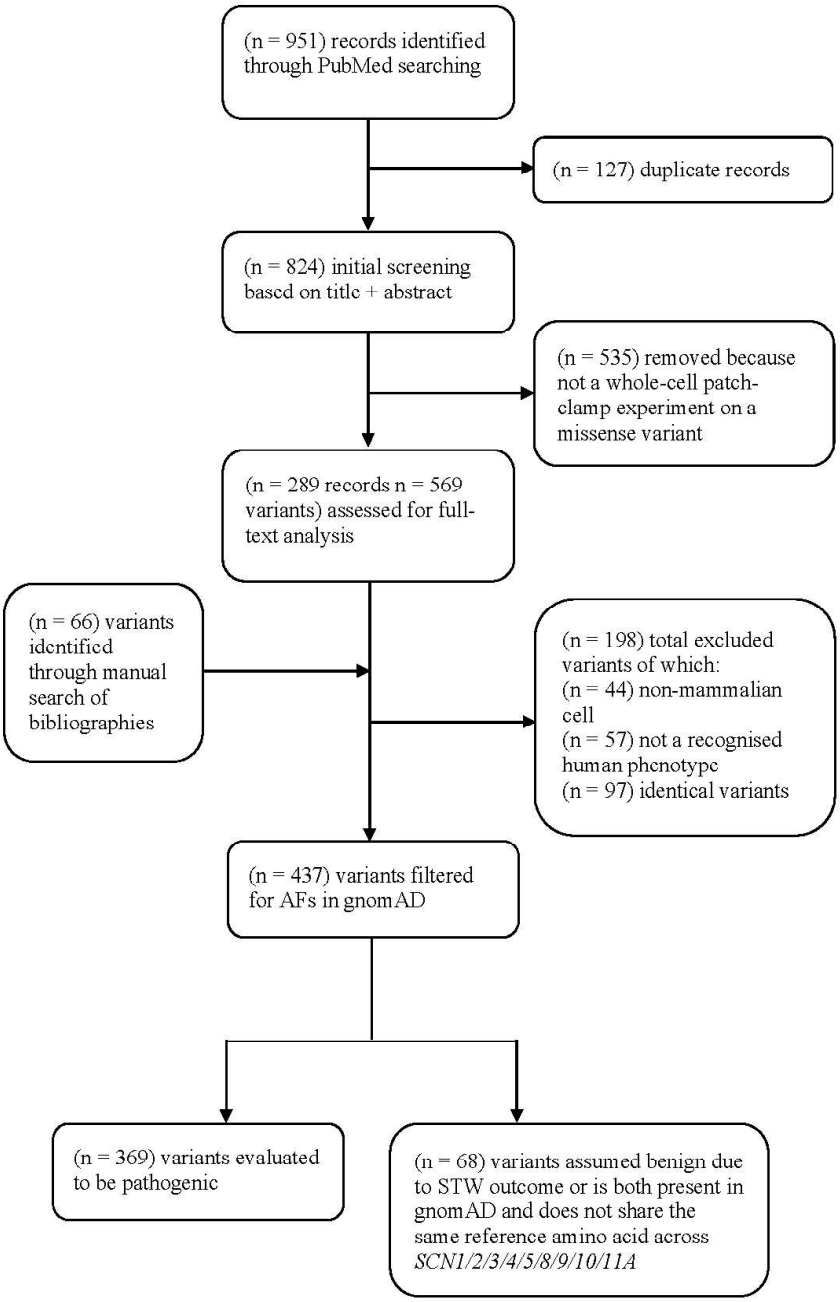
SCN1A				
Disorder	**No.**	**Functional Effects**	**Distribution**	**Clinical Context**
EPI	60 (85%)	• GoF – (3/60, 5%) • LoF – (47/60, 78%) • Mixed – (10/60, 17%)	*SCN1A variants clustered across pore-loop regions (S5, S5-6 & S6) and 88% of variants showed LoF (23/26)*	Variants displaying LoF were more frequently associated with DS, whereas variants showing GoF were FHM3-associated (p<0.001)
FHM3	11 (15%)	• GoF – (8/11, 73%) • LoF – (1/11, 9%) • Mixed – (2/11, 18%)	55% (6/11) of FHM3-associated variants occurred in sites implicated in channel inactivation, including the DIII-IV linker, DIVS4-5 and DIVS6, predominantly showing GoF effects. Of these, 83% were GoF (5/6). While most FHM3-associated variants are associated with GoF effects, the variant located in DIVS4 was clearly mixed. The other mixed variant was located in DIVS5	
SCN2A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
EPI	29 (81%)	• GoF – (17/29, 59%) • LoF – (7/29, 24%) • Mixed – (5/29, 17%)	*SCN2A* epilepsy-associated variants were evenly distributed across homologous domains but very few were found in pore-forming regions	Variants displaying LoF effects were frequently associated with ASD, while GoF variants were epilepsy-associated (p = 0.001)
ASD	7 (19%)	All variants showed LoF	71% of ASD-associated variants clustered in pore-loop regions (5/7), all displaying LoF	
SCN3A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
EPI	5 (38%)	• GoF – 3/5 (60%) • LoF – 1/5 (20%) • Mixed – 1/5 (20%)	Variants associated with an epilepsy or mixed phenotype were evenly distributed across the protein and only two were observed in pore-forming regions	Epilepsy and PMG variants appear to be mainly GoF
EPI/PMG	6 (46%)	All variants showed GoF		
ASD	1 (8%)	LoF	DIVS5	
Fetal Akinesia	1 (8%)	GoF	DIIS4	

SCN4A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
PMC	14 (36%)	• GoF – (11/14, 86%) • LoF – (1/14, 7%) • Mixed – (2/14, 7%)	43% occurred in the DIII-IV linker (6/14) and the remainder across the protein	GoF variants were frequently PAM/PMC-associated, whereas LoF variants were CMS-associated ($p<0.001$). Overall, 32% of variants occurred in inactivation sites (12/38), 92% of which showed GoF (11/12). 29% of variants were found in S4 sites (11/38), 70% of which showed LoF (8/11). 18% occurred in pore-forming regions (7/38) and the remainder across the protein.
CMS	10 (26%)	• LoF – (8/10, 80%) • Mixed – (2/10, 20%)	Two variants occurred in pore-loops and two in voltage-sensing regions.	
PAM	4 (11%)	All variants showed GoF	Three variants were found in inactivation sites and one in the C-terminus	
HypoPP	5 (13%)	Four variants displayed LoF and one variant showed mixed function	All four variants displaying LoF were found in S4 sites while the LoF variant was found in DIIIS4-5	
HyperPP	1 (3%)	Mixed	DIVS5	
NormoPP	1 (3%)	Mixed	DIIS4	
Mixed Phenotype	3 (8%)	All variants showed GoF	Variants occurred in DIS6, DIIS5 and DIIIS4-5	
SCN5A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
BrS	100 (69%)	96% of variants showed LoF (96/100), three were mixed effect and one displayed GoF	52% of BrS-associated variants occurred in pore-loop regions (S5, S5-6 & S6) (52/100), while only 6% were found at sites of channel inactivation (6/100)	Variants displaying LoF effects were more frequently BrS-associated, while GoF variants were LQT3-associated, ($p<0.001$)
LQT3	38 (26%)	• GoF – (30/38, 79%) • LoF – (2/38, 5%) • Mixed – (6/38, 16%)	47% of LQT3 variants clustered in sites of inactivation (18/38), showing predominantly GoF (94%, 17/18)	
Mixed	8 (5%)	• LoF – 4/8 (50%) • Mixed – 4/8 (50%)	Two variants occurred in the C-terminus, one in DIS4, one in DIIIS4 and the rest across the protein	

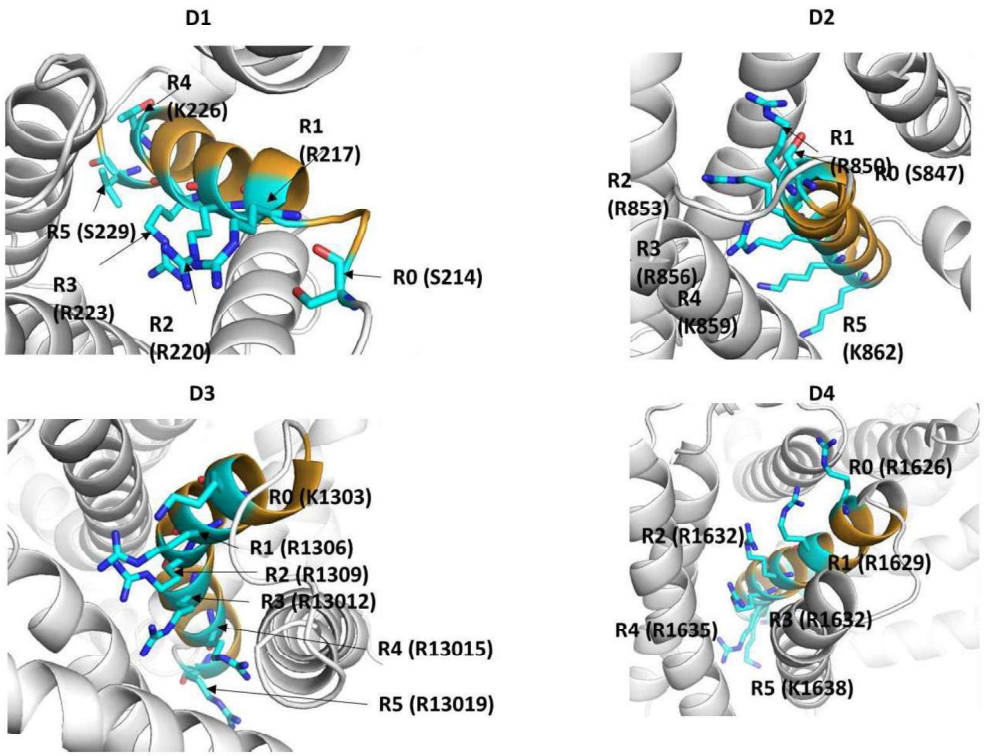
SCN8A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
EPI	14 (67%)	• GoF – (12/14, 86%) • LoF – (2/14, 14%)	29% occurred in S4 regions (4/14), 14% in the C-terminus (2/14), 14% in inactivation sites (2/14) and the remainder across the protein. None occurred in cytoplasmic regions.	Variants displaying LoF effects were frequently NDD/ASD-associated, while GoF variants were epilepsy-associated, (p=0.003).
NDD	5 (23%)	All variants showed LoF	Three were found in pore-forming regions, one in DIVS4-5 and one in DIIS3	
ASD	2 (9%)	One variant showed GoF while the other LoF	Both variants were located in DIVS4	
SCN9A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
PPN	34	• GoF – 33/34 (97%) • Mixed – 1/34 (3%)	Of 33 GoF variants, 33% were found in inactivation sites (11/33), 18% in DIIS5 (6/33), 15% in S4 regions (5/33) and the remainder across the protein. The mixed effect variant was located in DIIS4.	The majority of variants appear to be GoF
SCN10A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
AF	1	LoF	The LoF variant was observed in the N-terminus	GoF variants appear to be associated with PPN
SUD	2	Both variants showed LoF	Variants were observed in DIS6 and DIIS4	
PPN	3	All variants showed GoF	Variants were found at DIS4-S5, DIVS5-6 and DIVS6	
SCN11A				
Disorder	**No.**	**Functional Effect**	**Distribution**	**Clinical Context**
PPN	4	All variants showed GoF	PPN-associated variants were found in DIS4, DIIS4-5, DIIIS4 and DIIIS5	The majority of variants appear to be GoF.

Phenotypical features: FHM3 = familial hemiplegic migraine type 3, EPI = epilepsy, ASD = autism spectrum disorder, NDD = neurodevelopmental disorder, PMG = polymicrogyria, Hypo-PP = hypokalaemic periodic paralysis, Hyper-PP = hyperkalaemic periodic paralysis, PMC = paramyotonia congenita, CMS = congenital myasthenic syndrome, PAM = potassium-aggravated myotonia, BrS = Brugada syndrome, LQT3 = long QT3 syndrome, SCD = sudden cardiac death, PPN = peripheral painful neuropathy (including, PEPD = paroxysmal extreme pain disorder, IEM = inherited erythromelalgia and SFN = small fibre neuropathy), AF = atrial fibrillation, SUD = sudden unexpected death.

Supplementary Figure 1: Study Selection



Supplementary Figure 2: Voltage sensing regions (S4) structure zoom across D1-D4



Voltage sensing regions (S4) structure zoom across D1-D4. Close-up view of S4 voltage sensing regions across all four domains (D1-D4) illustrating conserved Arginines R0-R4.

6.3 Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes

Tobias Brünger, Eduardo Pérez-Palma, Ludovica Montanucci, Michael Nothnagel, Rikke S Møller, Stephanie Schorge, Sameer Zuberi, Joseph Symonds, Johannes R Lemke, Andreas Brunklaus, Stephen F Traynelis, Patrick May, Dennis Lal. **Brain**.
DOI:10.1093/brain/awac305. (published)

https://doi.org/10.1093/brain/awac305 BRAIN 2023: 146; 923–934 | 923

BRAIN
ORIGINAL ARTICLE

Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes

Tobias Brünger,[1] Eduardo Pérez-Palma,[2] Ludovica Montanucci,[3] Michael Nothnagel,[1,4] Rikke S. Møller,[5] Stephanie Schorge,[6] Sameer Zuberi,[7,8] Joseph Symonds,[7,8] Johannes R. Lemke,[9,10] Andreas Brunklaus,[7,8] Stephen F. Traynelis,[11] Patrick May[12] and Dennis Lal[1,3,13,14]

Clinically identified genetic variants in ion channels can be benign or cause disease by increasing or decreasing the protein function. As a consequence, therapeutic decision-making is challenging without molecular testing of each variant. Our biophysical knowledge of ion-channel structures and function is just emerging, and it is currently not well understood which amino acid residues cause disease when mutated.

We sought to systematically identify biological properties associated with variant pathogenicity across all major voltage and ligand-gated ion-channel families. We collected and curated 3049 pathogenic variants from hundreds of neurodevelopmental and other disorders and 12 546 population variants for 30 ion channel or channel subunits for which a high-quality protein structure was available. Using a wide range of bioinformatics approaches, we computed 163 structural features and tested them for pathogenic variant enrichment. We developed a novel 3D spatial distance scoring approach that enables comparisons of pathogenic and population variant distribution across protein structures. We discovered and independently replicated that several pore residue properties and proximity to the pore axis were most significantly enriched for pathogenic variants compared to population variants. Using our 3D scoring approach, we showed that the strongest pathogenic variant enrichment was observed for pore-lining residues and alpha-helix residues within 5Å distance from the pore axis centre and not involved in gating. Within the subset of residues located at the pore, the hydrophobicity of the pore was the feature most strongly associated with variant pathogenicity. We also found an association between the identified properties and both clinical phenotypes and functional in vitro assays for voltage-gated sodium channels (SCN1A, SCN2A, SCN8A) and N-methyl-D-aspartate receptor (GRIN1, GRIN2A, GRIN2B) encoding genes. In an independent expert-curated dataset of 1422 neurodevelopmental disorder pathogenic patient variants and 679 electrophysiological experiments, we show that pore axis distance is associated with seizure age of onset and cognitive performance as well as differential gain versus loss-of-channel function.

In summary, we identified biological properties associated with ion-channel malfunction and show that these are correlated with *in vitro* functional readouts and clinical phenotypes in patients with neurodevelopmental disorders. Our results suggest that clinical decision support algorithms that predict variant pathogenicity and function are feasible in the future.

1 Cologne Center for Genomics, University of Cologne, 50931 Cologne, Germany
2 Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana, Universidad de Desarrollo, Santiago 7590943, Chile
3 Lerner Research Institute Cleveland Clinic, Genomic Medicine Institute, Cleveland, OH 44195, USA
4 University Hospital Cologne, 50937 Cologne, Germany
5 Department of Epilepsy Genetics and Personalized Treatment, the Danish Epilepsy Center, DK 4293 Dianalund, Denmark

Received February 22, 2022. Revised July 29, 2022. Accepted August 09, 2022. Advance access publication August 29, 2022
© The Author(s) 2022. Published by Oxford University Press on behalf of the Guarantors of Brain.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

6 Department of Neuroscience, Physiology and Pharmacology, UCL, London WC1E 6BT, UK
7 The Paediatric Neurosciences Research Group, Royal Hospital for Children, Glasgow, UK
8 Institute of Health and Wellbeing, University of Glasgow, UK
9 Institute of Human Genetics, University of Leipzig Medical Center, 04103 Leipzig, Germany
10 Center for Rare Diseases, University of Leipzig Medical Center, 04103 Leipzig, Germany
11 Department of Pharmacology, Emory University School of Medicine, Rollins Research Center, Atlanta, GA 30322-3090, USA
12 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, L-4362 Esch-sur-Alzette, Luxembourg
13 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
14 Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, OH 44195, USA

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

Keywords: genetics; epilepsy; neurodevelopmental disorder; ion channel; bioinformatics

Introduction

Ion channels are membrane proteins that act as gated pathways for the movement of ions across cell membranes. They play essential roles in the physiology of all cells. At least 130 'channelopathies' have been associated with genetic variants causing ion-channel dysfunctions across the human body, including diseases in the nervous system (e.g. epilepsy, familial hemiplegic migraine), cardiovascular system (e.g. long QT syndrome, Brugada syndrome), respiratory system (e.g. cystic fibrosis) and endocrine system (e.g. neonatal diabetes mellitus).[1,2]

Ion channels malfunction due to missense variants that alter the protein sequence can cause a wide spectrum of disorders.[1,3] Missense variants even within the same gene can produce different molecular effects in ion channels: increased or decreased ion permeation; dysregulation of gating elements; changes in opening and closing probabilities; altered kinetics and protein trafficking.[4–9] Notably, not all variants that cause a molecular change from the wild-type cause necessarily a disease.[10,11] The location of missense variants on protein structure has been associated with variant pathogenicity.[12–14] In addition, several recent studies showed that among pathogenic variants, the variant position in specific functional units correlates with the patient subphenotype.[15–22] However, the knowledge about microdomain to phenotype associations is sparse and typically only available to experts of specific channels and cannot easily be adapted by a wider audience. Variant interpretation of less studied and recently identified channelopathies represents a major challenge.

In this study, we sought to identify biological properties associated with variant pathogenicity that are shared among all ion channels. These properties could be incorporated into future algorithms for predicting variant pathogenicity and, thus, provide insights into genotype–phenotype correlations in established and newly identified channelopathies regardless of the ion-channel class. To accomplish this goal, we performed a series of association analyses for biological properties with 12 546 population and 3049 pathogenic variants across 30 ion-channel structures. This work led to the development of a bioinformatic structure-based framework that showed that the distance of a residue from the central pore axis and pore hydrophobicity are the most pathogenicity-associated features shared across all ion channels. We applied our framework to two independent data sets comprising >1400 patient variants and ~700 functionally tested variants, for the NMDA receptor and voltage-gated sodium channels to demonstrate how correlations between the localization of variants and their associated molecular effect and clinical outcome can be captured.

Materials and methods

Collection of ion channels and protein structure selection

We collected all voltage and ligand-gated ion channels from IUPHAR[23] ($n = 172$ genes). Out of these channels, we included those in our study that harbour at least one known pathogenic variant in the ClinVar[24] database and whose encoded protein structures have been determined at <5Å resolution in the protein data bank[25] and form a pore. For each channel, the protein structure with the largest protein sequence coverage was chosen (Supplementary Table 1). All selected ion channels were divided into evolutionarily derived channel families according to HGNC[26] criteria.

Missense variant annotation

Canonical transcripts for proteins encoded by the ion-channel genes in our cohort were accessed from the UniProt[27] database. For these transcripts, variants were collected from multiple databases. In particular, protein-coding missense variants from the general population were retrieved from the genome aggregation Database[28] (https://gnomad.broadinstitute.org/, public release 2.0.2) including all ethnic backgrounds. Variant call format files were downloaded for all available exomes of ion-channel genes to access the variants. These 'population variants' served as control dataset in subsequent enrichment and association analyses. The extraction of missense variants (Filter = 'PASS') was performed with vcftools (v.0.1.12b) using the pre-annotated 'CSQ' field. ClinVar[24] variants were accessed from the ftp site (ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/, accessed July 2020). Only variants with a clinical consequence (annotated as 'Pathogenic'/'Likely Pathogenic') were extracted. Next, we obtained missense variants from the Human Gene Mutation Database (HGMD[29]) and filtered them for high-confidence calls (hgmd_confidence = 'HIGH' flag) as well as disease-causing states (hgmd_variant_Type = 'DM' flag) (HGMD, July 2020). Variants from ClinVar and HGMD databases

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

were combined into a pathogenic variant data set containing unique amino acid substitutions. We applied randomized sampling to create separate subsets of the collected variants to obtain discovery and validation cohorts each containing 70 and 30% of the pathogenic and population variants, respectively.

Collection of missense variants with detailed clinical phenotype or known molecular effect

We collected missense variants with known clinical phenotypes for three sodium channels (*SCN1A*, *SCN2A*, *SCN3A*) from the SCN Portal[30] (https://scn-portal.broadinstitute.org, accessed 1 December 2021). Missense variants with known clinical phenotype for three genes that encode ionotropic glutamate receptors (*GRIN1A*, *GRIN2A*, *GRIN2B*) were obtained from the GRIN Portal[31] (https://grin-portal.broadinstitute.org, access 1 December 2021). In addition, we obtained missense variants across all voltage-gated sodium channel encoding genes (*SCN1A*-*SCN11A*) with a reported molecular effect from Brunklaus *et al.*[22] Variants were categorized as either gain-of-function (GoF), loss-of-function (LoF) or mixed-function (mixed) function depending on the electrophysiological readouts. In addition, missense variants with a functional effect in any of the GRIN genes had been obtained (CFERV,[32] http://functionalvariants.emory.edu/, GRIN Portal[31]).

Collection and annotation of features

We annotated features that describe the localization of residues in functional regions of the protein as well as their protein structure (Supplementary Fig. 1). Therefore, we first mined the UniProt[27] database (https://www.uniprot.org/, accessed 2 May 2021) to collect the location of the amino acids inside the protein (intracellular, membrane-spanning or extracellular) and whether a residue is part of a specific unit (voltage sensor, allosteric-, agonist or other binding sites). The secondary structure of amino acid residues was calculated using the Dictionary of Protein Secondary Structure[33] program. Protein–protein interactions were obtained from the PDBsum[34] database (accessed 17 May 2021). We further considered two differently derived pore annotations to capture residues that are predicted to be located at the pore. First, a residue was considered to be part of the protein according to its annotation in the UniProt database. Second, we predicted the pore using the Mole2.0 webserver[35] (https://mole.upol.cz/, accessed 3 April 2021) and obtained the annotation of those residues that were predicted to be pore lining.

Statistical analysis

If not otherwise stated, we used a two-by-two contingency table to calculate the fold enrichment of pathogenic over population variants across different sets of residues, using the odds ratio (OR), augmented by the 95% confidence interval, as an estimate for the enrichment. We used a two-sided Fisher's exact test to test for the association while adjusting for multiple testing using Bonferroni correction. For the identification of important features (single features or feature combinations that describe a set of residues being enriched for pathogenic variants), a randomly subsampled set maintaining 70% of the population and pathogenic variants was considered as a discovery dataset. The remaining set consisting of each 30% population and pathogenic variants formed the validation dataset that was used to confirm the important single features and feature combinations identified in the discovery dataset.

A principal component analysis (PCA) was performed with the stats package (v.3.6.2) to aggregate the variance of the seven different physiochemical pore features across the pore residues. All statistical analyses were performed in R v.4.1.2.

Data availability

Residue-wise feature and distance annotations for all ion channels included in this study can be obtained from Supplementary Tables 2 and 3 and our GitHub repository (https://github.com/LalResearchGroup/Channel_Distances).

Results

Identification of pathogenic variant associated residue features across ion channels

We sought to identify features associated with variant pathogenicity in all ion channels (Fig. 1) Therefore, we collected 2411 pathogenic and 638 probably-pathogenic classified variants from ClinVar[24] and HGMD[29] and 12 546 population missense variants from the gnomAD database.[28] Of these, 2543 (83.1%) pathogenic variants and 5967 (47.6%) population variants could be mapped onto the protein structure of 30 ion-channel proteins (see 'Materials and methods' section for details). We calculated the fold enrichment of patient variants over population variants that served as controls in different residue sets.

For each amino acid residue, we computed 13 features that captured its structural and functional context (Supplementary Fig. 1 and Supplementary Table 2, also available in our GitHub repository: https://github.com/LalResearchGroup/Channel_Distances). These features are mainly independent of each other and include the secondary protein structure, the localization within a functional site (e.g. residues located at a ligand-binding site), and within structurally defined protein regions (e.g. residues located at the pore or inside the membrane) (Supplementary Figs 1A and 2). We then used these features to identify residue characteristics that are associated with pathogenic genetic variants compared to population variants (Fig. 1) across all channel proteins. We did observe a strong correlation across effective sizes for enrichment analyses of pathogenic versus likely-pathogenic variants in comparison with population variants (Supplementary Fig. 3 and Supplementary Table 4; Pearson correlation $\rho = 0.97$, $P = 1.7 \times 10^{-06}$). For this reason and to increase statistical power, pathogenic and likely-pathogenic variants were combined and from here on forward referred to as 'pathogenic' in the manuscript. Groups of residues that share the same features (either one or multiple features) are here referred to as 'residue sets' (Supplementary Fig. 1B, Supplementary material). For example, one residue set contains all residues that form helices (10 500 residues). Furthermore, a set of residues can be also described by a combination of features. For example, 104 residues are all (i) located at the pore; (ii) form helices; and (iii) form protein–protein interactions. The combinatorial assignment of the 13 features resulted in the identification of 163 different residue sets. We then investigated if there is an enrichment of pathogenic variants in each of the 163 residue sets. In the discovery screen, where we used a randomly assigned subset containing each 70% of the curated pathogenic and population variants, we found that 52 residue sets were enriched for pathogenic compared to population variants (Fisher's exact test, Bonferroni-adjusted $P < 0.05$ for

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

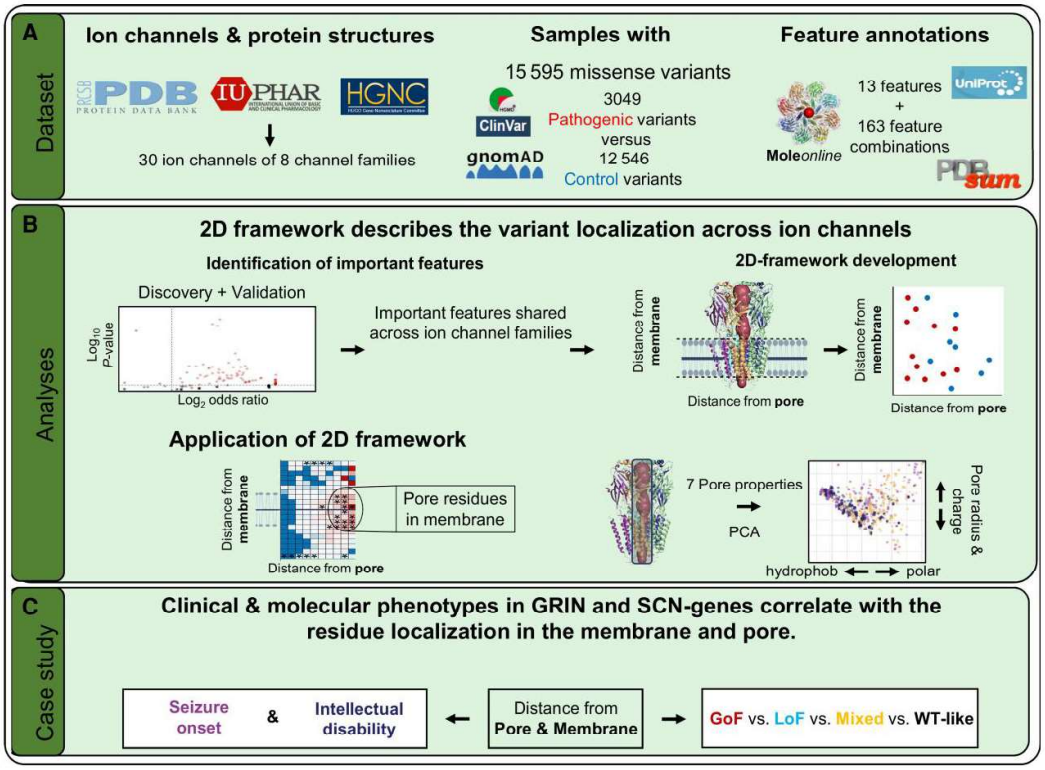


Figure 1 Graphical summary of the study. (**A**) Dataset. First, a subset of ion channels was selected from IUPHAR and screened for available protein structures. Second, missense variants were assembled from three different databases: gnomAD, ClinVar and HGMD. Third, structural features were annotated to describe the location and secondary structure of all residues comprising the ion channels. (**B**) Analyses. Sets of residues described by a single feature or a combination of features that were enriched for pathogenic variants were identified across all ion channels. A subset of these features was identified to show enrichment for pathogenic variants across all considered channel families. On the basis of these shared features a two-dimensional reference framework was developed for describing the location of a residue with respect to the distance from the membrane and the pore. Using this framework, the highest variant burden was observed in pore residues that were located inside the membrane. Finally, within these pore residues, the impact of the biophysical pore environment on the variant burden was investigated. (**C**) Case studies. Our framework identified correlations between functional and clinical phenotypes and the localization of missense variants in the protein structure. GRIN = *GRIN1*, *GRIN2A*, *GRIN2B* genes; SCN = *SCN1A*, *SCN2A*, *SCN8A* genes; Mixed = electrophysiological readouts showed conflicting functional changes; PCA = Principal component analysis; WT = wild-type.

163 tests; Fig. 2A). Of those 52 residue sets, 31 (60%) were enriched for pathogenic variants also in the validation screen, comprising the remaining 30% of variants (Fig. 2B). The highest variant enrichment of pathogenic variants was observed for the set of residues that combines 'pore lining', 'located in the membrane', 'in an alpha-helix secondary structure' and 'being involved in protein–protein interactions' (OR = 34.5, $P = 3.6 \times 10^{-15}$). This residue set contains residues from 11 out of the 30 channel genes (37%), including genes encoding for gamma-aminobutyric acid (GABA) receptors, glycine receptors, N-methyl-D-aspartate (NMDA)-receptors, potassium channels and transient receptor potential cation channels.

After identification of the residue sets most strongly associated with pathogenic variants across the full dataset of 30 ion channels, we sought to identify the subset of residue sets that are consistently associated with pathogenicity across all ion channels. Since the pathogenicity of homologous residues is similar across gene families,[36] we categorized our set of ion channels into eight evolutionarily derived families and assessed the enrichment of pathogenic versus population variants for each residue set in every channel family. As to residue sets, the lowest number of residue sets that were enriched for pathogenic variants was found in the glycine receptor family ($n = 9$), while the highest number was found in the glutamate receptor family ($n = 28$). We identified four features that contributed, individually or in combination, to all residues sets enriched for pathogenic variants in all eight ion-channel families, namely 'alpha-helix secondary structure', 'localization at the pore', 'alpha-helix secondary structure' and 'coiled secondary structure'.

Spatial distance to pore is associated with variant pathogenicity across all ion channels

The previous analysis identified four features consistently and prominently contributing to variant pathogenicity across all ion-channel families. Out of these, we observed the highest enrichment of pathogenic variants in residues that are located in the membrane (OR = 4.8, $P = 1.8 \times 10^{-273}$) and in residues that are pore lining (OR = 3.8, $P = 1.8 \times 10^{-28}$). To study the spatial variant localization in association with variant pathogenicity in more detail, we

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

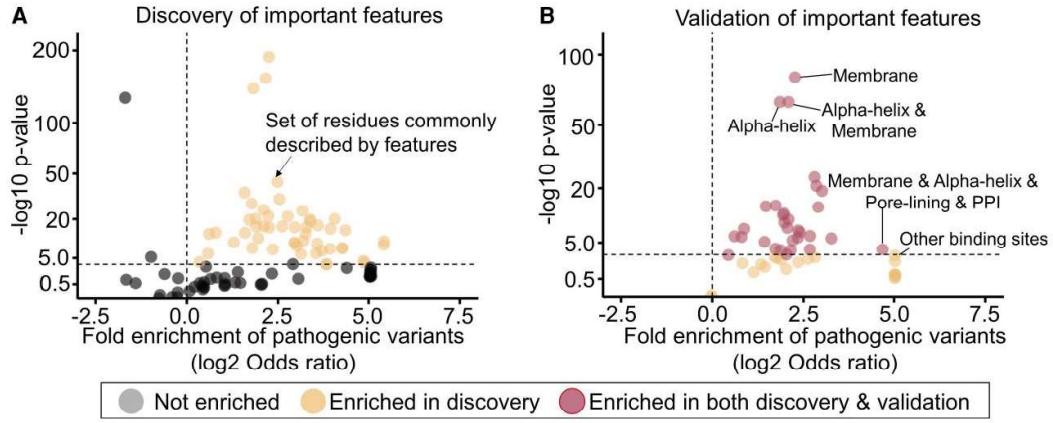


Figure 2 Identification of single features and feature combinations that are enriched for pathogenic variants. (**A**) Scatterplot of the odds ratio (*x*-axis) versus the *P*-value (*y*-axis) of the variant burden analysis of pathogenic versus population variants in each of the 163 different sets of residues described by one feature or a feature combination in the discovery cohort (Supplementary Fig. 1). Residue sets with significant enrichment of pathogenic variants after multiple-testing corrections (Fisher's exact test, OR > 1, $P < 0.05$) are displayed above the dotted line. (**B**) Scatterplot of variants in the validation cohort. Only the subset of the 55 features and feature combinations that were found to be enriched for pathogenic variants in the discovery cohort is shown. Residue sets above the dotted line are enriched for pathogenic variants in both the discovery and validation cohort.

developed a two-dimensional framework that normalizes the localization of each residue for each ion channel relative to its distance from the pore axis and the membrane centre (Fig. 3A, Supplementary material and Supplementary Table 3; also available in our GitHub repository: https://github.com/LalResearchGroup/Channel_Distances). This approach enabled us to compare the localization of variants in relation to pore and membrane distance across all ion channels, independently of their structural and sequence similarities. We identified in the combined ion-channel data set the most significant pathogenic variant enrichment in regions close to the pore axis and to the membrane centre, respectively (OR = 6.2, $P = 8.8 \times 10^{-08}$; Fig. 3B). To quantify the added value of our pore distance approach, we compared the enrichment of pathogenic variants with groups of residues defined by the normalized distances to the pore axis and their presence inside the membrane with the transmembrane (TM) domain residue annotation from UniProt. Residues closely located to the pore show a higher enrichment for pathogenic variants compared to population variants than residues covered by the UniProt TM annotation (Supplementary Fig. 4). For each ion-channel class, we observed a strong correlation between a greater relative number of pathogenic versus population variants with a closer distance to the pore axis (Pearson correlation $\rho = -0.98$, $P = 8.0 \times 10^{-07}$; Fig. 3C). In contrast, no significant correlation was found between the greater relative number of pathogenic versus population variants and the localization of the variant with respect to the membrane centre (Pearson correlation $\rho = -0.20$, $P = 1$).

The highest enrichment of pathogenic variants was observed closest to the pore axis. We next investigated the enrichment of pathogenic versus population variants in more refined subsets of residues that were located at or close to the pore, referred to as pore residues (see Supplementary material for details). We observed enrichment for pathogenic variants in all subsets of pore residues that were located inside the membrane and that form a helix or a coiled secondary structure (Fig. 3D). Among those, the highest enrichment of pathogenic variants was observed in helix forming residues that were close to the pore surface and close to the pore axis (OR = 5.6, $P = 1.44 \times 10^{-24}$). However, a similar enrichment of pathogenic variants was observed in membrane-located pore-lining residues forming alpha helices (OR = 5.5, $P = 2.6 \times 10^{-13}$; Fig. 3D), indicating that channel residues close to the pore may be similarly prone to disease-causing variants than pore-lining residues. Interestingly, the enrichment for pathogenic variants in these pore residues differed highly between channel families, having the lowest enrichment in sodium channels (OR = 2.25, $P = 2.1 \times 10^{-05}$) and the highest one in GABA receptors (OR = 54.5, $P = 3.2 \times 10^{-11}$).

Notably, we observed that the fold enrichment for pathogenic compared to population variants is up to five times higher in a subset of ion-channel genes that are depleted for missense variants ($n = 13$ genes, missense-z score[37] > 3.09; Fig. 3F). This subset comprises pore residues that are close to both the pore axis and the membrane surface, but which are not pore lining (OR = 27.3, $P = 1.9 \times 10^{-07}$). Within the set of constraint ion-channel genes, we identified a total of 116 pathogenic variants (7.1% out of 1641) in pore residues that are located inside the membrane whereas only 13 population variants (0.3% out of 4179) were found at these residue positions.

Hydrophobic pore sections contain a higher density of pathogenic versus population variants

Next, after observing that the enrichment of pathogenic versus population variants correlates with the distance from the pore, we asked whether the pathogenicity of a variant also correlates with more detailed biophysical properties of the pore. We computed seven biophysical pore properties, such as hydrophobicity, for each pore region and assigned these properties to each pore residue (see Supplementary material for details, Fig. 4A). Given the high correlation between the seven biophysical pore properties (Supplementary Fig. 5), we performed a PCA to transform them into a set of equivalent but non-correlated variables (the principal components, PCs). The first two PCs discriminated majorly hydrophobic and polar pore sections (PC1, 60.4% variance explained) and the

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

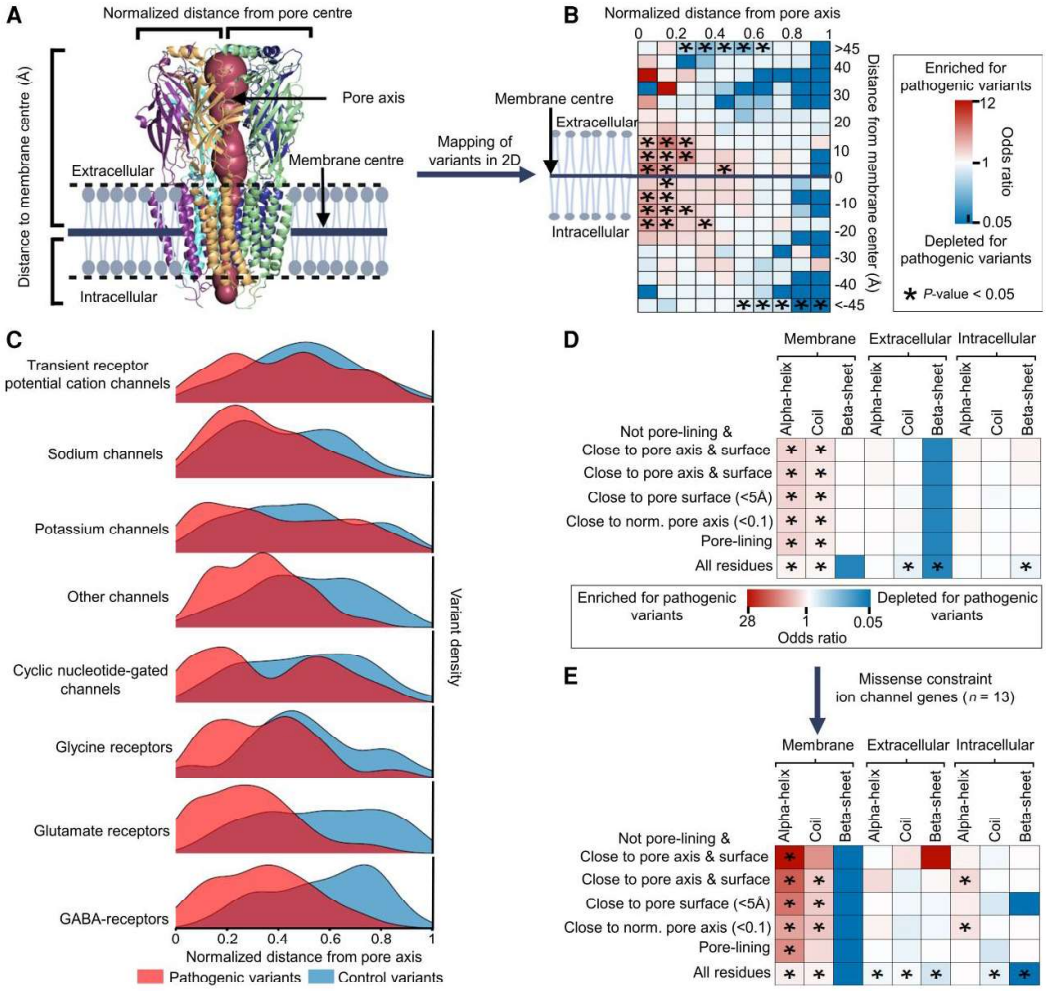


Figure 3 Residue distance from pore and membrane correlates with the pathogenic burden, harbouring the highest burden at membrane-spanning pore residues. (**A**) Graphical representation of the framework that describes the localization of each residue in ion-channel proteins on two dimensions (2D). The localization of each residue was depicted by the gene-wise normalized distance from the pore axis and the distance from the membrane centre, thereby allowing a comparison of the variant distribution across ion channels. (**B**) Table of the combined enrichment or depletion of pathogenic variants across all ion-channel proteins summarized over 200 different 2D regions. Each bin represents a distinct localization in the protein structure that is described by the normalized distance from the pore axis (*x*-axis) and the distance from the membrane centre (*y*-axis). Significant enrichments or depletions of pathogenic variants are indicated with a star. (**C**) The variant densities of pathogenic variants and population variants are displayed along with the normalized distance from the pore axis, separately for each channel class ($n = 8$). (**D**) Table of the enrichment or depletion of pathogenic variants in different sets of pore residues. Each set of pore residues was defined by their location (membrane, extracellular, intracellular) and secondary structure (*x*-axis), together with a description of which residues were considered as pore residues (*y*-axis). Significant values are indicated by a star. (**E**) Table as in D but based on a subset of channel genes that are constrained for variants in the general population ($n = 13$ genes).

pore radius and charge (PC2, 14.7% variance explained). Among the seven biophysical properties, hydrophobicity, hydropathy and water solubility made the highest contribution to the first two PCs (as indicated by the length of the projection of the arrow on the x- and y-axes in Fig. 4B). Hence, small PC1 values describe hydrophobic pore sections whereas larger PC1 values characterize a polar and water-soluble pore environment. We mapped the pathogenic and population variants along PC1 and PC2 and observed enrichment of pathogenic variants compared to population variants at hydrophobic pore sections (median PC1 pathogenic variants: −0.12, median PC1 population variants: 0.14, $P = 1.9 \times 10^{-06}$; Fig. 4C), indicating a correlation between the hydrophobicity of the pore and the pathogenicity of the pore region.

Localization of missense variants correlates with functional effect and clinical phenotype

Next, we explored whether the localization of missense variants can inform the functional effect or clinical phenotype in two independent and structurally different examples. We

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

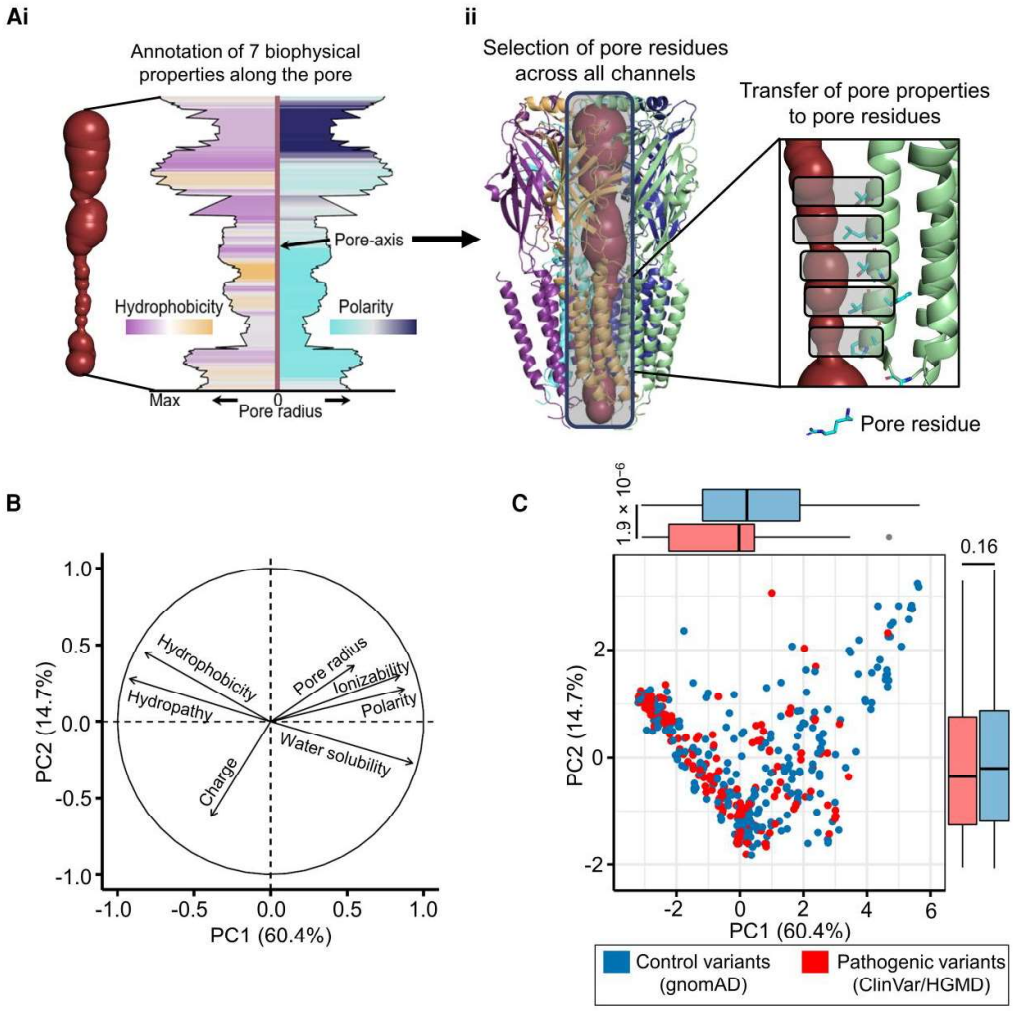


Figure 4 Biophysical pore properties discriminate pathogenic and population variants and identify residues that are most likely to harbour pathogenic variants in the hydrophobic pore sections. [**A**(i)] Graphical representation of the different biophysical pore properties for each pore section along the pore axis. [**A**(ii)] Each pore residue is assigned to the biophysical properties of its closest pore section. (**B**) Contribution of the seven considered biophysical pore properties to the first (*x*-axis, PC1) and second (*y*-axis, PC2) dimension of the PCA that was performed on the pore properties together for all ion-channel proteins. (**C**) Scatterplot along the first two dimensions of the PCA (PC1 and PC2). Each dot represents a pore residue where a pathogenic variant (ClinVar/HGMD) or population variant (gnomAD) was observed. Pore residues where population and pathogenic variant were observed are not shown.

collected 1104 and 314 pathogenic missense variants with curated clinical data across three voltage-gated sodium channel encoding genes (SCN genes: *SCN1A*, *SCN2A*, *SCN8A*) and three genes that encode NMDA receptors (GRIN genes: *GRIN1*, *GRIN2A*, *GRIN2B*), respectively. Most of the residues in sodium channels are located inside the membrane, whereas NMDA receptors have two large extracellular domains on top of their membrane-spanning part. Both ion-channel families have been associated with clinical and molecular heterogeneous neurodevelopmental disorders with epilepsy, intellectual disability and other neurological disorders.[21,38] For the six selected channels, the patient variants clustered within and close to the pore whereas population variants tended to be located at the surface of the protein (Fig. 5A and D). Pathogenic variants in the selected GRIN genes that are associated with an early seizure onset were localized closer to the pore axis than variants associated with a late onset ($P = 0.001$; Fig. 5F). Pathogenic variants in SCN genes that were observed in patients reporting intellectual disability (ID) tend to be closer to the pore ($P = 3.8 \times 10^{04}$; Fig. 5B) than pathogenic variants observed in patients without ID. A similar pattern was observed for the GRIN genes ($P = 5.8 \times 10^{-04}$; Fig. 5E).

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

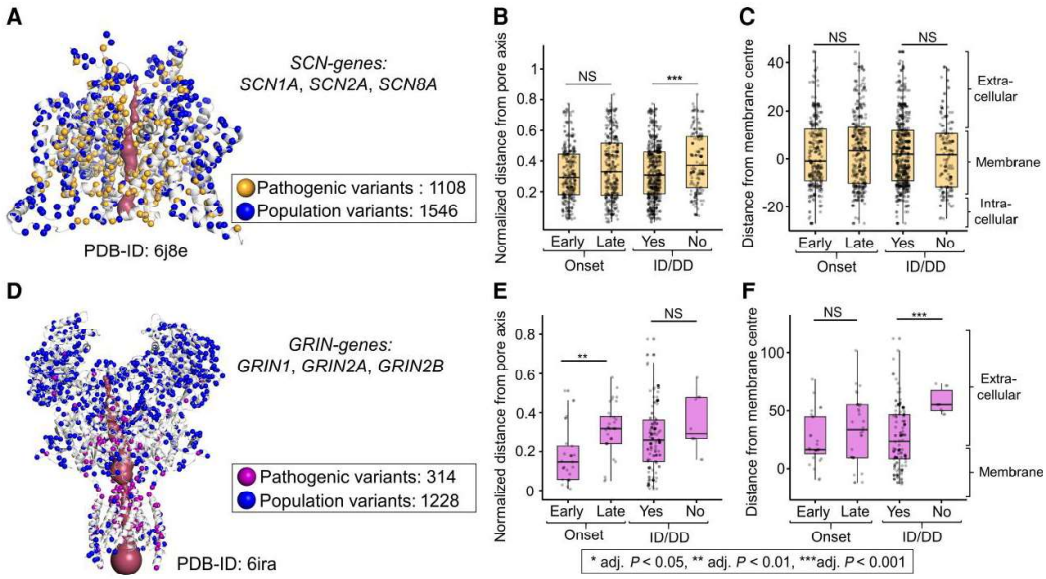


Figure 5 Distances to membrane and pore correlate with the clinical representation in sodium channels and NMDA receptors. (**A**) Nav1.2 protein structure (PDB-ID: 6j8e) encoded by *SCN2A* showing patient variants and population variants observed in *SCN1A*, *SCN2A* and *SCN8A* (SCN genes) that were aligned to *SCN2A*. Pathogenic SCN variants were curated from the SCN Portal, population variants from gnomAD. (**B**) Boxplot of the distribution of patients' variants grouped by their clinical phenotypes along with the normalized distance from the pore centre. Boxes represent patient variants that were associated with an early seizure onset (seizure onset < median onset) or a late seizure onset (seizure onset > median onset) or patient variants associated with intellectual disability (ID) or developmental delay (DD). (**C**) Boxplots showing the same groups as in **C**, but along the distance to the membrane centre. (**D**) Heterotetrameric protein complex consisting of two Glu1N and two Glu2NA subregions (PDB-ID: 6ira) encoded by *GRIN1* and *GRIN2A*, respectively. Patient and population variants observed in *GRIN1*, *GRIN2A* and *GRIN2B* (GRIN genes) are visualized on the structure. Patient variants were recruited from our internal variant database, population variants from gnomAD. Variants in *GRIN2B* were aligned to *GRIN2A* and were visualized on the GluN2A subregions. (**E** and **F**) Boxplots as in **B** and **C** of the clinical phenotypes observed in the GRIN patient cohort.

Furthermore, we investigated the relationship between the functional effects of variants that were tested in 689 electrophysiological experiments[6,22] in association with their localization in GRIN and SCN genes (Fig. 6A and D). On the basis of electrophysiological readouts variants were classified according to https://grin-portal.broadinstitute.org/ to have a GoF (combined: likely/potential-GoF), LOF (combined: likely/potential LoF), mixed (complex) or no effect. Variants in GRIN genes that were classified as GoF variants were located closer to the pore compared to variants classified as LoF (Wilcoxon rank-sum test: GRIN genes: $P = 8.0 \times 10^{-06}$; Fig. 6B and E). In contrast, no significant difference was observed in variants classified as GoF and LoF in SCN genes (Wilcoxon rank-sum test: SCN genes: $P = 0.075$). Similarly, in both gene families, GoF-associated variants were closer located to the membrane centre than variants classified as LoF (Wilcoxon rank-sum test: SCN genes: $P = 9.8 \times 10^{-11}$; GRIN genes: $P = 1.2 \times 10^{-05}$; Fig. 6C and F). In addition, functionally tested variants that showed no difference in their activity compared to the wild-type were only available for the set of GRIN genes and characterized by a larger distance from the membrane centre ($P = 6.9 \times 10^{-14}$) and pore axis ($P = 6.3 \times 10^{-13}$) than variants that were annotated with any functional difference (Fig. 6E and F). For functionally tested variants located at the pore, we further explored whether the variants of different functional effects cluster in distinct biophysical pore environments. Whereas the variants in GRIN genes were scattered along with the PCs of the biophysical pore properties (Supplementary Fig. 6A), GoF classified variants in SCN genes cluster at low PC1 values (median = −2.7) and high PC2 values (median = 1.0) indicating a hydrophobic pore environment, whereas LoF variants were predominantly present at a less hydrophobic pore and charged pore environment (PC1 values median = −0.2; PC2 values median = −0.1; Supplementary Fig. 6B).

Discussion

We performed the first approach to systematically determine the most functionally essential structural features of residues across disease-associated ion channels. Some of the identified features are conserved across evolutionary diverse ion channels. In a proof-of-concept analysis of six neurodevelopmental disorder-associated channelopathies, we show that the newly identified conserved features are correlated with variant pathogenicity, molecular function and clinical phenotype.

Spatial constraints have been shown as a powerful tool to predict pathogenicity in ion channels.[40] We observe that residues that are close to the pore, are located in the membrane, and form an alpha-helix, are most enriched for pathogenic variants across ion channels. Our results are in line with previous molecular biological studies on small sets of genes, which reported ion-channel residues at the pore to be prone to disease-causing mutations and correlate with molecular mutations.[16,41,42] For example, clusters of pathogenic variants had been found in the pore region in some specific ion channels, including voltage-gated sodium channels,[44] NMDA receptors,[41] potassium channels[16] and GABA receptors.[45] Here, we systematically show that across 163 feature combinations characterizing amino acid residues, the pore

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

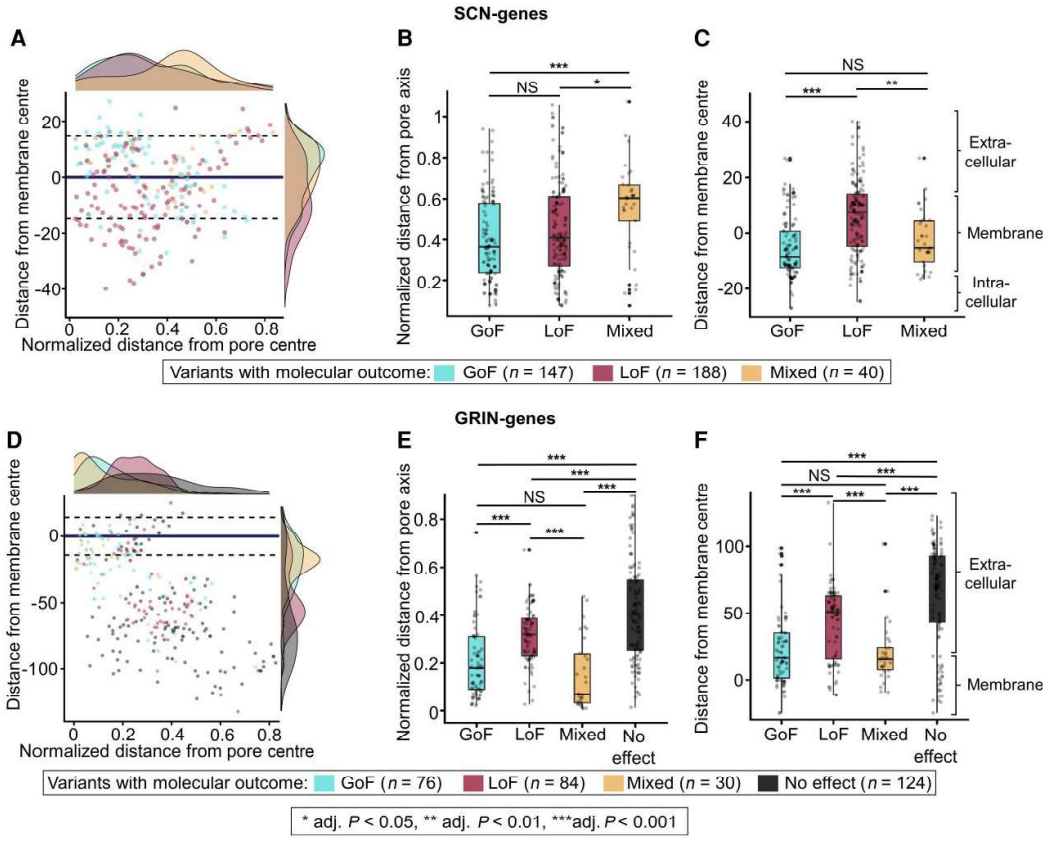


Figure 6 Distances to membrane and pore correlate with the molecular effect in voltage-gated sodium channels and NMDA receptors. (**A**) Scatterplot of variants associated with a molecular effect plotted with respect to the normalized distance to the pore axis (*x*-axis) and the distance from the membrane centre (*y*-axis). All readouts were assigned to the complementary SCN2A protein position based on a multiple sequence alignment. Variant with a mixed effect had contrary effects in different electrophysiological measurements.[39] (**B**) Box plot of the molecular effects scattered across the normalized distance from the pore axis in SCN genes. (**C**) Box plot of the molecular effects scattered across the distance from the membrane centre in SCN genes. (**D**) Scatterplot as in **A** visualizing variants associated with a molecular effect in GRIN genes. (**E** and **F**) Box plots as in **B** and **C** show the variants classified with a molecular effect along with the normalized distance from the pore axis and membrane centre in GRIN genes.

features are most important. We were able to map these most important pore features and identify for the first time a spatially conserved pattern across eight different channel families. Although there are well characterized functional domains, such as the voltage sensor in sodium channels, that are distant from the pore and known to be enriched for pathogenic variants,[42,46,47] the overall observation that variants which are spatially more distant from the pore axis are less often pathogenic, remained stable across all eight channel families. The core function of the pore-forming channel subunits is to allow a passive ion flow through the pore.[3] Residues comprising the water-filled pore have stringent requirements in terms of their precise location relative to other residues to maintain pore dimensions and allow channel opening and closing to proceed with normal stimuli.[48,49] The pore region is under strong evolutionary selection, and pathogenic variants probably alter the side chain size or characteristics of the pore and interfere with the pore function.[41,50] Our observation that pathogenic variants compared to populations variants are enriched at residues close to the pore could thus be explained by the increased likelihood of the mutations interfering with the pore function. We make our generated distance annotations available in Supplementary Table 3 as well as our GitHub repository (https://github.com/LalResearchGroup/Channel_Distances) and discuss how our annotations can be used in variant classification or research projects (see Supplementary material).

Structure-based bioinformatic methods that score physicochemical residue properties at the pore, such as hydrophobicity, have been shown to improve the prediction of the channel conformational states and to provide insights into the channel gating processes.[51] In our study, we observed that hydrophobicity showed the strongest association among biophysical properties with variant pathogenicity for pore residues across all investigated ion channels. Previous studies of the voltage-gated sodium channel $Na_V1.1$[44] and $Na_V1.7$[52] identified a correlation between pore hydrophobicity and variant pathogenicity. It has been suggested that a hydrophobic pore section can present an energetic barrier to ion permeation interrupting the channel gating, without having the pore physically closed due to de-wetting of the hydrophobic environment, a mechanism known as hydrophobic gating.[53,54] As a

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

consequence, the variants we identified to be located at the hydrophobic pore sections may not only affect the channel activity through a possible change in the pore size but also the alterations of the hydrophobic gating.[55]

We investigated the identified pathogenicity features associated with pore axis and membrane centre distance in molecular and clinical datasets. Prediction of phenotypes will guide patients and families and enable a more individualized prognosis. We showed the variant position in voltage-gated sodium channel genes (SCN genes) and NMDA receptors (GRIN genes), which are both associated with early-onset epilepsies and neurodevelopmental disorders,[47,56–59] correlates with the functional effect and clinical phenotype. Overall, our spatial distance scoring approach agrees with previously reported observations that have been made on association with specific domains. For example, we find that LoF variants in SCN genes are close to the pore axis in the extracellular side of the membrane, which is the region where the selectivity filter is located, whereas GoF variants cluster in the intracellular site of the membrane.[42] Such knowledge of the molecular functional consequences of a variant may inform treatment decisions.[60–63] In contrast to the previous analysis, which compared the functional effect across predefined protein regions such as the voltage sensor, the S5 and S6 Linker, or the selectivity filter,[42] our approach has a higher spatial resolution for association analysis. Our approach can be applied to proteins in different conformational states and to ion-channel proteins where the function of domains is not well studied.

Still, our study has several limitations. First, while hundreds of genes that encode voltage- and ligand-gated ion channels have so far been identified in humans, we limited our study to only those 30 disease-associated ion channels for which 3D structural data describing proteins and protein complexes were available. Nevertheless, our study represents the most comprehensive assessment of important features across ion channels originating from eight different protein families.[26] This diversity of ion channels that are captured in our dataset suggests that our findings may also be generalizable to other ion channels, such as aquaporins, chloride channels or piezo channels that were not part of the present study. Second, to validate generalizability and to explore gene-specific important protein features, future gene-level association analyses should be performed. Given the exponential increase in sequencing data generation, gene-level analysis should be sufficiently powered in the next few years. Third, ion channels can be observed in different conformational states, namely open, closed or inactivated conformation,[64] and this affects the location of the residues in 3D. We selected a variety of protein structures with different conformational changes, indicating that our results are probably valid for a spectrum of conformational changes. However, once human protein structures of many ion channels will be available in several conformations, our approach can be used to systematically study the structural changes and effects of the variant localization in different conformations.

We introduced a new potentially powerful approach to identify pathogenic enriched regions and identify molecular and clinical correlates across diverse ion channels. In the foreseeable future, high-quality data on protein structures and protein complexes will be available for most ion channels in the light of recent improvements in *in silico* structure prediction tools,[65] and this will allow a wider application of our approach. Further, with the availability of upcoming high-throughput mutagenesis screens for ion channels, we will refine our analysis by providing high-quality assessments of the effect of variants. Finally, our method could be expanded to other protein classes, where two features represent a horizontal and vertical axis along with the protein structure, such as membrane-located transporter, to study pathogenicity and molecular and clinical phenotypes in context variant localization.

Funding

Funding for this work was provided from the *Bundesministrium für Bildung und Forschung* (German Federal Ministry for Education and Research Treat-ION, BMBF 01GM1907D) to D.L., T.B. and P.M., the Fonds National de la Recherche Luxembourg (Research Unit FOR-2715, FNR grant INTER/DFG/21/16394868 MechEPI2) to P.M., the *Bundesministrium für Bildung und Forschung* (German Federal Ministry for Education and Research Treat-Ion2, 01GM2210B) to P.M., the Agencia Nacional de Investigación y Desarrollo (ANID, PAI77200124) of Chile to E.P., the Familie SCN2A foundation 2020 Action Potential (website: FamilieSCN2A Foundation) Grant to E.P., the Dravet Syndrome Foundation (grant number, 272016) to D.L. and the National Institute of Neurological Disorders and Stroke (NIH NINDS) (Channelopathy-Associated Epilepsy Research Center, 5-U54-NS108874) to D.L.

Competing interests

The authors report no competing interests.

Supplementary material

Supplementary material is available at *Brain* online.

References

1. Kim JB. Channelopathies. *Korean J Pediatr*. 2014;57:1–18.
2. Bajaj S, Ong ST, Chandy KG. Contributions of natural products to ion channel pharmacology. *Nat Prod Rep*. 2020;37:703–16.
3. Nayak S, Batalov S, Jegla T, Zmasek C. Evolution of the human ion channel set. *Comb Chem High Throughput Screen*. 2009;12:2–23.
4. Addis L, Virdee JK, Vidler LR, Collier DA, Pal DK, Ursu D. Epilepsy-associated GRIN2A mutations reduce NMDA receptor trafficking and agonist potency—Molecular profiling and functional rescue. *Sci Rep*. 2017;7:66.
5. Rivaud MR, Delmar M, Remme CA. Heritable arrhythmia syndromes associated with abnormal cardiac sodium channel function: Ionic and non-ionic mechanisms. *Cardiovasc Res*. 2020;116:1557–70.
6. Hansen KB, Wollmuth LP, Bowie D, *et al*. Structure, function, and pharmacology of glutamate receptor ion channels. *Pharmacol Rev*. 2021;73:298–487.
7. Allen NM, Weckhuysen S, Gorman K, King MD, Lerche H. Genetic potassium channel-associated epilepsies: Clinical review of the Kv family. *Eur J Paediatr Neurol*. 2019;24:105–16.
8. Chuang SH, Reddy DS. Genetic and molecular regulation of extrasynaptic GABA-A receptors in the brain: Therapeutic insights for epilepsy. *J Pharmacol Exp Ther*. 2018;364:180–97.
9. Catterall WA. Sodium channel mutations and epilepsy. In: Noebels JL, Avoli M, Rogawski MA, Olsen RW, Delgado-Escueta AV, eds. *Jasper's basic mechanisms of the epilepsies*. 4th edn. National Center for Biotechnology Information (US); 2012.

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

10. Klassen T, Davis C, Goldman A, *et al.* Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell.* 2011;145:1036–48.
11. Baez-Nieto D, Allen A, Akers-Campbell S, *et al.* Analysing an allelic series of rare missense variants of CACNA1I in a Swedish schizophrenia cohort. *Brain.* 2022;145:1839–53.
12. Silk M, Pires DEV, Rodrigues CHM, *et al.* MTR3D: Identifying regions within protein tertiary structures under purifying selection. *Nucleic Acids Res.* 2021;49:W438–45.
13. Kamburov A, Lawrence MS, Polak P, *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc Natl Acad Sci USA.* 2015;112:E5486–95.
14. 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; 102:415–26.
15. Reynolds C, King MD, Gorman KM. The phenotypic spectrum of SCN2A-related epilepsy. *Eur J Paediatr Neurol.* 2020;24:117–22.
16. Goto A, Ishii A, Shibata M, Ihara Y, Cooper EC, Hirose S. Characteristics of KCNQ2 variants causing either benign neonatal epilepsy or developmental and epileptic encephalopathy. *Epilepsia.* 2019;60:1870–80.
17. Johannesen KM, Liu Y, Koko M, *et al.* Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. *Brain.* 2022;145:2991–3009.
18. Strehlow V, Heyne HO, Vlaskamp DRM, *et al.* GRIN2A-related disorders: Genotype and functional consequence predict phenotype. *Brain.* 2019;142:80–92.
19. Kelly M, Park M, Mihalek I, *et al.* Spectrum of neurodevelopmental disease associated with the GNAO1 guanosine triphosphate-binding region. *Epilepsia.* 2019;60:406–18.
20. Wu RH, Tang WT, Qiu KY, *et al.* Identification of novel CSNK2A1 variants and the genotype–phenotype relationship in patients with Okur–Chung neurodevelopmental syndrome: A case report and systematic literature review. *J Int Med Res.* 2021;49: 03000605211017063.
21. Brunklaus A, Du J, Steckler F, *et al.* Biological concepts in human sodium channel epilepsies and their relevance in clinical practice. *Epilepsia.* 2020;61:387–99.
22. Brunklaus A, Feng T, Brünger T, *et al.* Gene variant effects across sodium channelopathies predict function and guide precision therapy. *Brain.* Published online 17 January 2022. doi:10.1093/brain/awac006
23. Armstrong JF, Faccenda E, Harding SD, *et al.* The IUPHAR/BPS guide to pharmacology in 2020: Extending immunopharmacology content and introducing the IUPHAR/MMV guide to malaria pharmacology. *Nucleic Acids Res.* 2020;48:D1006–21.
24. Landrum MJ, Lee JM, Benson M, *et al.* Clinvar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 2018;46:D1062–7.
25. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res.* 2000;28:235–42.
26. Yates B, Gray KA, Jones TEM, Bruford EA. Updates to HCOP: The HGNC comparison of orthology predictions tool. *Briefings in Bioinformatics.* 2021;22:bbab155.
27. Consortium TU. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506–15.
28. Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581:434–43.
29. Stenson PD, Ball EV, Mort M, *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat.* 2003;21:577–81.
30. SCN Portal. Accessed 1 December 2021. https://scn-portal.broadinstitute.org/
31. GRIN Portal. Accessed 1 December 2021. https://grin-portal.broadinstitute.org/
32. CFERV. Accessed 1 December 2021. http://functionalvariants.emory.edu/
33. Touw WG, Baakman C, Black J, *et al.* A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 2015;43:D364–8.
34. Laskowski RA, Chistyakov VV, Thornton JM. PDBsum more: New summaries and analyses of the known 3D structures of proteins and nucleic acids. *Nucleic Acids Res.* 2005;33(Suppl 1): D266–8.
35. Sehnal D, Svobodová Vařeková R, Berka K, *et al.* MOLE 2.0: Advanced approach for analysis of biomacromolecular channels. *J Cheminform.* 2013;5:39.
36. 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.
37. Lek M, Karczewski KJ, Minikel EV, *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–91.
38. Benke TA, Park K, Krey I, *et al.* Clinical and therapeutic significance of genetic variation in the GRIN gene family encoding NMDARs. *Neuropharmacology.* 2021;199:108805.
39. Brunklaus A, Schorge S, Smith AD, *et al.* SCN1A variants from bench to bedside—Improved clinical prediction from functional characterization. *Human Mutat.* 2020;41:363–74.
40. Li B, Roden DM, Capra JA, *et al.* The 3D mutational constraint on amino acid sites in the human proteome. *Nature Communications.* 2022;13:3273.
41. Li J, Zhang J, Tang W, *et al.* De novo GRIN variants in NMDA receptor M2 channel pore-forming loop are associated with neurological diseases. *Human Mutat.* 2019;40:2393–413.
42. 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:eaay6848.
43. Glazer AM, Wada Y, Li B, *et al.* High-throughput reclassification of SCN5A variants. *Am J Hum Genet.* 2020;107:111–23.
44. Zuberi SM, Brunklaus A, Birch R, Reavey E, Duncan J, Forbes GH. Genotype-phenotype associations in SCN1A-related epilepsies. *Neurology.* 2011;76:594–600.
45. Hernandez CC, Zhang Y, Hu N, *et al.* GABA A receptor coupling junction and pore GABRB3 mutations are linked to early-onset epileptic encephalopathy. *Sci Rep.* 2017;7:15903.
46. Skrenkova K, Song JM, Kortus S, *et al.* The pathogenic S688Y mutation in the ligand-binding domain of the GluN1 subunit regulates the properties of NMDA receptors. *Sci Rep.* 2020;10:18576.
47. Endele S, Rosenberger G, Geider K, *et al.* Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat Genet.* 2010;42:1021–6.
48. Schlegel AM, Haswell ES. Charged pore-lining residues are required for normal channel kinetics in the eukaryotic mechanosensitive ion channel MSL1. *Channels (Austin).* 2020;14:310–25.
49. Liao M, Cao E, Julius D, Cheng Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature.* 2013;504:107–12.
50. Pan X, Li Z, Jin X, *et al.* Comparative structural analysis of human Nav1.1 and Nav1.5 reveals mutational hotspots for sodium channelopathies. *Proc Natl Acad Sci USA.* 2021;118:e2100066118.
51. Rao S, Klesse G, Stansfeld PJ, Tucker SJ, Sansom MSP. A heuristic derived from analysis of the ion channel structural proteome permits the rapid identification of hydrophobic gates. *Proc Natl Acad Sci USA.* 2019;116:13989–95.
52. Xenakis MN, Kapetis D, Yang Y, *et al.* Hydropathicity-based prediction of pain-causing NaV1.7 variants. *BMC Bioinformatics.* 2021;22:212.

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

53. Beckstein O, Biggin PC, Sansom MSP. A hydrophobic gating mechanism for nanopores. *J Phys Chem B*. 2001;105:12902–5.
54. Bertil H. *Ionic channels of excitable membranes*. 3rd edn. Sinauer; 2001.
55. Yazdani M, Jia Z, Chen J. Hydrophobic dewetting in gating and regulation of transmembrane protein ion channels. *J Chem Phys*. 2020;153:110901.
56. Claes L, Del-Favero J, Ceulemans B, Lagae L, Van Broeckhoven C, De Jonghe P. *De novo* mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy. *Am J Hum Genet*. 2001;68:1327–32.
57. Veeramah KR, O'Brien JE, Meisler MH, *et al*. *De novo* pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am J Hum Genet*. 2012;90:502–10.
58. Matalon D, Goldberg E, Medne L, Marsh ED. Confirming an expanded spectrum of SCN2A mutations: A case series. *Epileptic Disord*. 2014;16:13–8.
59. Hamdan FF, Gauthier J, Araki Y, *et al*. Excess of de novo deleterious mutations in genes associated with glutamatergic systems in nonsyndromic intellectual disability. *Am J Hum Genet*. 2011; 88:306–16.
60. Krey I, von Spiczak S, Johannesen KM, *et al*. L-serine treatment is associated with improvements in behavior, EEG, and seizure frequency in individuals with GRIN-related disorders due to null variants. *Neurotherapeutics*. 2022;19:334–41.
61. Soto D, Olivella M, Grau C, *et al*. L-Serine dietary supplementation is associated with clinical improvement of loss-of-function GRIN2B-related pediatric encephalopathy. *Sci Signal*. 2019;12: eaaw0936.
62. Pierson TM, Yuan H, Marsh ED, *et al*. GRIN2A mutation and early-onset epileptic encephalopathy: Personalized therapy with memantine. *Ann Clin Transl Neurol*. 2014;1:190–8.
63. Wolff M, Johannesen KM, Hedrich UBS, *et al*. Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders. *Brain*. 2017;140:1316–36.
64. Yoder N, Yoshioka C, Gouaux E. Gating mechanisms of acid sensing ion channels. *Nature*. 2018;555:397–401.
65. Jumper J, Richard E, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–9.

Downloaded from https://academic.oup.com/brain/article/146/3/923/6678476 by Universitaetsbibliothek Cologne user on 29 April 2023

Supplementary Material

Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes

Tobias Brünger, Eduardo Pérez-Palma, Ludovica Montanucci, Michael Nothnagel, Rikke S Møller, Stephanie Schorge, Sameer Zuberi, Joseph Symonds, Johannes R Lemke, Andreas Brunklaus, Stephen F Traynelis, Patrick May, Dennis Lal. **Brain**.
DOI:10.1093/brain/awac305. (published)

The supplementary material is listed below and accessible at https://doi.org/10.1093/brain/awac305. Due to their large size Supplementary Table 2 and Supplementary Table 3 are only available online at https://doi.org/10.1093/brain/awac305 and https://github.com/LalResearchGroup/Channel_Distances.

Supplementary data

Application of the established pore distance thresholds

We observed that several annotations that capture residues close to the pore show greater pathogenic vs. population variant enrichment than the widely used UniProt 'transmembrane' annotation (Supplementary Fig. 4). The strongest fold enrichment (OR = 4.4, P = 2.7e-35) was observed for the group of residues with a normalized pore distance between 0.1-0.2 (see methods for detail) and also located within the membrane. We provide the computed distances to the membrane and pore axis in Supplementary Table 3 and our GitHub repository (https://github.com/LalResearchGroup/Channel_Distances) for every residue. We also made residue annotations and associated pathogenic vs. population variant enrichment effect size available. The developed annotations can be integrated into bioinformatic pipelines for variant interpretation using a combination of the genomic position, transcript ID, and amino acid position. Extensive additional annotations that can enhance variant interpretation were generated for all variants on genomic-, cDNA-, and protein-level (Supplementary Table 3). Our framework, informed by extensive annotations, identified several VUS variants as likely-pathogenic (e.g.Thr288Ala in GABRB3, Tyr481His in HCN4, and Gly313Trp in KCNQ2 classified as VUS in the ClinVar database, accessed 07/2022, with PM2 and PM5 criteria[1]). We demonstrate that all three variants additionally fulfill the PM1 criterium by being located in a region enriched with patient variants / depleted of benign variants: Inside the membrane & near the pore axis (Figure 3b, Supplementary Table 3).

Supplementary Methods

Combination of features to create residue sets

The 13 annotated features (see Methods, Supplementary Fig. 1) were arranged into five main categories: whether a residue is located at the pore, whether a residue is located in the membrane, intracellular or extracellular space, the secondary structure type of the residue, the association of a residue to a functional site (e.g. a binding site), and whether a residue is reported to be involved in any protein-protein interaction. Using the features, we created different sets of residues, where all residues within one set were described by the same feature, e.g. all residues being located in the membrane. Furthermore, we created additional

residue sets as a combination of multiple features, e.g. a set of residues being located intracellular and simultaneously forming an alpha-helix secondary structure. To this end, we considered all possible combinations of two or more features, resulting in 163 different residue sets harboring at least one residue.

Quantification of the residue location

To describe the localization of each channel residue with respect to its distance to the membrane center and pore axis, we predicted the membrane for each ion channel protein with the PPM[2] server (https://opm.phar.umich.edu/ppm_server, version PPM 2.0) and obtained the location of the non-membrane spanning part (intracellular or extracellular) from Uniprot[3]. We then calculated the distance of the C-alpha atom of each protein residue to the membrane center and assigned a positive distance to all residues oriented towards the intracellular part of the protein as well as negative distance values for residues located in the extracellular space (Fig. 3A). Residues with an assigned distance of 0 were located at the membrane center. Similarly, we used the previously predicted pore dimensions (see annotation of features above) to calculate the distance of each C-alpha atom to the pore axis and normalize the distance for each gene between 0 and 1. We then divided each protein structure into 200 2D bins that group residues together based on their distance from the pore axis and from the membrane center and analyzed the enrichment of pathogenic compared to population variants in the residues of each bin.

Calculation of the fold enrichment of pathogenic vs. population variants for different pore thresholds

We created 20 groups of residues using different pore distance thresholds and their presence inside the membrane. We split the residues into 10 normalized pore distance groups: 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, 0.9-1. Each of these distance groups was evaluated with and without additional filtering for a localization of the residues inside the membrane. We then performed a Fisher's Exact Test to test for enrichment of pathogenic vs. population variants in each of these groups of residues. Similarly, a Fisher's Exact Test was performed to obtain the enrichment of pathogenic vs. population variants in residues that were annotated in Uniprot to be located within the transmembrane region of the protein.

Annotations of different pore parts

To evaluate the variant distribution across different pore parts, we defined pore residues as 1) residues that are pore-lining as defined by Mole2.0 webserver[4] (https://mole.upol.cz/, access 03.04.2021), 2) residues that are close to the pore axis (normalized pore distance <0.1), and 3) residues that are close to the pore surface (pore surface distance <5A). To obtain the distance to the pore surface, we considered, as before, C-alpha atoms and calculated the closest distance of each residue to the pore surface. We then combined these pore annotations with the descriptions of the secondary structure and the general localization (membrane, intracellular, extracellular).

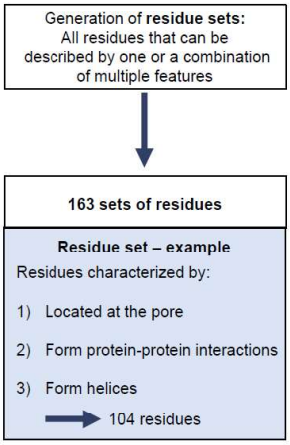
We defined a pore residue as a residue that is either pore-lining, close to the pore center (normalized pore distance <0.1), or close to the pore surface (pore surface distance < 5). Important pore properties should be enriched for patient variants and depleted for population variants. To test this enrichment of pathogenic variants we assigned to each of the pore residues seven biophysical properties annotated with the Mole webserver[4]: pore radius, lipophilicity, hydropathy, hydrophobicity, polarity, water-solubility, and charge. Enrichment of patient variants over population variants was tested using Fisher's Exact test.

Supplementary Figures

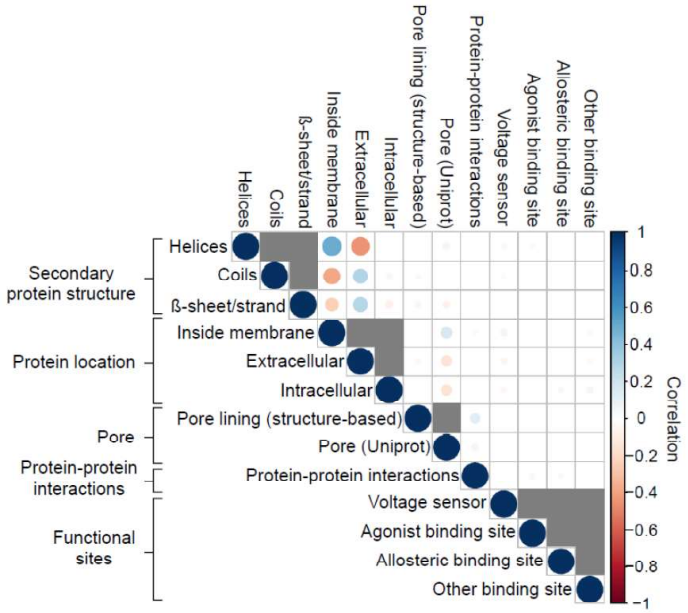
A)

Feature category	Feature	Source for collection or calculation
Protein location	Extracellular	Uniprot
	Intracellular	Uniprot
	Membrane	PPM webserver
Pore	Pore-lining (structure-based)	Mole 2.0
	Pore (Uniprot)	Uniprot
Protein-Protein Interactions	Protein-Protein Interactions	PDBsum
Secondary protein structure	ß-strand/sheets	DSSP
	helices	DSSP
	coils	DSSP
Functional sites	Voltage sensor	Uniprot
	Allosteric binding site	Uniprot
	Agonist binding site	Uniprot
	Other binding site	Uniprot

B)

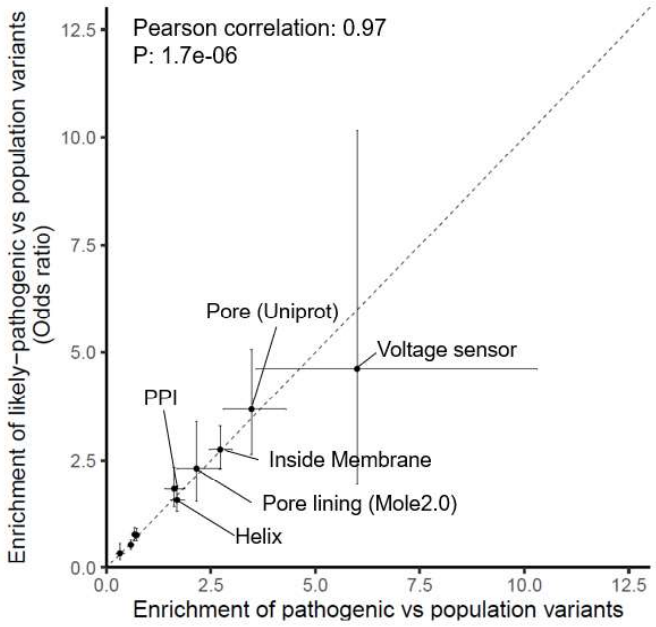


Supplementary Figure 1: Feature-defined residue sets. A) List of the 13 considered features along with their feature category (column one) and their source of annotation (column three). **B)** The features were used to generate sets of residues, where all residues are commonly characterized by a single feature or a combination of two or multiple features of different categories. An example for one set of residues is displayed.

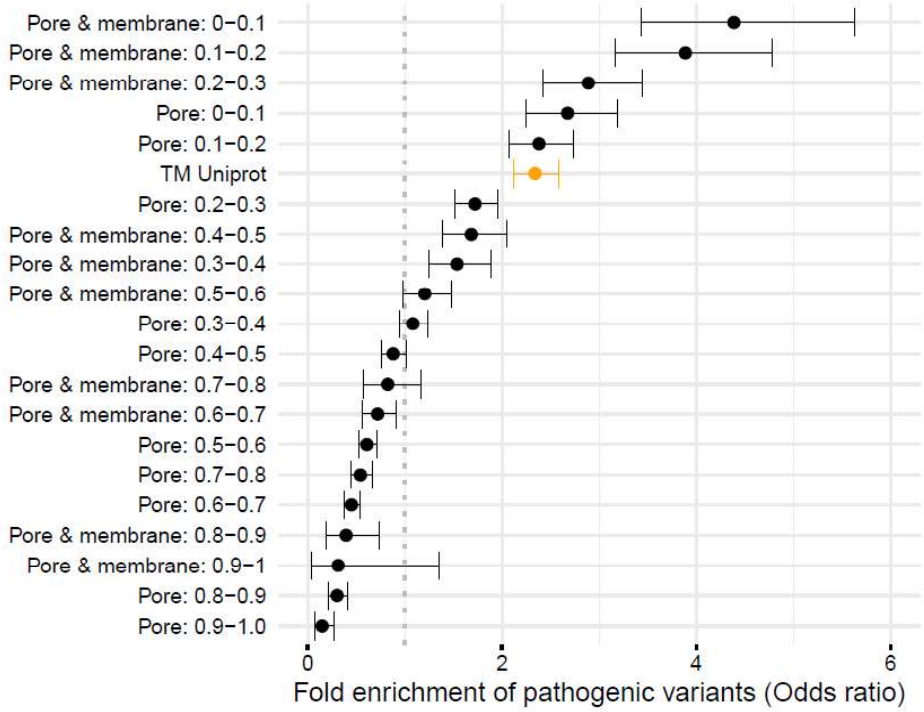


Supplementary Figure 2: Correlation of the residue features. The matrix depicts the correlations between the individual residue features. Correlations between features within the same feature category are not computed because they have no biological sense (for example "being in the extracellular region" and at the same time "being in the intracellualar region") and their corresponding boxes are colored in gray (see Supplementary Figure 1, Methods).

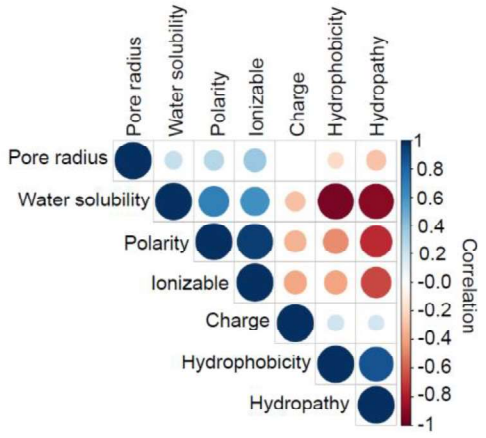
Larger spheres indicate a higher correlations. Positive correlations are shown in blue, negative correlations are shown in red. Only significant correlations (p-value < 0.05) are shown.



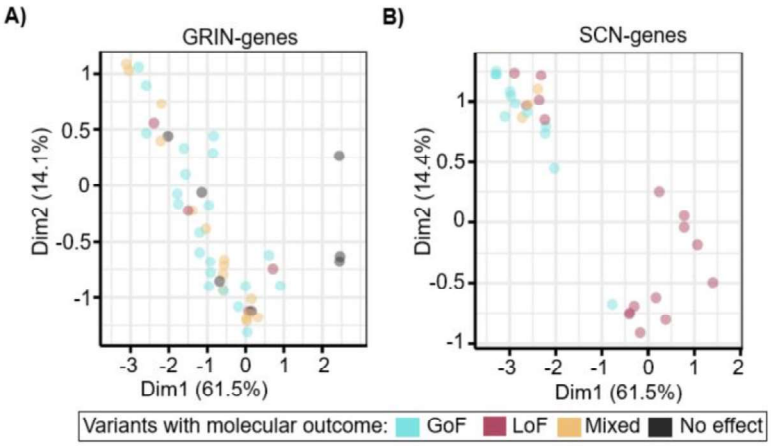
Supplementary Figure 3: Concordance of the effect sizes of pathogenic and likely pathogenic variants in nine out of 13 features that describe the localization of residues in functional regions (see Supplementary Methods). The odds ratio of the enrichment of pathogenic variants from ClinVar and HGMD over population variants (Fisher's Exact Test) is displayed on the x-axis. Similarly, the odds ratios of the enrichment of likely-pathogenic variants from ClinVar over population variants from gnomAD are displayed on the y-axis. Of the previously 13 annotated features, only nine features were included which were captured by >10 likely-pathogenic/pathogenic variants. The Pearson correlation between the enrichment of pathogenic and likely-pathogenic variants over population variants is displayed on top of the graph.



Supplementary Figure 4: Fold enrichment of pathogenic vs. population variants for different pore thresholds and the Uniprot 'transmembrane' annotation. Residues were grouped based on the normalized distances to the pore axis (Pore) and their presence inside the membrane (membrane) using our 3D annotations. For the normalized pore distance we split the positions in 10 distance groups: 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, 0.9-1. Each of these distance groups was tested with and without additional filtering for membrane location of residues (Pore & membrane, see Supplementary Methods for details). For example, the group 'Pore: 0.1-0.2' contains residues that have a normalized pore distance between 0.1 and 0.2. Additionally, we tested the residues annotated as 'transmembrane' by Uniprot (TM Uniprot, orange color).



Supplementary Figure 5: Correlation of the biophysical pore properties. The matrix shows the correlation between the individual pore properties. Larger spheres indicate a higher correlation. Positive correlations are shown in blue, and negative correlations are shown in red. Only significant correlations (p-value < 0.05) are shown.



Supplementary Figure 6: Molecular effects at pore residues differ across pore properties. A) Scatterplot of the molecular effects in SCN genes along the first two dimensions of the PCA of the pore properties that have been performed across all ion channels (Figure 4). Variants with available electrophysiological readouts have been collected across all sodium channels and assigned to the Nav1.2 subunit, which is encoded by *SCN2A*, using a multiple sequence alignment. **B)** Scatterplot as in A, of the molecular effects in pore residues that have been collected for *GRIN2A, GRIN2B* and *GRIN1*.

References

1. 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).

2. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* **40**, D370-376 (2012).

3. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2019).

4. Sehnal, D. *et al.* MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *J Cheminform* **5**, 39 (2013).

Supplementary Table 1: List of all voltage-gated and ligand-gated ion channel encoding genes used in this study. Only those ion channels for which a structure of the human protein is available in the Protein Data Bank have been selected. The representative structure (PDB-ID, see methods for details) and the missense-z-score is provided for each gene.

Genes	PDB-ID	Missense-z-score (gnomAD database)	Resolution(Å)
CACNA1G	6kzp	3.1	4.64
CHRNA4	6cnj	3.7	0.34
CHRNB2	6cnj	3.7	2.1
GABRA1	6huj	3	3.15
GABRA5	6a96	3.5	3.23
GABRB2	6d6u	3.9	3.4
GABRB3	6i53	3.2	3.39
GABRG2	6hug	3.1	2.99
GLRA1	2m6i	NMR	1.42
GRIN1	6ira	4.5	6.22
GRIN2A	6ira	4.5	2.83
HCN1	5u6p	3.51	3.66
HCN4	6gyo	3.4	2.75
KCNK3	6rv3	2.9	3.55
KCNK4	4wfg	3	0.75
KCNMA1	6v3g	4	5.06
KCNQ1	6uzz	3.1	1.83
KCNQ2	7cr7	3.7	4.04
MCOLN1	5wj9	3.5	1.61
PKD2	5mkf	4.2	0.27
SCN2A	6j8e	3	6.46
SCN4A	6agf	3.2	1.56
SCN9A	6j8g	3.2	1.06
TRPA1	6pqo	2.9	0.04
TRPC6	6uza	3.1	2.12
TRPM4	5wp6	3.8	0.97
TRPV3	6mhw	4	0.17
SCN1A	7dtd	3.3	6.04
SCN5A	7dtc	3.3	2.75
GRIN2B	7eu8	4.1	5.42

Supplementary Table 4: Enrichment of patient and over population variants in nine features for three different groups of patient variants. Either variants classified as pathogenic, variants classified as likle-y pathogenic or the combined set of variants were considered. lCl: lower confidence Intervall, uCl: upper confidence intervall, PPI: Protein-Protein Interactions

Feature	Variant group	Odds ratio	lCl	uCl	P-value
Beta-sheet	Pathogenic	0.3	0.2	0.4	7.00E-20
Beta-sheet	Likely-pathogenic	0.3	0.2	0.6	2.70E-06
Beta-sheet	Pathogenic + likely-Pathogenic	0.3	0.3	0.4	1.30E-20
Coil	Pathogenic	0.7	0.6	0.8	6.10E-10
Coil	Likely-pathogenic	0.8	0.6	0.9	4.80E-03
Coil	Pathogenic + likely-Pathogenic	0.7	0.7	0.8	9.10E-11
Cytosol	Pathogenic	0.6	0.5	0.6	5.50E-24
Cytosol	Likely-pathogenic	0.5	0.4	0.7	2.40E-10
Cytosol	Pathogenic + likely-Pathogenic	0.6	0.5	0.6	1.60E-25
Extracellular	Pathogenic	0.7	0.6	0.8	1.10E-12
Extracellular	Likely-pathogenic	0.8	0.6	0.9	1.10E-02
Extracellular	Pathogenic + likely-Pathogenic	0.7	0.6	0.8	2.20E-13
Helix	Pathogenic	1.7	1.5	1.9	1.80E-24
Helix	Likely-pathogenic	1.6	1.3	1.9	5.50E-07
Helix	Pathogenic + likely-Pathogenic	1.7	1.5	1.9	3.50E-26
Membrane	Pathogenic	2.7	2.5	3	1.40E-87
Membrane	Likely-pathogenic	2.8	2.3	3.3	5.00E-30
Membrane	Pathogenic + likely-Pathogenic	2.7	2.4	3	1.70E-93
Pore-lining (structure)	Pathogenic	2.2	1.7	2.7	5.20E-10
Pore-lining (structure)	Likely-pathogenic	2.3	1.5	3.4	5.20E-05
Pore-lining (structure)	Pathogenic + likely-Pathogenic	2.2	1.7	2.7	6.90E-11
Pore-lining (Uniprot)	Pathogenic	3.5	2.8	4.3	2.70E-31
Pore-lining (Uniprot)	Likely-pathogenic	3.7	2.6	5.1	2.50E-13
Pore-lining (Uniprot)	Pathogenic + likely-Pathogenic	3.4	2.8	4.2	1.20E-32
PPI	Pathogenic	1.6	1.4	1.9	8.40E-11
PPI	Likely-pathogenic	1.8	1.4	2.3	1.40E-06
PPI	Pathogenic + likely-Pathogenic	1.6	1.4	1.9	1.10E-11
Voltage-sensor	Pathogenic	6	3.6	10.3	9.40E-14
Voltage-sensor	Likely-pathogenic	4.6	2	10.2	3.40E-04
Voltage-sensor	Pathogenic + likely-Pathogenic	5.7	3.5	9.8	1.10E-13

6.4 Delineation of functionally essential protein regions for 242 neurodevelopmental genes

Sumaiya Iqbal[+], **Tobias Brünger[+]**, Eduardo Pérez-Palma, Marie Macnee, Andreas Brunklaus, Mark J Daly, Arthur J Campbell, David Hoksza, Patrick May, Dennis Lal. **Brain**. 10.1093/brain/awac381. (published)

[+]These authors contributed equally to the work

https://doi.org/10.1093/brain/awac381 BRAIN 2023: 146; 519–533 | 519

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
© The Author(s) 2022. Published by Oxford University Press on behalf of the Guarantors of Brain.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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 variants 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,

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

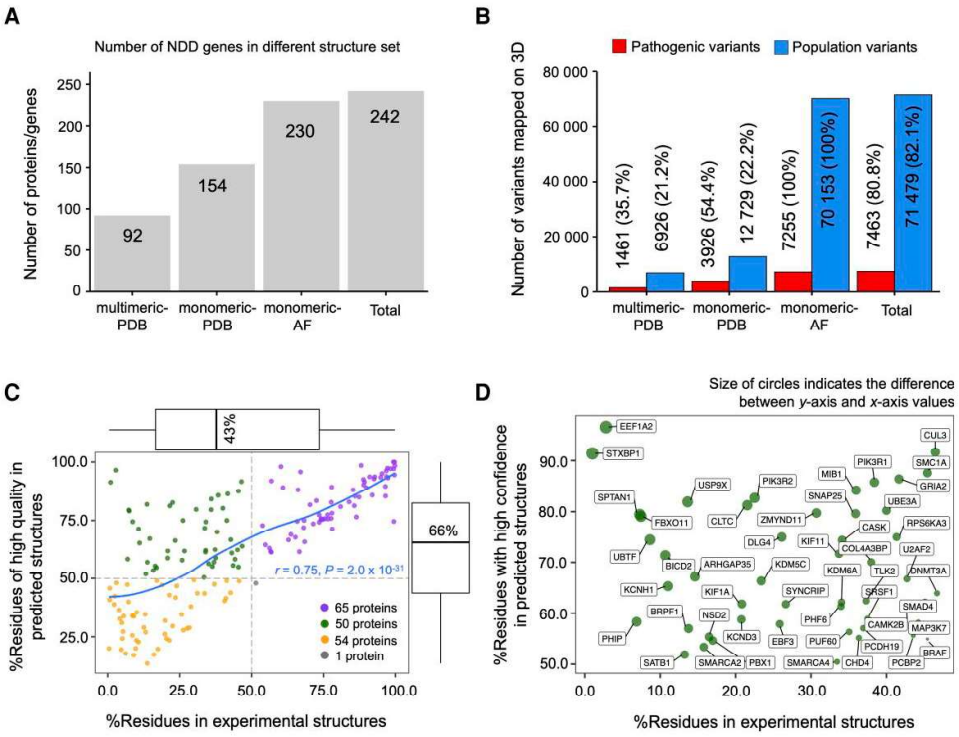


Figure 1 Overview of structure and variant level data in our dataset and related statistics. (**A**) Number of NDD genes with available protein structures in three different structure sets, and in total 3D structures. (**B**) Number of population (gnomAD) and pathogenic (ClinVar and HGMD) variants that 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

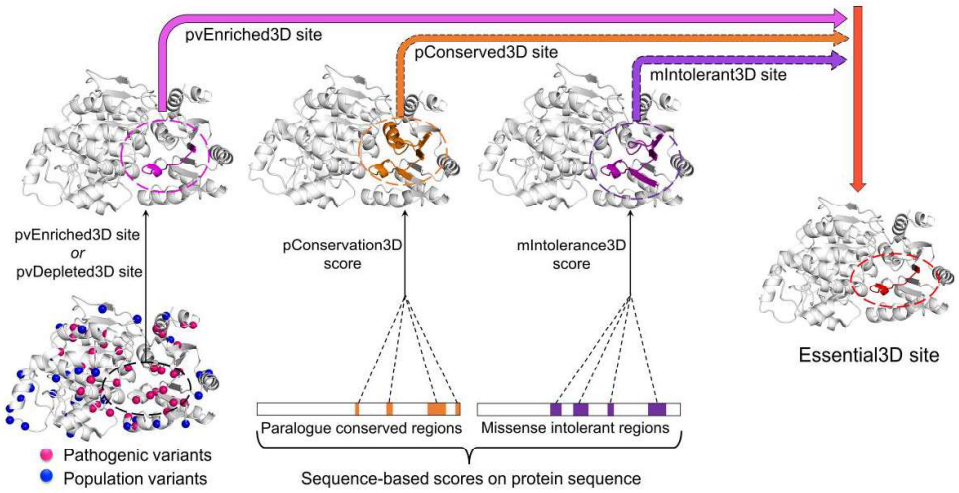


Figure 2 Schematic overview of the method to identify Essential3D sites. *Left*: Pathogenic (source: ClinVar and HGMD databases) and population (source: gnomAD database) missense variants are mapped onto protein structures. For each residue, a burden analysis (two-sided Fisher's exact test) is performed 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

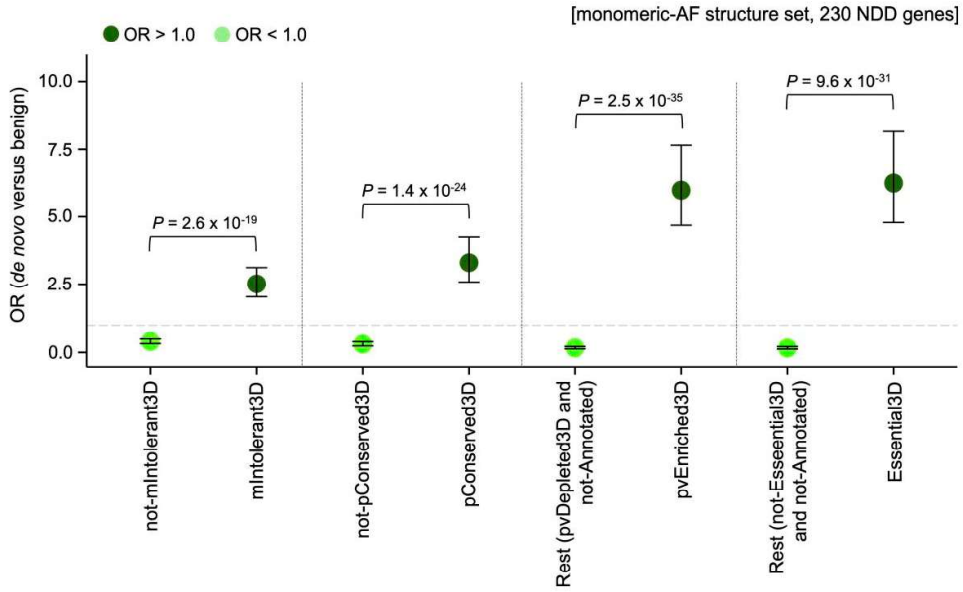


Figure 3 Residues annotated as Essential3D sites by a consensus approach are more enriched in *de novo* variants compared to the individual criterion, used to derive the consensus. OR values (*y*-axis) were calculated using two-sided Fisher's exact test with 432 NDD-associated *de novo* variants and 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

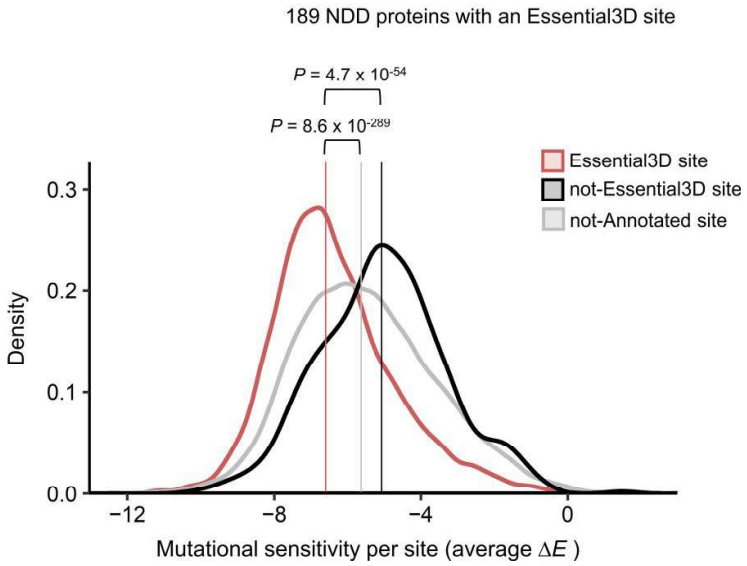


Figure 4 Essential3D sites highlight positions in proteins with high mutational sensitivity (average Δ*E*) as quantified by EVmutation.[46] Average Δ*E* is the mean change in energy for substituting the reference amino acid in the wild type protein with 19 alternate amino acids. The median mutational 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, $P < 4.7 \times 10^{-51}$; 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.

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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 Essential3D 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

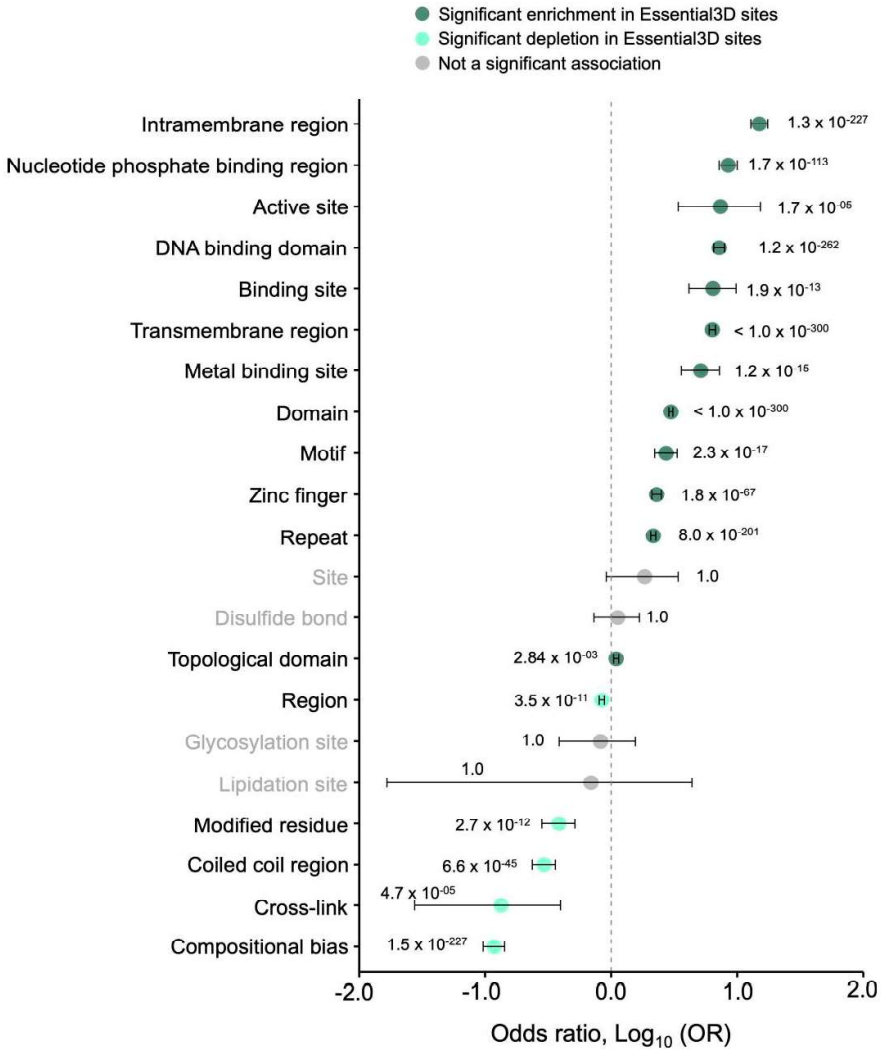


Figure 5 Association of Essenetial3D sites with function features from the UniProt database. The plot shows the results of two-sided Fisher's exact tests of association between Essential3D sites and the remaining protein residues with the features for the 189 NDD proteins with an Essential3D site. The 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% *CI*s. 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

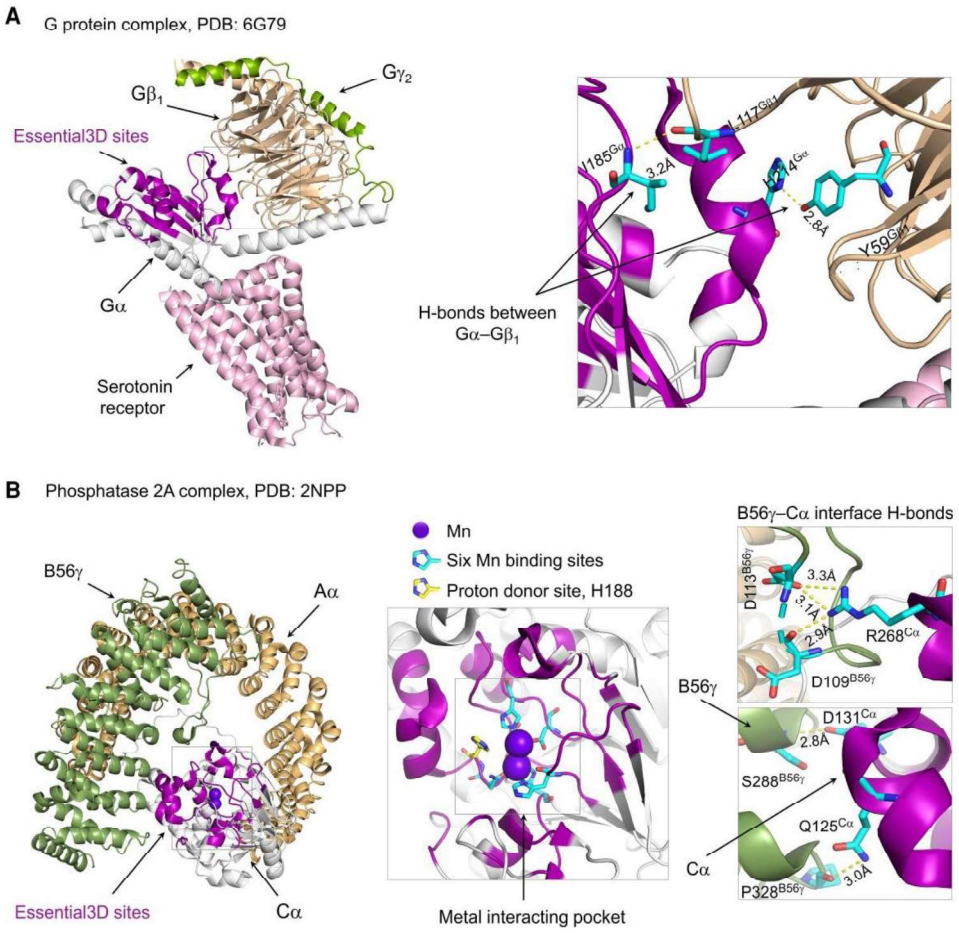


Figure 6 Essential3D site annotates important protein-protein interactions in protein complexes. (**A**, *left*) In the G-protein complex structure comprising multiple NDD genes (*GNAO1*: Gα subunit, *GNB1*: $G\beta_1$ subunit), Essential3D sites highlight two potential mechanisms: Gα-Gβ1 and Gα-GTP 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.

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

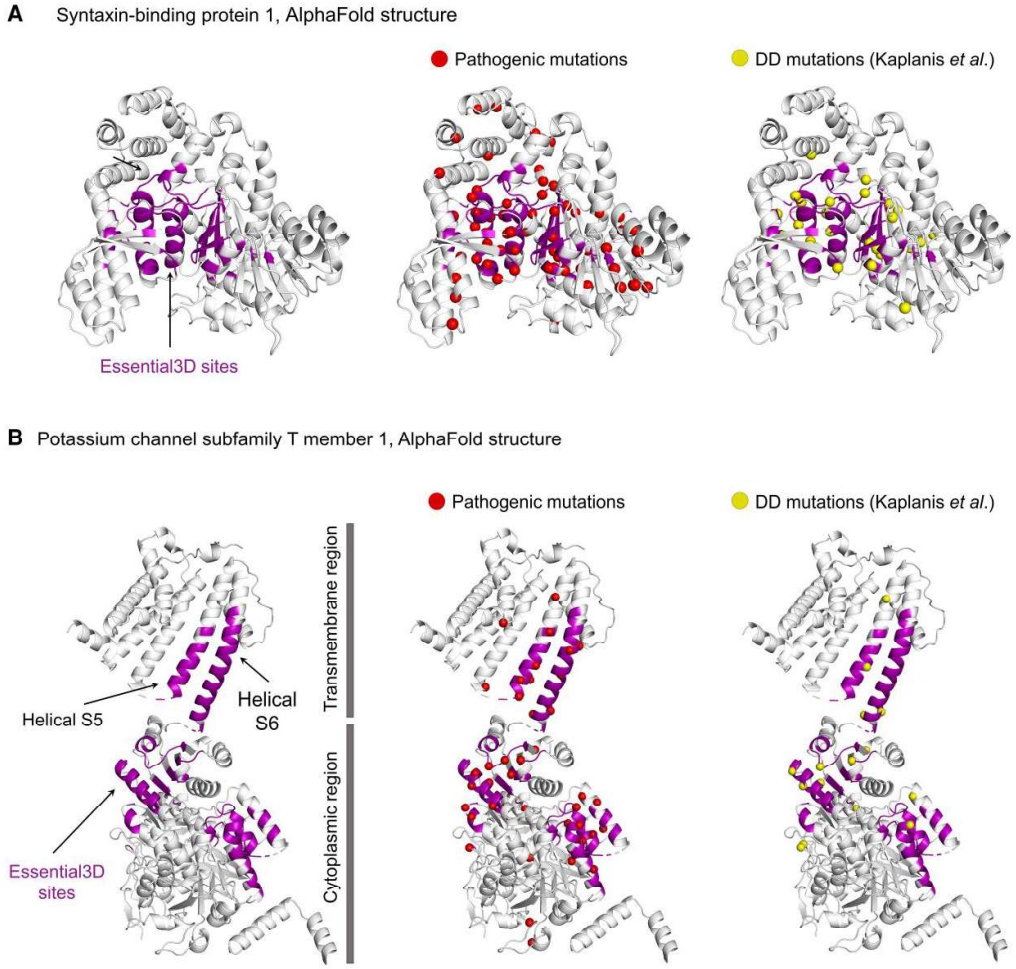


Figure 7 Use of AlphaFold-predicted structures enabled identification of Essential3D sites in NDD proteins with little to no experimental structure available and these Essential3D sites capture *de novo* DD-associated variants. (**A**, *left*) Predicted structure of syntaxin-1 protein (*STXBP1*) with 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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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.

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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.

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

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.

Downloaded from https://academic.oup.com/brain/article/146/2/519/6763201 by Bibliothek des Staatswissenschaftlichen Seminars der Universitaet zu Koeln user on 29 April 2023

Supplementary material

Delineation of functionally essential protein regions for 242 neurodevelopmental genes

Sumaiya Iqbal[+], **Tobias Brünger[+]**, Eduardo Pérez-Palma, Marie Macnee, Andreas Brunklaus, Mark J Daly, Arthur J Campbell, David Hoksza, Patrick May, Dennis Lal. **Brain**. 10.1093/brain/awac381. (published)

[+]These authors contributed equally to the work

The supplementary material is listed below and accessible at https://doi.org/10.1093/brain/awac381. Due to its large size Supplementary Table 1 and Supplementary Table 2 are only available online at https://doi.org/10.1093/brain/awac381. Precalculated annotation can be obtained from our GitHub repository https://github.com/dlal-group/Essential3D_NDD as well as from our web tool that is available at https://es-ndd.broadinstitute.org/index.html.

Supplemental file for

Delineation of functionally essential protein regions for 242 neurodevelopmental disorders

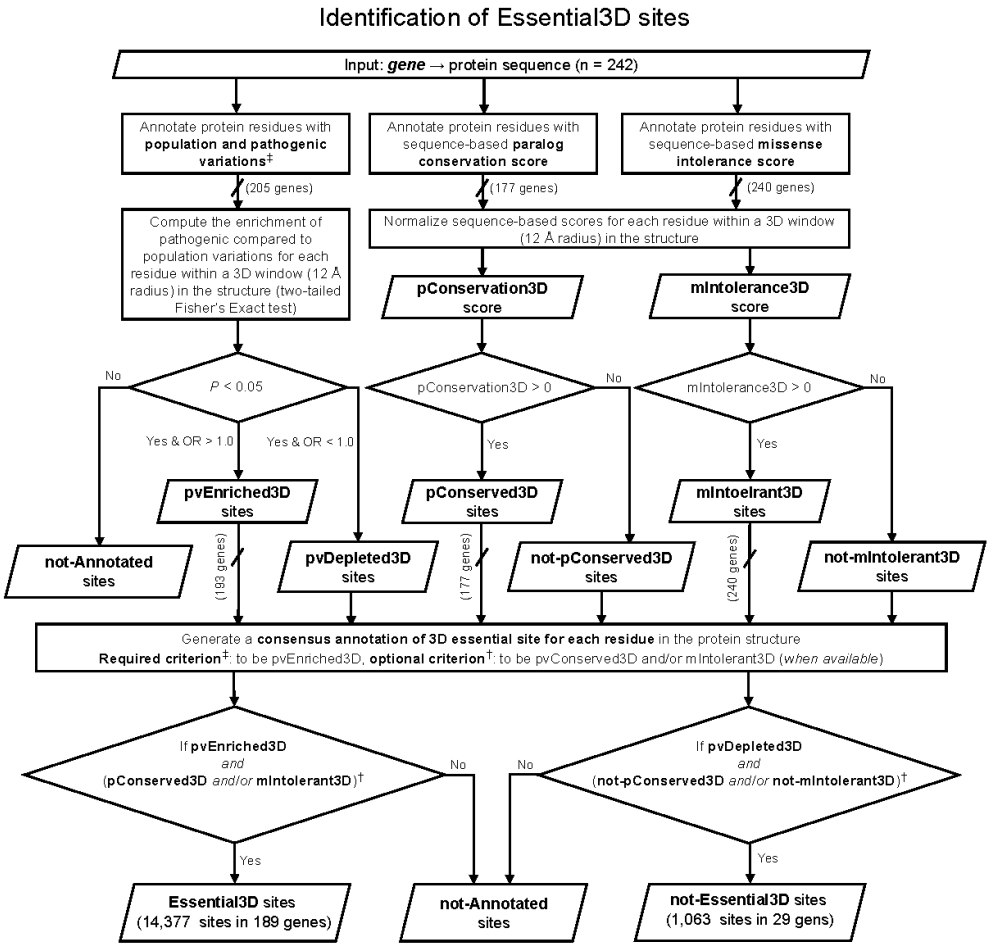
Sumaiya Iqbal[1,2,3,†,*], Tobias Brünger[4,†], Eduardo Pérez-Palma[5], Marie Macnee[4], Andreas Brunklaus[6,7], Mark J. Daly[2,3,8], Arthur J. Campbell[1,2], David Hoksza[9], Patrick May[10], Dennis Lal[2,4,11,12,*]

[†]**These authors contributed equally to this work.**

[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, Cologne, Germany
[5] Universidad del Desarrollo, Centro de Genética y Genómica, Facultad de Medicina Clínica Alemana.
[6] The Paediatric Neurosciences Research Group, Royal Hospital for Children, Glasgow, United Kingdom.
[7] Institute of Health and Wellbeing, University of Glasgow, UK.
[8] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00100 Helsinki, Finland
[9] Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[10] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[11] Epilepsy Center, Neurological Institute, Cleveland Clinic, Cleveland, USA
[12] Genomic Medicine Institute, Lerner Research Institute Cleveland Clinic, Cleveland, USA

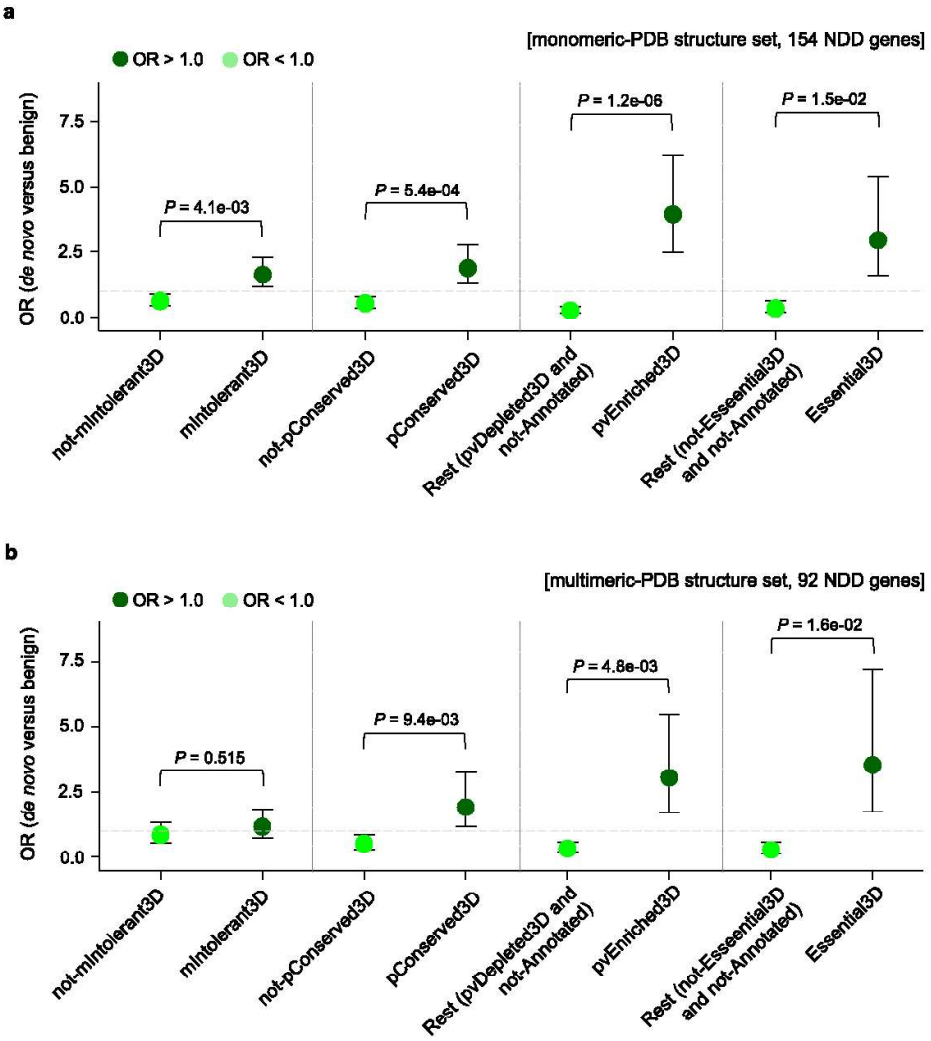
*Corresponding authors (sumaiya@broadinstitute.org, dlal@broadinstitute.org)
[+]these authors contributed equally to this work

Supplementary Figures

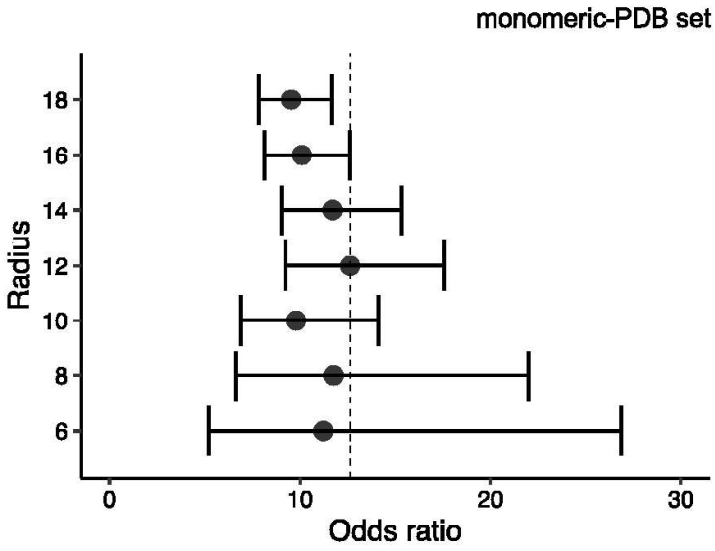


Supplementary Figure 1: Flowchart of the algorithm to identify Essential3D sites. Parallelograms indicate data (input/output); rectangles indicate processes, with arrows labelled with the number of genes that have data after execution of the process; diamonds indicate conditions, with arrows labelled with the output of the decision (Yes/No) and additional conditions (if any). '‡': To be a pvEnriched3D site is a required criterion for a residue to be Essential3D site. '†': To be a pConserved3D and/or mIntolerant3D site is an optional criterion for a residue to be Essential3D site, *i.e.*, when pConservation3D or mIntolerance3D score is not available, the corresponding criterion is not considered in the algorithm. For 242 NDD genes, an experimental or predicted structure of the corresponding protein was available. Out of them, we identified 82,747 pConserved3D sites in 177 genes, 112,777 mIntolerant3D sites in 240 genes, 19,343 pvEnriched3D sites in 193 genes, and 14,377 Essnetial3D sites in 189 NDD genes.

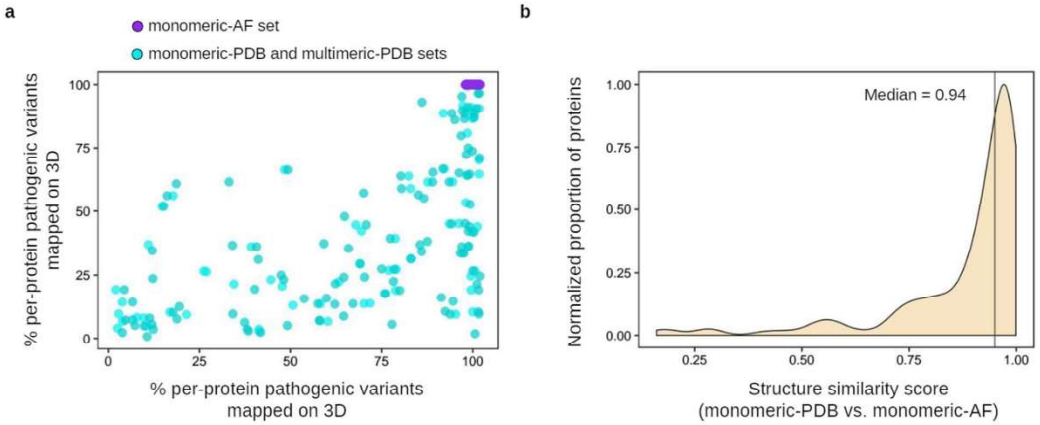
Glossary of terms: Essential3D – a 3D essential site, not-Essential3D – not a 3D essential site, not-Annotated – not annotated as Essential3D or not-Essenetial3D site; pvEnriched3D – a pathogenic variant enriched 3D site, pvDepleted3D – a pathogenic variant depleted 3D site; pConservation3D – paralog conservation 3D score, pConserved3D – a paralog conserved 3D site, not-pConserved3D – not a paralog conserved 3D site; mIntolerance3D – missense intolerance 3D score, mIntolerant3D – a missense intolerant 3D site, not-mIntolerant3D – not a missense intolerant 3D site.



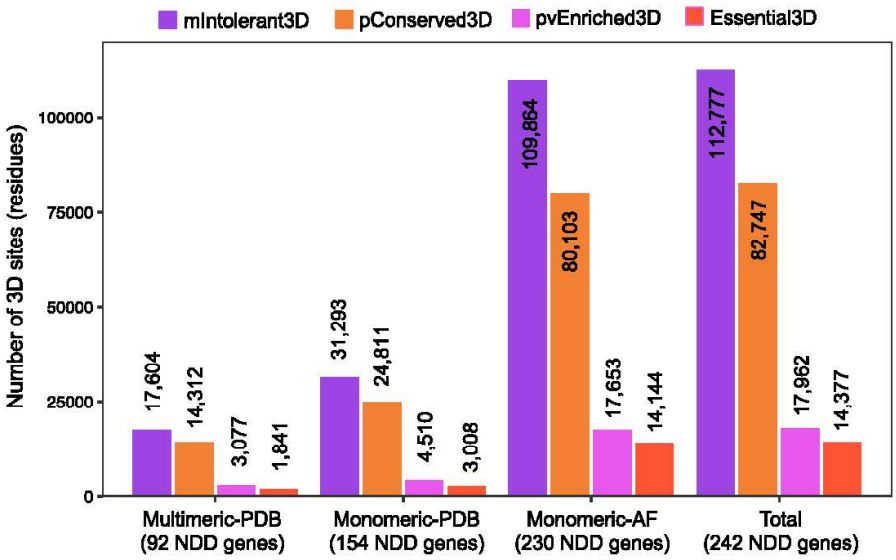
Supplementary Figure 2: Consensus annotation of Essential3D sites improves prioritization of disease mutations over single criteria in monomeric and complex experimental structures. a, OR values (*y*-axis) were calculated using two-sided Fisher's Exact test with 163 NDD-associated *de novo* variants and 1,874 benign variants in 154 NDD genes (monomeric-PDB structure set). **b,** OR values (*y*-axis) were calculated using two-sided Fisher's Exact test with 93 NDD-associated *de novo* variants and 1,043 benign variants in 92 NDD genes (multimeric-PDB structure set). Point colors indicate OR > 1.0 (enrichment, *dark green*) and OR < 1.0 (depletion, *light green*). Error bars indicate 95% confidence intervals and *P* values represent the significance of test results. The horizontal dashed line represents OR = 1.0.



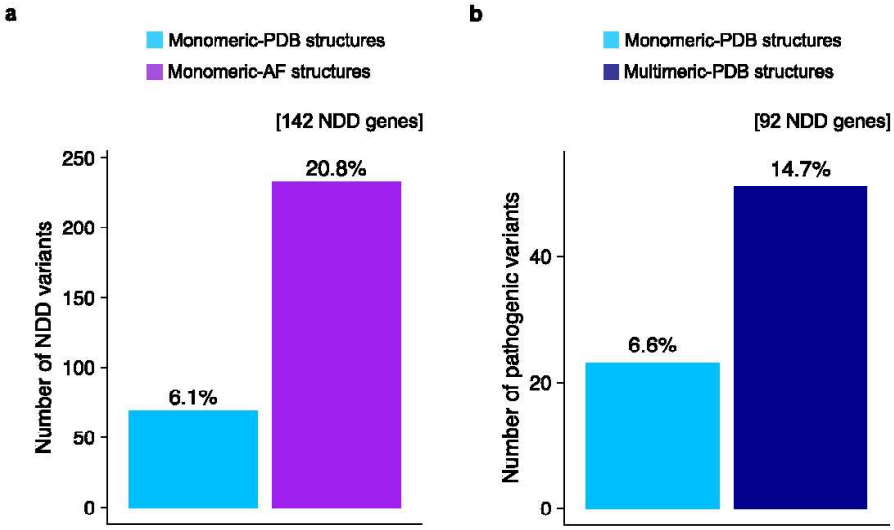
Supplementary Figure 3: Selection of the radius (*r*) of 3D window. The plot shows the results of two-sided Fisher's Exact test of enrichment of pathogenic versus population variants within the 3D window of Essential3D sites. 50% of the available pathogenic (ClinVar and HGMD) and population (gnomAD) variants were randomly selected for generating Essential3D site annotation and the enrichment test was performed using the remaining 50% of the data. The value of *r* was tuned in between 6 to 18 Å to find the value of *r* that maximizes the enrichment. Circles indicate the odds ratio of enrichment, and error bars indicate 95% confidence intervals. The vertical dashed line represents the *r* value of 12 Å resulting the highest enrichment (P = 4.3e-74). The analysis was performed using the experimental structure in the monomeric-PDB set.



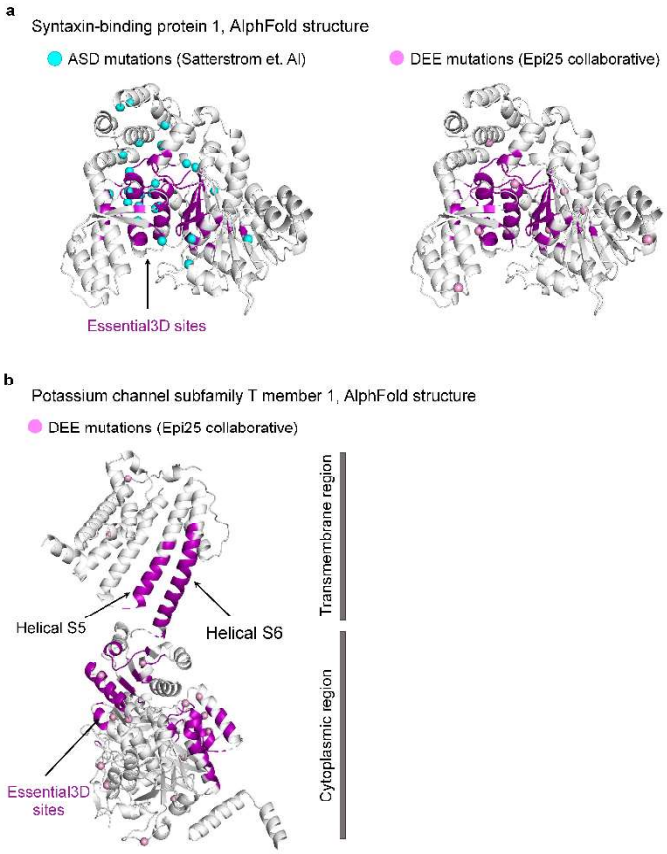
Supplementary Figure 4: Comparison of predicted structures of NDD proteins from AlphaFold (AF) database with their corresponding experimental structures from Protein Data Bank. a, Comparison of per-protein fraction of pathogenic (y-axis) and population variants (x-axis) mappable onto experimental (monomeric-PDB and multimeric-PDB set, cyan) and AlphaFold-predicted structures (monomeric-AF set, purple). The average ratio of per-protein mappable pathogenic to population variants for experimental structures was 3.1, while that for AlphaFold-predicted structures was 1.0, as the predicted structures covered full proteins. **b,** Distribution of pair-wise structural similarity between experimental structures (monomeric-PDB set) and corresponding predicted structures (monomeric-AF). The similarity score was quantified by superimposing the experimental and predicted structures per protein using TM-align. The median structural similarity was 0.94; to compare, the authors' recommended threshold for two structures being similar and adopting the same fold is 0.5.



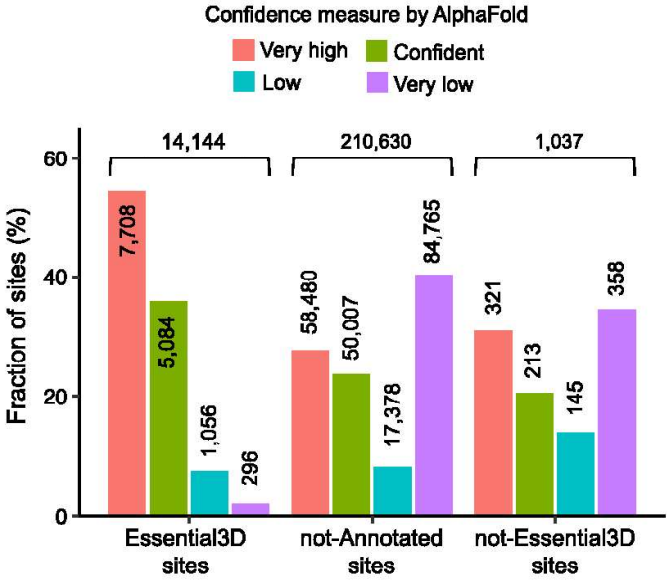
Supplementary Figure 5: Number of mIntolerant3D, pConserved3D, pvEnriched3D, and Essential3D sites in different structure sets. Multimeric-PDB and monomeric-PDB sets include experimental structures of proteins corresponding to 92 and 154 NDD genes, respectively, and the monomeric-AF set includes predicted structures for 230 NDD proteins/genes. For 12 genes, only experimental structures were analyzed (*see* **Methods** for structure collection and filtering steps), amounting to a total of 242 unique genes for which Essential3D sites were identified. Bars are labeled with number of identified sites of different types. The consensus annotation of Essential3D sites of residues were derived based on three lines of evidence, i.e., pvEnriched3D, pConserved3D, mIntolerant3D (see annotation methodology in **Supplementary Fig. 1**).



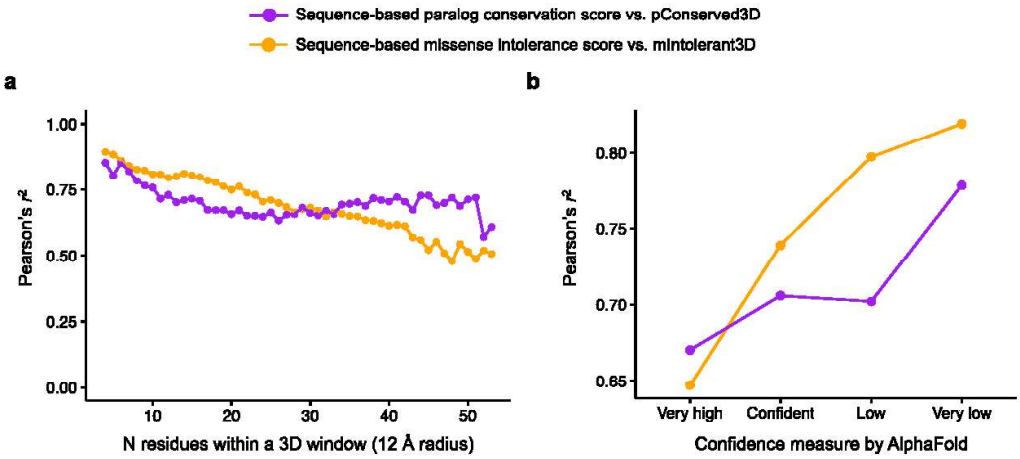
Supplementary Figure 6: **a**, Essential3D sites capture a higher fraction of *de novo* NDD variants in predicted monomeric structures (monomeric-AF set) compared to experimental monomeric structures (monomeric-PDB set) of 142 NDD proteins. Here, the NDD variants include NDD associated missense variants from denovo-db database, and those that are found associated with ASD, DD and DEE in independent cohort studies (see **Table 1**), amounting to 1,122 variants that were mappable onto 3D. **b**, Essential3D sites capture a higher proportion of *de novo* missense mutations in experimental protein complexes (multimeric-PDB) than in experimental monomeric structures (monomeric-PDB set) of 92 NDD proteins, with a total of 347 *de novo* NDD mutations that were mappable onto protein structures.



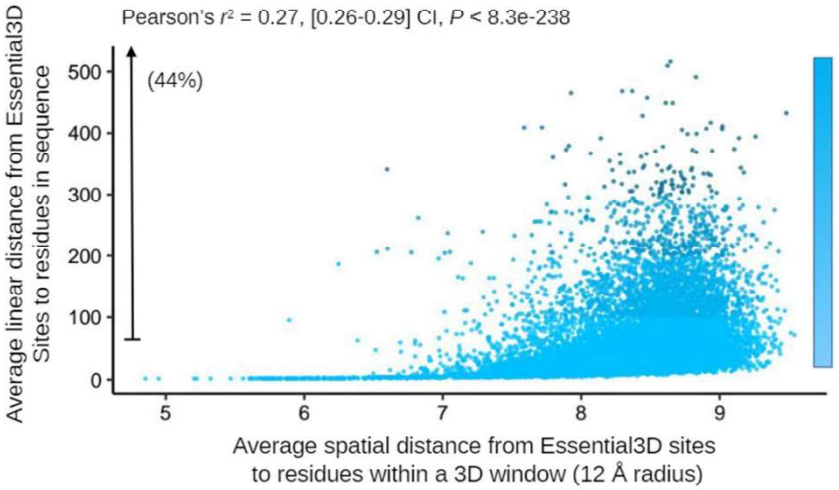
Supplementary Figure 7: De novo mutation associated with autism (ASD) and developmental epileptic encephalopathy (DEE) showed differential spatial location with respect to Essenetia3D sites. a, AlphaFold-predicted structure of Syntaxin-1 protein (*STXBP1*) with annotated Essnetial3D sites, and de novo mutations associated with ASD (left, cyan spheres) and DEE (right, pink spheres). 13 out of 26 rare, de novo ASD mutations are located in Essential3D sites, whereas only 2 out of 9 DEE rare, de novo mutations were captured by Essential3D sites. **b**, AlphaFold-predicted structure of Potassium channel protein, subfamily T member 1 (*KCNT1*) with annotated Essnetial3D sites and de novo mutations associated with DEE (pink spheres). In contrast with the DD mutations (Fig. 7, main file) which predominantly include ASD mutations, only 5 out of 40 DEE mutations as found by the Epi25 collaborative were located in Esseential3D sites. These case-studies suggest that Essential3D site annotation may preferably hint at the mechanisms associated with ASD and other DDs compared to DEE.



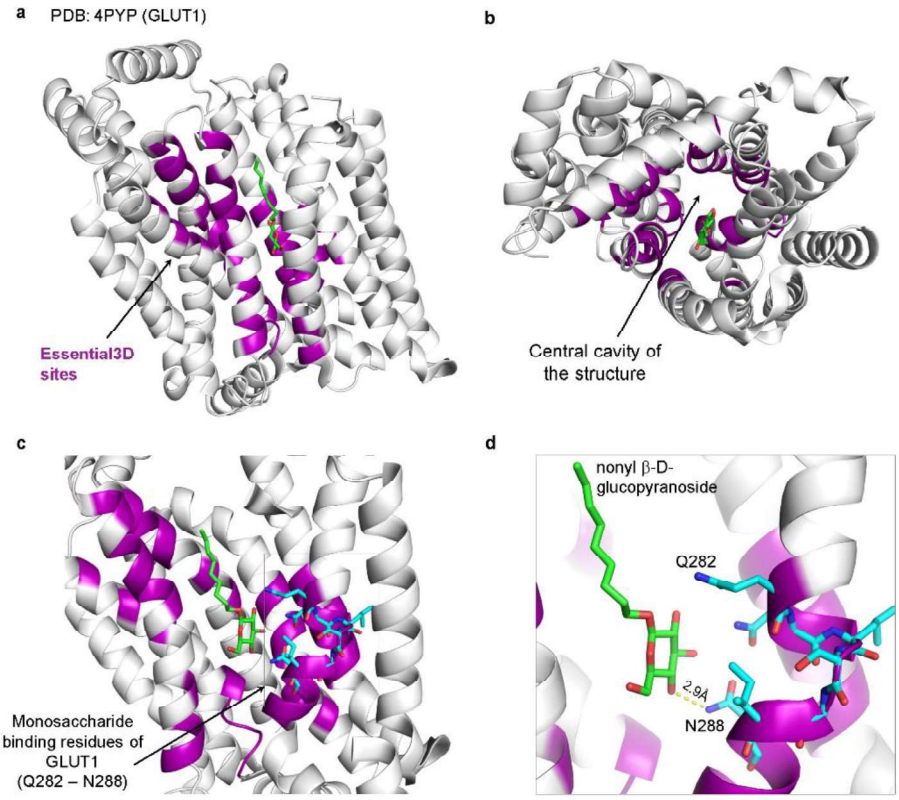
Supplementary Figure 8: The majority of Essential3D sites identified in AlphaFold-predicted structures (230 NDD genes) are of high prediction quality. Fraction of Essential3D, not-Essential3D, and not-Annotated (not an Essential3D or not-Essential3D site) sites of different prediction quality, as given by AlphaFold (very high: pLDDT > 90, confident: 70 < pLDDT < 90, low: 50 < pLDDT < 70, very low: pLDDT < 50). Over 90% of all Essential3D sites were located at residues with high predicted reliability compared to that of about 50% for not-Essential3D and not-Annotated sites.



Supplementary Figure 9: Structure-based conservation and constraint scores are complementary to sequence-based score for the residues located in compact, structured regions of proteins. a, Pearson's correlation between sequence-based conservation score and paralog conservation 3D score (pConserved3D), and sequence-based missense intolerance score (MTR) and missense intolerance 3D score (mIntolerant3D), as a function of number of residues (N) within the local 3D window (12 Å radius) around residues in protein structures. **b,** Pearson's correlation between sequence-based conservation score and pConserved3D, and sequence-based MTR and mIntolerant3D, as a function of prediction quality of residues generated by AlphaFold[ref] (very high: pLDDT > 90, confident: 70 < pLDDT < 90, low: 50 < pLDDT < 70, very low: pLDDT < 50).



Supplementary Figure 10: Linearly dispersed amino acid residues come close in 3D space to define a compact fold in the structure contributing to different and complementary outcome from sequence-based and structure-guided method. The plot shows the relationship between the average spatial distance (in Angstroms) and linear distance (positional distance between residues in the protein sequence) from Essential3D sites to all residues within the local 3D window (12 Å radius) around the sites. Linear distance and spatial distance of neighboring residues of 3D essential sites from the site itself were found weekly correlated (Pearson's $r^2 = 0.27$, [0.26-0.29] CI, P < 8.3e-238). Additionally, about 44% of all Essential3D sites have neighboring residues (within their local 3D window) that are on an average 50 residues apart from the site in the sequence, which can go up to maximum of 895 residues. For visualization purpose, y-axis was limited to 550. The darker the point color is, the greater the linear distance between the Essential3D sites and it's neighbor residues in the protein sequence.



Supplementary Figure 11: Essential3D sites of GLUT1 glucose transporter (PDB id: 4PYP) identifies residues essential for recognition and transportation of glucose. a, Transmembrane view of the structure (top: intracellular, bottom: extracellular) with Essential3D sites highlighted in *purple*. **b**, Essential3D sites are closely packed and forming the inner surface of the central cavity of the structure. **c**, A stretch of seven monosaccharide binding residues, Q282 – N288 (*cyan*), have been identified as Essential3D sites. **d**, Residue N288, part of the monosaccharide binding region (Q282 – N288, *cyan*) and an Essential3D site, forming a hydrogen bond with the nonyl β-D-glucopyranoside (*green*).

Supplementary Table legends

Supplementary Table 1: Overview of genes, encoded proteins, and variant level data for three different structure sets. a, monomeric-PDB set: experimentally solved structures from the Protein Data Bank (PDB) where an NDD protein's structure is present as a single molecule. For this set, we report the total number of available structures that fulfill the filtering criteria (see **Methods**) and the representative structure (PDB code and the relevant chain identifier) selected to compute structure-based scores. **b,** monomeric-AF: predicted monomeric structures from the AlphaFold (AF) protein structure database. Here, we report the per-protein fraction of residues with high prediction quality, pLDDT or predicted local-distance difference test > 70, and whether this protein has an analogous experimental structure (full or partial) in PDB. **c,** multimeric-PDB: experimentally solved multimeric structures (homomeric or heteromeric) from the Protein Data Bank (PDB) where an NDD protein is in complex with itself or other partner proteins. For this set, we report the selected structure (PDB code and whether it's a homomeric or heteromeric structure) and the partner proteins in the complex biological assemblies. For each gene in different structure sets, we further report the length of the encoded protein, coverage of protein residues on structure, the number of total and mappable variants (gnomAD, ClinVar, HGMD, DiscovEHR, and denovo-db databases; see **Methods** for details) onto protein structures.

Supplementary Table 2: List of 242 neurodevelopmental disorder (NDD) associated genes (NDD genes or proteins) and reference of their association with NDDs. The table includes per-gene missense Z score quantifying whether the gene is constrained for missense variations, HGNC gene family name, HGNC name of the protein encoded by the gene, Panther protein class name, and the reference identifying the association of the gene with NDDs (DOIs and/or PubMed Identifiers). A missense Z score > 1.96 indicates a missense constrained gene, as reported in Lek, M. et al, Nature, (2016).

Supplementary Table 3: Number of pathogenic variant enriched sites in 3D or pvEnriched3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no pvEnriched3D site.

Supplementary Table 4: Number of paralog conserved 3D sites or pConserved3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no pConserved3D site.

Supplementary Table 5: Number of missense intolerant 3D sites or mIntolerant3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally

solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no mIntolerant3D site.

Supplementary Table 6: Number of 3D essential sites or Essential3D sites as well as not-Essential3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no Essential3D or not-Esseential3D site.

Supplementary Table 7: Characteristic functional features associated with Essential3D sites. The table lists 12 functional features (position-specific feature annotations collected from the UniProt database) that are significantly enriched for Essentiial3D sites in 189 NDD proteins compared to the rest of the protein residues. The name and definition of the features are listed in the 1st column, the odds ratio (OR) enrichment and *P* values from two-tailed Fisher's Exact test are given in the 2nd column, and the 3rd column lists the genes that have an Esseetial3D site in the structure of the encoded proteins overlapped with the corresponding feature.

Supplementary Table 3: Number of pathogenic variant enriched sites in 3D or pvEnriched3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complxes from Protein Data Bank). *NA* indicates no pvEnriched3D site.

Genes	Monomeric-PDB	Monomeric-AF	Multimeric-PDB	Uniprot
ABL1	84	139	NA	P00519
ADCY5	NA	99	NA	O95622
ADNP	NA	38	NA	Q9H2P0
AGO1	NA	NA	NA	Q9UL18
AHDC1	NA	22	NA	Q5TGY3
AP2S1	23	23	NA	P53680
ARF1	1	5	3	P84077
ARHGAP35	NA	NA	NA	Q9NRY4
ARHGEF9	NA	60	NA	O43307
ARID1A	42	87	NA	O14497
ARID1B	NA	87	NA	Q8NFD5
ARID2	NA	NA	NA	Q68CP9
ASH1L	NA	NA	NA	Q9NR48
ATP1A3	NA	310	NA	P13637
ATP6V0A1	NA	NA	NA	Q93050
ATP6V1A	55	59	NA	P38606
ATRX	35	431	NA	P46100
AUTS2	NA	16	NA	Q8WXX7
BCL11A	NA	48	NA	Q9H165
BCL11B	NA	45	NA	Q9C0K0
BICD2	NA	54	NA	Q8TD16
BMPR2	72	172	72	Q13873
BRAF	127	151	NA	P15056
BRPF1	NA	42	NA	P55201
CACNA1A	NA	474	NA	O00555
CACNA1C	NA	133	NA	Q13936
CACNA1E	NA	123	NA	Q15878
CACNA1G	52	80	NA	O43497
CAMK2B	17	63	17	Q13554
CAMTA1	NA	44	NA	Q9Y6Y1
CASK	19	60	NA	O14936
CDK13	50	130	62	Q14004
CDK8	36	44	46	P49336
CDKL5	61	224	NA	O76039
CHD2	NA	200	NA	O14647
CHD3	NA	148	NA	Q12873
CHD4	56	155	56	Q14839
CHD8	NA	83	NA	Q9HCK8

CLCN4	NA	84	NA	P51793
CLTC	NA	26	NA	Q00610
CNKSR2	NA	NA	NA	Q8WXI2
CNOT3	NA	51	NA	O75175
COL2A1	NA	148	NA	P02458
COL4A1	1	33	1	P02462
COL4A3BP	NA	NA	NA	Q9Y5P4
CREBBP	21	410	NA	Q92793
CSNK2A1	40	65	55	P68400
CSNK2B	2	2	NA	P67870
CTBP1	NA	NA	NA	Q13363
CTCF	12	74	NA	P49711
CTNNB1	44	41	NA	P35222
CUL3	NA	25	NA	Q13618
CUL4B	15	49	84	Q13620
DDX23	NA	NA	NA	Q9BUQ8
DDX3X	56	145	NA	O00571
DLG4	NA	NA	NA	P78352
DNM1	94	97	100	Q05193
DNM1L	56	79	46	O00429
DNMT3A	26	101	77	Q9Y6K1
DPF2	4	53	NA	Q92785
DPYSL5	NA	NA	NA	Q9BPU6
DYNC1H1	171	NA	141	Q14204
DYRK1A	51	119	51	Q13627
EBF3	NA	85	NA	Q9H4W6
EEF1A2	NA	34	NA	Q05639
EEF2	NA	NA	NA	P13639
EFTUD2	56	53	115	Q15029
EIF5A	NA	NA	NA	P63241
EP300	47	217	NA	Q09472
ERF	NA	58	NA	P50548
FBN1	30	NA	33	P35555
FBXO11	NA	135	NA	Q86XK2
FBXW7	47	59	NA	Q969H0
FOXG1	NA	92	NA	P55316
FOXP1	NA	63	NA	Q9H334
GABRA1	81	90	101	P14867
GABRB2	89	92	107	P47870
GABRB3	107	110	112	P28472
GATAD2B	NA	14	NA	Q8WXI9
GNAI1	NA	NA	87	P63096
GNAO1	10	126	104	P09471
GNAS	NA	293	NA	Q5JWF2

GNB1	74	73	64	P62873
GNB2	NA	NA	NA	P62879
GRIA2	NA	39	1	P42262
GRIN1	97	108	93	Q05586
GRIN2A	157	226	172	Q12879
GRIN2B	1	281	1	Q13224
H3F3A	NA	NA	NA	P84243
HDAC2	NA	NA	NA	Q92769
HDAC4	NA	21	NA	P56524
HDAC8	44	72	44	Q9BY41
HECW2	NA	106	NA	Q9P2P5
HK1	31	34	31	P19367
HNRNPD	NA	NA	NA	Q14103
HNRNPK	NA	33	NA	P61978
HNRNPU	NA	28	NA	Q00839
HUWE1	24	NA	NA	Q7Z6Z7
IQSEC2	14	113	14	Q5JU85
KAT6A	NA	65	NA	Q92794
KAT6B	NA	50	NA	Q8WYB5
KCNA2	NA	76	NA	P16389
KCNB1	NA	170	NA	Q14721
KCND3	NA	91	NA	Q9UK17
KCNH1	NA	85	NA	O95259
KCNK3	18	75	25	O14649
KCNQ2	86	349	NA	O43526
KCNQ3	NA	90	NA	O43525
KCNT1	NA	153	NA	Q5JUK3
KDM5C	33	230	33	P41229
KDM6A	51	92	49	O15550
KIDINS220	NA	NA	NA	Q9ULH0
KIF11	29	66	29	P52732
KIF1A	111	254	2	Q12756
KIF5C	NA	46	NA	O60282
KMT2A	8	NA	NA	Q03164
KMT2B	NA	NA	NA	Q9UMN6
KMT2C	NA	NA	NA	Q8NEZ4
KMT2D	4	NA	NA	O14686
KMT5B	3	62	3	Q4FZB7
KRAS	76	91	NA	P01116
MAP2K1	91	80	124	Q02750
MAP3K7	36	93	46	O43318
MED12	NA	144	NA	Q93074
MED13L	NA	105	NA	Q71F56
MEF2C	NA	52	NA	Q06413

MEIS2	NA	34	NA	O14770
MIB1	7	5	4	Q86YT6
MN1	NA	NA	NA	Q10571
MSL2	NA	NA	NA	Q9HCI7
MTOR	181	204	NA	P42345
MYT1L	NA	65	NA	Q9UL68
NAA10	19	50	123	P41227
NAA15	29	27	16	Q9BXJ9
NACC1	NA	7	5	Q96RE7
NF1	69	NA	26	P21359
NFIA	NA	23	NA	Q12857
NFIX	NA	61	NA	Q14938
NONO	NA	NA	NA	Q15233
NOTCH1	NA	111	NA	P46531
NR2F1	NA	76	NA	P10589
NR4A2	NA	25	NA	P43354
NSD1	40	494	49	Q96L73
NSD2	NA	NA	NA	O96028
OGT	14	14	9	O15294
PACS1	NA	NA	NA	Q6VY07
PACS2	NA	9	NA	Q86VP3
PBX1	NA	24	NA	P40424
PCBP2	NA	NA	NA	Q15366
PCDH19	90	225	94	Q8TAB3
PDHA1	131	147	186	P08559
PHF12	NA	NA	NA	Q96QT6
PHF21A	NA	NA	NA	Q96BD5
PHF6	3	54	NA	Q8IWS0
PHIP	NA	107	NA	Q8WWQ0
PIK3CA	207	217	193	P42336
PIK3R1	19	56	NA	P27986
PIK3R2	NA	47	NA	O00459
POGZ	NA	21	NA	Q7Z3K3
POU3F3	NA	9	NA	P20264
PPM1D	NA	NA	NA	O15297
PPP1CB	NA	NA	NA	P62140
PPP2CA	56	57	109	P67775
PPP2R1A	41	44	41	P30153
PPP2R5D	NA	17	NA	Q14738
PPP3CA	62	70	67	Q08209
PRPF8	100	104	74	Q6P2Q9
PRR12	NA	NA	NA	Q9ULL5
PSMC5	NA	NA	NA	P62195
PTEN	77	144	NA	P60484

PTPN11	170	165	NA	Q06124
PUF60	33	55	31	Q9UHX1
PURA	NA	74	NA	Q00577
QRICH1	NA	NA	NA	Q2TAL8
RAB14	NA	NA	NA	P61106
RAC1	26	22	54	P63000
RERE	NA	23	NA	Q9P2R6
RHOBTB2	NA	54	NA	Q9BYZ6
RIT1	32	40	NA	Q92963
RPS6KA3	25	90	26	P51812
SATB1	NA	NA	NA	Q01826
SATB2	NA	69	NA	Q9UPW6
SCN1A	243	957	NA	P35498
SCN2A	346	543	NA	Q99250
SCN8A	NA	455	NA	Q9UQD0
SET	2	20	NA	Q01105
SETD1B	NA	87	NA	Q9UPS6
SETD2	5	132	5	Q9BYW2
SHOC2	NA	36	NA	Q9UQ13
SIN3A	NA	NA	NA	Q96ST3
SLC2A1	127	150	NA	P11166
SLC35A2	NA	42	NA	P78381
SLC6A1	NA	103	NA	P30531
SMAD4	44	101	NA	Q13485
SMARCA2	87	264	87	P51531
SMARCA4	69	252	NA	P51532
SMARCB1	60	54	NA	Q12824
SMC1A	16	72	31	Q14683
SMC3	28	62	6	Q9UQE7
SNAP25	NA	11	NA	P60880
SOX10	NA	85	NA	P56693
SOX11	NA	69	NA	P35716
SOX5	NA	69	NA	P35711
SPEN	NA	NA	NA	Q96T58
SPOP	51	52	51	O43791
SPTAN1	NA	71	NA	Q13813
SRCAP	NA	NA	NA	Q6ZRS2
SRSF1	NA	NA	NA	Q07955
STAG1	60	79	NA	Q8WVM7
STAG2	20	53	18	Q8N3U4
STXBP1	NA	128	NA	P61764
SYNCRIP	NA	NA	NA	O60506
SYNGAP1	NA	134	NA	Q96PV0
SYT1	30	33	27	P21579

TAF1	17	102	NA	P21675
TAOK1	NA	40	NA	Q7L7X3
TBL1XR1	121	133	NA	Q9BZK7
TCF4	NA	57	NA	P15884
TCF7L2	NA	NA	NA	Q9NQB0
TFE3	NA	NA	NA	P19532
TLK2	4	46	NA	Q86UE8
TRAF7	NA	17	NA	Q6Q0C0
TRIO	34	NA	NA	O75962
TRIP12	NA	39	NA	Q14669
U2AF2	NA	NA	NA	P26368
UBE3A	42	82	NA	Q05086
UBTF	NA	NA	NA	P17480
UPF1	NA	NA	NA	Q92900
USP7	46	49	30	Q93009
USP9X	13	101	14	Q93008
VCP	68	73	71	P55072
WDR26	NA	56	NA	Q9H7D7
WDR37	NA	17	NA	Q9Y2I8
WDR45	NA	NA	NA	Q5MNZ6
YY1	NA	48	NA	P25490
ZBTB18	NA	72	NA	Q99592
ZBTB20	NA	84	NA	Q9HC78
ZEB2	NA	70	NA	O60315
ZMIZ1	NA	21	NA	Q9ULJ6
ZMYM2	NA	NA	NA	Q9UBW7
ZMYND11	NA	17	4	Q15326
ZNF148	NA	NA	NA	Q9UQR1

Supplementary Table 4: Number of paralog conserved 3D sites or pConserved3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no pConserved3D site.

Genes	Monomeric-PDB	Monomeric-AF	Multimeric-PDB	Uniprot
ABL1	NA	NA	NA	P00519
ADCY5	NA	680	NA	O95622
ADNP	NA	790	NA	Q9H2P0
AGO1	405	436	398	Q9UL18
AHDC1	NA	NA	NA	Q5TGY3
AP2S1	NA	NA	NA	P53680
ARF1	100	104	100	P84077
ARHGAP35	51	382	46	Q9NRY4
ARHGEF9	NA	283	NA	O43307
ARID1A	167	620	NA	O14497
ARID1B	NA	NA	NA	Q8NFD5
ARID2	NA	NA	NA	Q68CP9
ASH1L	NA	NA	45	Q9NR48
ATP1A3	NA	NA	NA	P13637
ATP6V0A1	NA	NA	NA	Q93050
ATP6V1A	NA	NA	NA	P38606
ATRX	NA	900	NA	P46100
AUTS2	NA	505	NA	Q8WXX7
BCL11A	NA	435	NA	Q9H165
BCL11B	NA	443	NA	Q9C0K0
BICD2	NA	NA	NA	Q8TD16
BMPR2	180	424	180	Q13873
BRAF	208	268	NA	P15056
BRPF1	NA	NA	NA	P55201
CACNA1A	NA	1108	NA	O00555
CACNA1C	NA	NA	NA	Q13936
CACNA1E	NA	1062	NA	Q15878
CACNA1G	663	1092	NA	O43497
CAMK2B	155	293	155	Q13554
CAMTA1	NA	1022	NA	Q9Y6Y1
CASK	NA	NA	NA	O14936
CDK13	211	437	257	Q14004
CDK8	215	317	234	P49336
CDKL5	NA	NA	NA	O76039
CHD2	NA	771	NA	O14647
CHD3	NA	NA	NA	Q12873
CHD4	394	791	394	Q14839
CHD8	NA	1024	NA	Q9HCK8
CLCN4	NA	418	NA	P51793
CLTC	NA	NA	NA	Q00610
CNKSR2	NA	314	NA	Q8WXI2
CNOT3	NA	NA	NA	O75175
COL2A1	NA	886	NA	P02458

COL4A1	112	876	120	P02462
COL4A3BP	NA	NA	NA	Q9Y5P4
CREBBP	99	957	NA	Q92793
CSNK2A1	234	327	216	P68400
CSNK2B	NA	NA	NA	P67870
CTBP1	200	263	NA	Q13363
CTCF	90	457	NA	P49711
CTNNB1	304	553	NA	P35222
CUL3	218	436	106	Q13618
CUL4B	356	652	580	Q13620
DDX23	266	510	363	Q9BUQ8
DDX3X	255	390	NA	O00571
DLG4	NA	NA	NA	P78352
DNM1	443	502	440	Q05193
DNM1L	NA	NA	NA	O00429
DNMT3A	257	408	257	Q9Y6K1
DPF2	61	230	NA	Q92785
DPYSL5	259	423	258	Q9BPU6
DYNC1H1	1452	NA	33	Q14204
DYRK1A	192	341	192	Q13627
EBF3	NA	NA	NA	Q9H4W6
EEF1A2	NA	275	NA	Q05639
EEF2	NA	540	NA	P13639
EFTUD2	655	742	656	Q15029
EIF5A	NA	NA	NA	P63241
EP300	268	946	NA	Q09472
ERF	NA	159	NA	P50548
FBN1	105	NA	69	P35555
FBXO11	NA	615	NA	Q86XK2
FBXW7	231	350	NA	Q969H0
FOXG1	NA	120	NA	P55316
FOXP1	NA	NA	NA	Q9H334
GABRA1	231	295	212	P14867
GABRB2	203	293	213	P47870
GABRB3	213	281	206	P28472
GATAD2B	NA	325	NA	Q8WXI9
GNAI1	182	179	174	P63096
GNAO1	113	179	100	P09471
GNAS	NA	381	NA	Q5JWF2
GNB1	215	212	207	P62873
GNB2	NA	213	NA	P62879
GRIA2	203	425	212	P42262
GRIN1	NA	NA	124	Q05586
GRIN2A	350	801	348	Q12879
GRIN2B	194	795	194	Q13224
H3F3A	61	81	NA	P84243
HDAC2	201	349	201	Q92769
HDAC4	217	611	NA	P56524
HDAC8	164	229	164	Q9BY41
HECW2	53	593	NA	Q9P2P5
HK1	NA	NA	NA	P19367
HNRNPD	NA	226	NA	Q14103
HNRNPK	NA	NA	NA	P61978

HNRNPU	NA	491	NA	Q00839
HUWE1	296	NA	NA	Q7Z6Z7
IQSEC2	190	307	193	Q5JU85
KAT6A	134	552	49	Q92794
KAT6B	50	550	50	Q8WYB5
KCNA2	NA	291	NA	P16389
KCNB1	NA	366	NA	Q14721
KCND3	81	324	94	Q9UK17
KCNH1	54	431	54	O95259
KCNK3	122	211	120	O14649
KCNQ2	213	450	NA	O43526
KCNQ3	NA	457	NA	O43525
KCNT1	NA	NA	NA	Q5JUK3
KDM5C	189	964	189	P41229
KDM6A	266	524	261	O15550
KIDINS220	NA	661	NA	Q9ULH0
KIF11	179	370	180	P52732
KIF1A	NA	NA	NA	Q12756
KIF5C	NA	358	NA	O60282
KMT2A	NA	NA	NA	Q03164
KMT2B	101	NA	99	Q9UMN6
KMT2C	98	NA	NA	Q8NEZ4
KMT2D	84	NA	NA	O14686
KMT5B	NA	NA	NA	Q4FZB7
KRAS	93	105	NA	P01116
MAP2K1	189	241	53	Q02750
MAP3K7	165	266	170	O43318
MED12	NA	1541	NA	Q93074
MED13L	NA	1353	NA	Q71F56
MEF2C	NA	NA	NA	Q06413
MEIS2	NA	132	NA	O14770
MIB1	159	432	156	Q86YT6
MN1	NA	NA	NA	Q10571
MSL2	NA	NA	NA	Q9HCI7
MTOR	1261	1518	NA	P42345
MYT1L	NA	653	NA	Q9UL68
NAA10	90	166	160	P41227
NAA15	430	495	389	Q9BXJ9
NACC1	81	223	61	Q96RE7
NF1	NA	NA	NA	P21359
NFIA	NA	NA	NA	Q12857
NFIX	NA	NA	NA	Q14938
NONO	143	315	154	Q15233
NOTCH1	43	705	NA	P46531
NR2F1	NA	251	NA	P10589
NR4A2	149	289	NA	P43354
NSD1	107	536	106	Q96L73
NSD2	NA	NA	NA	O96028
OGT	261	492	262	O15294
PACS1	NA	704	NA	Q6VY07
PACS2	NA	NA	NA	Q86VP3
PBX1	NA	292	NA	P40424
PCBP2	NA	NA	NA	Q15366

PCDH19	229	628	223	Q8TAB3
PDHA1	NA	NA	78	P08559
PHF12	NA	NA	NA	Q96QT6
PHF21A	NA	460	NA	Q96BD5
PHF6	68	209	NA	Q8IWS0
PHIP	88	1135	86	Q8WWQ0
PIK3CA	481	495	421	P42336
PIK3R1	154	423	NA	P27986
PIK3R2	65	420	NA	O00459
POGZ	NA	541	NA	Q7Z3K3
POU3F3	NA	172	NA	P20264
PPM1D	NA	342	NA	O15297
PPP1CB	NA	248	NA	P62140
PPP2CA	195	194	193	P67775
PPP2R1A	355	353	355	P30153
PPP2R5D	NA	412	NA	Q14738
PPP3CA	221	402	258	Q08209
PRPF8	NA	NA	23	Q6P2Q9
PRR12	NA	1287	NA	Q9ULL5
PSMC5	NA	NA	1	P62195
PTEN	172	217	NA	P60484
PTPN11	214	230	NA	Q06124
PUF60	113	293	121	Q9UHX1
PURA	NA	220	NA	Q00577
QRICH1	NA	247	NA	Q2TAL8
RAB14	73	99	57	P61106
RAC1	NA	NA	NA	P63000
RERE	NA	606	NA	Q9P2R6
RHOBTB2	NA	NA	NA	Q9BYZ6
RIT1	NA	NA	NA	Q92963
RPS6KA3	85	317	142	P51812
SATB1	NA	NA	NA	Q01826
SATB2	NA	457	NA	Q9UPW6
SCN1A	696	985	NA	P35498
SCN2A	692	989	NA	Q99250
SCN8A	NA	982	NA	Q9UQD0
SET	100	165	98	Q01105
SETD1B	NA	NA	NA	Q9UPS6
SETD2	116	514	117	Q9BYW2
SHOC2	NA	279	NA	Q9UQ13
SIN3A	NA	871	NA	Q96ST3
SLC2A1	246	301	NA	P11166
SLC35A2	NA	NA	NA	P78381
SLC6A1	NA	396	NA	P30531
SMAD4	146	294	NA	Q13485
SMARCA2	156	1117	74	P51531
SMARCA4	321	1229	NA	P51532
SMARCB1	NA	NA	NA	Q12824
SMC1A	297	680	76	Q14683
SMC3	NA	NA	33	Q9UQE7
SNAP25	36	142	NA	P60880
SOX10	NA	119	NA	P56693
SOX11	NA	104	NA	P35716

SOX5	NA	471	NA	P35711
SPEN	160	NA	105	Q96T58
SPOP	174	215	183	O43791
SPTAN1	NA	NA	NA	Q13813
SRCAP	179	NA	232	Q6ZRS2
SRSF1	20	112	NA	Q07955
STAG1	529	825	221	Q8WVM7
STAG2	NA	NA	127	Q8N3U4
STXBP1	NA	NA	NA	P61764
SYNCRIP	84	330	NA	O60506
SYNGAP1	NA	562	NA	Q96PV0
SYT1	141	261	145	P21579
TAF1	NA	NA	NA	P21675
TAOK1	NA	337	NA	Q7L7X3
TBL1XR1	194	304	NA	Q9BZK7
TCF4	NA	NA	NA	P15884
TCF7L2	NA	NA	NA	Q9NQB0
TFE3	NA	295	NA	P19532
TLK2	196	580	NA	Q86UE8
TRAF7	NA	338	NA	Q6Q0C0
TRIO	169	NA	NA	O75962
TRIP12	NA	NA	NA	Q14669
U2AF2	104	249	NA	P26368
UBE3A	139	381	NA	Q05086
UBTF	NA	NA	NA	P17480
UPF1	319	732	327	Q92900
USP7	297	352	302	Q93009
USP9X	NA	NA	NA	Q93008
VCP	NA	NA	NA	P55072
WDR26	NA	NA	NA	Q9H7D7
WDR37	NA	314	NA	Q9Y2I8
WDR45	NA	NA	NA	Q5MNZ6
YY1	72	232	NA	P25490
ZBTB18	NA	NA	NA	Q99592
ZBTB20	NA	286	NA	Q9HC78
ZEB2	NA	534	NA	O60315
ZMIZ1	NA	506	NA	Q9ULJ6
ZMYM2	NA	364	NA	Q9UBW7
ZMYND11	81	329	60	Q15326
ZNF148	NA	259	NA	Q9UQR1

Supplementary Table 5: Number of missense intolerant 3D sites or mIntolerant3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). *NA* indicates no mIntolerant3D site.

Genes	Monomeric-PDB	Monomeric-AF	Multimeric-PDB	Uniprot
ABL1	213	509	NA	P00519
ADCY5	NA	622	NA	O95622
ADNP	NA	512	NA	Q9H2P0
AGO1	338	373	353	Q9UL18
AHDC1	NA	851	NA	Q5TGY3
AP2S1	62	66	NA	P53680
ARF1	98	95	97	P84077
ARHGAP35	103	676	105	Q9NRY4
ARHGEF9	NA	288	NA	O43307
ARID1A	236	1100	NA	O14497
ARID1B	60	1140	NA	Q8NFD5
ARID2	NA	744	NA	Q68CP9
ASH1L	105	NA	137	Q9NR48
ATP1A3	NA	635	NA	P13637
ATP6V0A1	408	443	NA	Q93050
ATP6V1A	276	290	NA	P38606
ATRX	76	1064	NA	P46100
AUTS2	NA	621	NA	Q8WXX7
BCL11A	NA	398	NA	Q9H165
BCL11B	NA	471	NA	Q9C0K0
BICD2	NA	407	NA	Q8TD16
BMPR2	124	507	124	Q13873
BRAF	150	356	NA	P15056
BRPF1	78	596	71	P55201
CACNA1A	NA	1245	NA	O00555
CACNA1C	NA	NA	NA	Q13936

CACNA1E	NA	1163	NA	Q15878
CACNA1G	499	1108	NA	O43497
CAMK2B	151	324	151	Q13554
CAMTA1	NA	862	NA	Q9Y6Y1
CASK	187	419	NA	O14936
CDK13	183	661	97	Q14004
CDK8	153	199	237	P49336
CDKL5	126	486	NA	O76039
CHD2	NA	906	NA	O14647
CHD3	NA	1028	NA	Q12873
CHD4	354	888	354	Q14839
CHD8	NA	1240	NA	Q9HCK8
CLCN4	NA	358	NA	P51793
CLTC	187	818	NA	Q00610
CNKSR2	NA	588	NA	Q8WXI2
CNOT3	59	416	NA	O75175
COL2A1	NA	772	NA	P02458
COL4A1	129	837	122	P02462
COL4A3BP	131	309	NA	Q9Y5P4
CREBBP	106	1148	NA	Q92793
CSNK2A1	152	200	165	P68400
CSNK2B	79	88	NA	P67870
CTBP1	159	200	NA	Q13363
CTCF	108	361	NA	P49711
CTNNB1	242	381	NA	P35222
CUL3	174	367	173	Q13618
CUL4B	354	465	424	Q13620
DDX23	235	448	327	Q9BUQ8
DDX3X	220	393	NA	O00571
DLG4	89	368	NA	P78352
DNM1	401	441	394	Q05193
DNM1L	250	346	250	O00429
DNMT3A	192	417	227	Q9Y6K1

DPF2	73	166	NA	Q92785
DPYSL5	248	285	235	Q9BPU6
DYNC1H1	1519	NA	761	Q14204
DYRK1A	188	428	188	Q13627
EBF3	86	344	77	Q9H4W6
EEF1A2	NA	262	NA	Q05639
EEF2	NA	492	NA	P13639
EFTUD2	517	559	240	Q15029
EIF5A	64	69	38	P63241
EP300	342	1095	NA	Q09472
ERF	NA	324	NA	P50548
FBN1	71	NA	92	P35555
FBXO11	NA	471	NA	Q86XK2
FBXW7	209	384	NA	Q969H0
FOXG1	NA	203	NA	P55316
FOXP1	NA	320	NA	Q9H334
GABRA1	186	239	189	P14867
GABRB2	174	281	181	P47870
GABRB3	188	306	229	P28472
GATAD2B	NA	304	NA	Q8WXI9
GNAI1	188	183	216	P63096
GNAO1	124	194	202	P09471
GNAS	NA	397	NA	Q5JWF2
GNB1	193	200	198	P62873
GNB2	NA	171	NA	P62879
GRIA2	172	395	174	P42262
GRIN1	353	438	466	Q05586
GRIN2A	390	705	211	Q12879
GRIN2B	204	718	204	Q13224
H3F3A	54	76	NA	P84243
HDAC2	207	331	207	Q92769
HDAC4	234	570	NA	P56524
HDAC8	206	222	206	Q9BY41

HECW2	52	840	NA	Q9P2P5
HK1	430	448	432	P19367
HNRNPD	NA	143	NA	Q14103
HNRNPK	NA	210	NA	P61978
HNRNPU	NA	413	NA	Q00839
HUWE1	221	NA	NA	Q7Z6Z7
IQSEC2	169	790	129	Q5JU85
KAT6A	157	855	62	Q92794
KAT6B	55	980	53	Q8WYB5
KCNA2	NA	274	NA	P16389
KCNB1	NA	441	NA	Q14721
KCND3	72	258	96	Q9UK17
KCNH1	57	498	57	O95259
KCNK3	143	236	148	O14649
KCNQ2	180	370	NA	O43526
KCNQ3	NA	380	NA	O43525
KCNT1	NA	652	NA	Q5JUK3
KDM5C	226	738	226	P41229
KDM6A	213	711	214	O15550
KIDINS220	NA	867	NA	Q9ULH0
KIF11	139	445	125	P52732
KIF1A	203	753	89	Q12756
KIF5C	NA	483	NA	O60282
KMT2A	100	NA	NA	Q03164
KMT2B	105	NA	159	Q9UMN6
KMT2C	75	NA	NA	Q8NEZ4
KMT2D	79	NA	NA	O14686
KMT5B	114	398	114	Q4FZB7
KRAS	90	102	NA	P01116
MAP2K1	173	183	180	Q02750
MAP3K7	136	259	135	O43318
MED12	NA	1115	NA	Q93074
MED13L	NA	1046	NA	Q71F56

MEF2C	NA	229	NA	Q06413
MEIS2	NA	211	NA	O14770
MIB1	144	445	146	Q86YT6
MN1	NA	652	NA	Q10571
MSL2	73	306	NA	Q9HCI7
MTOR	1011	1176	NA	P42345
MYT1L	NA	533	NA	Q9UL68
NAA10	88	133	160	P41227
NAA15	303	338	239	Q9BXJ9
NACC1	78	242	59	Q96RE7
NF1	172	NA	129	P21359
NFIA	NA	263	NA	Q12857
NFIX	NA	272	NA	Q14938
NONO	124	265	114	Q15233
NOTCH1	145	1309	NA	P46531
NR2F1	NA	244	NA	P10589
NR4A2	93	191	NA	P43354
NSD1	121	1086	116	Q96L73
NSD2	141	652	141	O96028
OGT	318	462	321	O15294
PACS1	NA	522	NA	Q6VY07
PACS2	NA	477	NA	Q86VP3
PBX1	NA	193	NA	P40424
PCBP2	71	241	NA	Q15366
PCDH19	230	641	230	Q8TAB3
PDHA1	216	240	262	P08559
PHF12	NA	474	NA	Q96QT6
PHF21A	NA	349	NA	Q96BD5
PHF6	42	171	NA	Q8IWS0
PHIP	61	880	61	Q8WWQ0
PIK3CA	533	552	501	P42336
PIK3R1	138	379	NA	P27986
PIK3R2	77	407	NA	O00459

POGZ	NA	665	NA	Q7Z3K3
POU3F3	NA	200	NA	P20264
PPM1D	NA	254	NA	O15297
PPP1CB	NA	196	NA	P62140
PPP2CA	200	206	288	P67775
PPP2R1A	341	349	341	P30153
PPP2R5D	NA	275	NA	Q14738
PPP3CA	215	311	132	Q08209
PRPF8	1186	1239	241	Q6P2Q9
PRR12	NA	1023	NA	Q9ULL5
PSMC5	226	233	363	P62195
PTEN	146	211	NA	P60484
PTPN11	247	283	NA	Q06124
PUF60	85	250	91	Q9UHX1
PURA	NA	202	NA	Q00577
QRICH1	NA	319	NA	Q2TAL8
RAB14	89	117	116	P61106
RAC1	102	99	136	P63000
RERE	NA	819	NA	Q9P2R6
RHOBTB2	NA	308	NA	Q9BYZ6
RIT1	64	120	NA	Q92963
RPS6KA3	140	420	152	P51812
SATB1	51	350	NA	Q01826
SATB2	NA	397	NA	Q9UPW6
SCN1A	596	1082	NA	P35498
SCN2A	512	931	NA	Q99250
SCN8A	NA	896	NA	Q9UQD0
SET	77	165	76	Q01105
SETD1B	NA	900	NA	Q9UPS6
SETD2	117	1122	124	Q9BYW2
SHOC2	NA	296	NA	Q9UQ13
SIN3A	NA	552	NA	Q96ST3
SLC2A1	268	283	NA	P11166

SLC35A2	NA	176	NA	P78381
SLC6A1	NA	363	NA	P30531
SMAD4	99	381	NA	Q13485
SMARCA2	120	875	40	P51531
SMARCA4	325	827	NA	P51532
SMARCB1	148	199	NA	Q12824
SMC1A	316	687	180	Q14683
SMC3	317	597	13	Q9UQE7
SNAP25	36	103	NA	P60880
SOX10	NA	191	NA	P56693
SOX11	NA	215	NA	P35716
SOX5	NA	351	NA	P35711
SPEN	139	NA	74	Q96T58
SPOP	143	170	154	O43791
SPTAN1	76	1141	76	Q13813
SRCAP	174	NA	31	Q6ZRS2
SRSF1	70	170	NA	Q07955
STAG1	458	601	173	Q8WVM7
STAG2	450	620	441	Q8N3U4
STXBP1	NA	345	NA	P61764
SYNCRIP	82	344	NA	O60506
SYNGAP1	NA	694	NA	Q96PV0
SYT1	130	245	128	P21579
TAF1	330	887	NA	P21675
TAOK1	NA	549	NA	Q7L7X3
TBL1XR1	182	293	NA	Q9BZK7
TCF4	NA	308	NA	P15884
TCF7L2	NA	288	NA	Q9NQB0
TFE3	NA	267	NA	P19532
TLK2	134	389	NA	Q86UE8
TRAF7	NA	369	NA	Q6Q0C0
TRIO	153	NA	160	O75962
TRIP12	NA	1003	NA	Q14669

U2AF2	124	266	NA	P26368
UBE3A	195	377	NA	Q05086
UBTF	NA	462	NA	P17480
UPF1	356	704	405	Q92900
USP7	375	584	257	Q93009
USP9X	167	1087	166	Q93008
VCP	367	392	361	P55072
WDR26	NA	352	NA	Q9H7D7
WDR37	NA	302	NA	Q9Y2I8
WDR45	136	156	NA	Q5MNZ6
YY1	62	186	NA	P25490
ZBTB18	NA	NA	NA	Q99592
ZBTB20	NA	372	NA	Q9HC78
ZEB2	NA	459	NA	O60315
ZMIZ1	61	558	NA	Q9ULJ6
ZMYM2	NA	605	NA	Q9UBW7
ZMYND11	101	333	65	Q15326
ZNF148	NA	379	NA	Q9UQR1

Supplementary Table 6: Number of 3D essential sites or Essential3D sites as well as not-Essential3D sites identified for each gene in our cohort (N = 242 genes). Counts are separately reported for structures in three structure sets: monomeric-PDB (experimentally solved monomeric structures from Protein Data Bank), monomeric-AF (predicted monomeric structures from AlphaFold database), and multimeric-PDB (experimentally solved homomeric or heteromeric protein complexes from Protein Data Bank). NA indicates no Essential3D or not-Esseential3D site.

	monomeric-PDB		monomeric-AF		multimeric-PDB		Uniprot
Genes	Essential3D	not-Essential3D	Essential3D	not-Essential3D	Essential3D	not-Essential3D	
ABL1	71	2	138	NA	NA	NA	P00519
ADCY5	NA	NA	54	NA	NA	NA	O95622
ADNP	NA	NA	26	NA	NA	NA	Q9H2P0
AGO1	NA	NA	NA	NA	NA	NA	Q9UL18
AHDC1	NA	NA	13	NA	NA	NA	Q5TGY3
AP2S1	11	NA	11	NA	NA	NA	P53680
ARF1	1	NA	5	NA	3	NA	P84077
ARHGAP35	NA	NA	NA	NA	NA	NA	Q9NRY4
ARHGEF9	NA	NA	28	NA	NA	NA	O43307
ARID1A	2	NA	23	NA	NA	NA	O14497
ARID1B	NA	NA	84	NA	NA	NA	Q8NFD5
ARID2	NA	NA	NA	NA	NA	NA	Q68CP9
ASH1L	NA	NA	NA	NA	NA	NA	Q9NR48
ATP1A3	NA	NA	205	86	NA	NA	P13637
ATP6V0A1	NA	NA	NA	NA	NA	NA	Q93050
ATP6V1A	40	NA	46	NA	NA	NA	P38606
ATRX	18	7	306	NA	NA	NA	P46100
AUTS2	NA	NA	NA	NA	NA	NA	Q8WXX7
BCL11A	NA	NA	35	NA	NA	NA	Q9H165
BCL11B	NA	NA	40	NA	NA	NA	Q9C0K0
BICD2	NA	NA	29	NA	NA	NA	Q8TD16
BMPR2	45	6	137	NA	45	6	Q13873
BRAF	14	16	76	NA	NA	NA	P15056
BRPF1	NA	NA	42	NA	NA	NA	P55201

CACNA1A	NA	NA	415	NA	NA	NA	O00555
CACNA1C	NA	NA	133	NA	NA	NA	Q13936
CACNA1E	NA	NA	119	NA	NA	NA	Q15878
CACNA1G	48	NA	79	NA	NA	NA	O43497
CAMK2B	12	NA	62	NA	12	NA	Q13554
CAMTA1	NA	NA	35	NA	NA	NA	Q9Y6Y1
CASK	14	NA	46	NA	NA	NA	O14936
CDK13	30	3	130	NA	33	NA	Q14004
CDK8	28	NA	37	NA	39	NA	P49336
CDKL5	34	20	211	13	NA	NA	O76039
CHD2	NA	NA	134	NA	NA	NA	O14647
CHD3	NA	NA	140	NA	NA	NA	Q12873
CHD4	56	NA	153	NA	56	NA	Q14839
CHD8	NA	NA	63	NA	NA	NA	Q9HCK8
CLCN4	NA	NA	82	NA	NA	NA	P51793
CLTC	NA	NA	25	NA	NA	NA	Q00610
CNKSR2	NA	NA	NA	NA	NA	NA	Q8WXI2
CNOT3	NA	NA	51	NA	NA	NA	O75175
COL2A1	NA	NA	28	56	NA	NA	P02458
COL4A1	NA	NA	6	NA	1	NA	P02462
COL4A3BP	NA	NA	NA	NA	NA	NA	Q9Y5P4
CREBBP	21	NA	193	NA	NA	NA	Q92793
CSNK2A1	22	NA	36	NA	26	NA	P68400
CSNK2B	NA	NA	NA	NA	NA	NA	P67870
CTBP1	NA	NA	NA	NA	NA	NA	Q13363
CTCF	3	NA	70	NA	NA	NA	P49711
CTNNB1	26	NA	29	NA	NA	NA	P35222
CUL3	NA	NA	19	NA	NA	NA	Q13618
CUL4B	11	NA	40	NA	70	NA	Q13620
DDX23	NA	NA	NA	NA	NA	NA	Q9BUQ8
DDX3X	35	18	136	NA	NA	NA	O00571
DLG4	NA	NA	NA	NA	NA	NA	P78352
DNM1	62	NA	75	NA	67	NA	Q05193

DNM1L	43	1	70	NA	34	NA	O00429
DNMT3A	20	NA	58	NA	45	NA	Q9Y6K1
DPF2	NA	NA	11	NA	NA	NA	Q92785
DPYSL5	NA	NA	NA	NA	NA	NA	Q9BPU6
DYNC1H1	149	NA	NA	NA	3	NA	Q14204
DYRK1A	40	NA	117	NA	40	NA	Q13627
EBF3	NA	NA	85	NA	NA	NA	Q9H4W6
EEF1A2	NA	NA	31	NA	NA	NA	Q05639
EEF2	NA	NA	NA	NA	NA	NA	P13639
EFTUD2	47	NA	47	NA	37	NA	Q15029
EIF5A	NA	NA	NA	NA	NA	NA	P63241
EP300	5	NA	101	NA	NA	NA	Q09472
ERF	NA	NA	33	NA	NA	NA	P50548
FBN1	11	NA	NA	NA	9	14	P35555
FBXO11	NA	NA	114	NA	NA	NA	Q86XK2
FBXW7	19	NA	50	NA	NA	NA	Q969H0
FOXG1	NA	NA	90	2	NA	NA	P55316
FOXP1	NA	NA	63	NA	NA	NA	Q9H334
GABRA1	44	26	75	NA	60	21	P14867
GABRB2	47	13	90	NA	58	15	P47870
GABRB3	58	35	104	NA	68	16	P28472
GATAD2B	NA	NA	9	NA	NA	NA	Q8WXI9
GNAI1	NA	NA	NA	NA	NA	60	P63096
GNAO1	7	NA	64	46	75	NA	P09471
GNAS	NA	NA	293	NA	NA	NA	Q5JWF2
GNB1	41	6	40	4	41	NA	P62873
GNB2	NA	NA	NA	NA	NA	NA	P62879
GRIA2	NA	NA	36	NA	NA	NA	P42262
GRIN1	73	17	87	16	63	NA	Q05586
GRIN2A	112	31	225	NA	95	35	Q12879
GRIN2B	NA	NA	280	NA	NA	NA	Q13224
H3F3A	NA	NA	NA	NA	NA	NA	P84243
HDAC2	NA	NA	NA	NA	NA	NA	Q92769

HDAC4	NA	NA	16	NA	NA	NA	P56524
HDAC8	39	NA	61	5	39	NA	Q9BY41
HECW2	NA	NA	72	NA	NA	NA	Q9P2P5
HK1	10	NA	12	NA	11	NA	P19367
HNRNPD	NA	NA	NA	NA	NA	NA	Q14103
HNRNPK	NA	NA	29	NA	NA	NA	P61978
HNRNPU	NA	NA	11	NA	NA	NA	Q00839
HUWE1	16	NA	NA	NA	NA	NA	Q7Z6Z7
IQSEC2	1	NA	60	NA	1	NA	Q5JU85
KAT6A	NA	NA	23	NA	NA	NA	Q92794
KAT6B	NA	NA	22	NA	NA	NA	Q8WYB5
KCNA2	NA	NA	65	NA	NA	NA	P16389
KCNB1	NA	NA	168	NA	NA	NA	Q14721
KCND3	NA	NA	91	NA	NA	NA	Q9UK17
KCNH1	NA	NA	82	NA	NA	NA	O95259
KCNK3	18	NA	73	NA	24	1	O14649
KCNQ2	25	32	247	80	NA	NA	O43526
KCNQ3	NA	NA	89	NA	NA	NA	O43525
KCNT1	NA	NA	142	NA	NA	NA	Q5JUK3
KDM5C	1	3	129	NA	1	3	P41229
KDM6A	31	NA	73	NA	28	NA	O15550
KIDINS220	NA	NA	NA	NA	NA	NA	Q9ULH0
KIF11	25	NA	50	NA	23	NA	P52732
KIF1A	55	36	253	NA	2	NA	Q12756
KIF5C	NA	NA	46	NA	NA	NA	O60282
KMT2A	4	NA	NA	NA	NA	NA	Q03164
KMT2B	NA	NA	NA	NA	NA	NA	Q9UMN6
KMT2C	NA	NA	NA	NA	NA	NA	Q8NEZ4
KMT2D	1	NA	NA	NA	NA	NA	O14686
KMT5B	3	NA	62	NA	3	NA	Q4FZB7
KRAS	26	15	33	24	NA	NA	P01116
MAP2K1	23	15	35	7	1	72	Q02750
MAP3K7	26	NA	75	NA	33	NA	O43318

MED12	NA	NA	127	NA	NA	NA	Q93074
MED13L	NA	NA	80	NA	NA	NA	Q71F56
MEF2C	NA	NA	50	NA	NA	NA	Q06413
MEIS2	NA	NA	34	NA	NA	NA	O14770
MIB1	2	NA	NA	NA	2	NA	Q86YT6
MN1	NA	NA	NA	NA	NA	NA	Q10571
MSL2	NA	NA	NA	NA	NA	NA	Q9HCI7
MTOR	105	NA	123	NA	NA	NA	P42345
MYT1L	NA	NA	65	NA	NA	NA	Q9UL68
NAA10	1	6	48	NA	123	NA	P41227
NAA15	28	NA	26	NA	16	NA	Q9BXJ9
NACC1	NA	NA	6	NA	NA	NA	Q96RE7
NF1	33	26	NA	NA	17	NA	P21359
NFIA	NA	NA	23	NA	NA	NA	Q12857
NFIX	NA	NA	61	NA	NA	NA	Q14938
NONO	NA	NA	NA	NA	NA	NA	Q15233
NOTCH1	NA	NA	58	NA	NA	NA	P46531
NR2F1	NA	NA	53	NA	NA	NA	P10589
NR4A2	NA	NA	20	NA	NA	NA	P43354
NSD1	15	7	329	NA	15	16	Q96L73
NSD2	NA	NA	NA	NA	NA	NA	O96028
OGT	3	NA	8	NA	4	NA	O15294
PACS1	NA	NA	NA	NA	NA	NA	Q6VY07
PACS2	NA	NA	9	NA	NA	NA	Q86VP3
PBX1	NA	NA	21	NA	NA	NA	P40424
PCBP2	NA	NA	NA	NA	NA	NA	Q15366
PCDH19	39	19	188	NA	45	17	Q8TAB3
PDHA1	81	49	100	45	61	1	P08559
PHF12	NA	NA	NA	NA	NA	NA	Q96QT6
PHF21A	NA	NA	NA	NA	NA	NA	Q96BD5
PHF6	NA	2	9	NA	NA	NA	Q8IWS0
PHIP	NA	NA	107	NA	NA	NA	Q8WWQ0
PIK3CA	95	46	89	53	74	43	P42336

PIK3R1	10	NA	56	NA	NA	NA	P27986
PIK3R2	NA	NA	47	NA	NA	NA	O00459
POGZ	NA	NA	18	NA	NA	NA	Q7Z3K3
POU3F3	NA	NA	8	NA	NA	NA	P20264
PPM1D	NA	NA	NA	NA	NA	NA	O15297
PPP1CB	NA	NA	NA	NA	NA	NA	P62140
PPP2CA	31	14	38	12	84	NA	P67775
PPP2R1A	36	NA	39	NA	36	NA	P30153
PPP2R5D	NA	NA	2	NA	NA	NA	Q14738
PPP3CA	57	NA	69	NA	36	NA	Q08209
PRPF8	31	NA	33	NA	6	NA	Q6P2Q9
PRR12	NA	NA	NA	NA	NA	NA	Q9ULL5
PSMC5	NA	NA	NA	NA	NA	NA	P62195
PTEN	21	10	50	40	NA	NA	P60484
PTPN11	22	25	26	21	NA	NA	Q06124
PUF60	20	NA	51	NA	23	NA	Q9UHX1
PURA	NA	NA	69	4	NA	NA	Q00577
QRICH1	NA	NA	NA	NA	NA	NA	Q2TAL8
RAB14	NA	NA	NA	NA	NA	NA	P61106
RAC1	25	1	21	1	54	NA	P63000
RERE	NA	NA	7	NA	NA	NA	Q9P2R6
RHOBTB2	NA	NA	28	NA	NA	NA	Q9BYZ6
RIT1	3	1	31	NA	NA	NA	Q92963
RPS6KA3	NA	NA	43	4	6	6	P51812
SATB1	NA	NA	NA	NA	NA	NA	Q01826
SATB2	NA	NA	67	NA	NA	NA	Q9UPW6
SCN1A	100	72	521	334	NA	NA	P35498
SCN2A	150	90	384	78	NA	NA	Q99250
SCN8A	NA	NA	381	37	NA	NA	Q9UQD0
SET	1	NA	19	NA	NA	NA	Q01105
SETD1B	NA	NA	87	NA	NA	NA	Q9UPS6
SETD2	2	NA	81	NA	4	NA	Q9BYW2
SHOC2	NA	NA	NA	NA	NA	NA	Q9UQ13

SIN3A	NA	NA	NA	NA	NA	NA	Q96ST3
SLC2A1	58	20	73	26	NA	NA	P11166
SLC35A2	NA	NA	21	NA	NA	NA	P78381
SLC6A1	NA	NA	91	1	NA	NA	P30531
SMAD4	33	4	87	NA	NA	NA	Q13485
SMARCA2	43	18	258	1	NA	40	P51531
SMARCA4	64	NA	246	NA	NA	NA	P51532
SMARCB1	30	13	41	NA	NA	NA	Q12824
SMC1A	3	8	45	17	23	NA	Q14683
SMC3	22	NA	59	NA	NA	NA	Q9UQE7
SNAP25	NA	NA	11	NA	NA	NA	P60880
SOX10	NA	NA	85	NA	NA	NA	P56693
SOX11	NA	NA	68	NA	NA	NA	P35716
SOX5	NA	NA	67	NA	NA	NA	P35711
SPEN	NA	NA	NA	NA	NA	NA	Q96T58
SPOP	24	NA	21	NA	24	NA	O43791
SPTAN1	NA	NA	53	NA	NA	NA	Q13813
SRCAP	NA	NA	NA	NA	NA	NA	Q6ZRS2
SRSF1	NA	NA	NA	NA	NA	NA	Q07955
STAG1	56	NA	72	NA	NA	NA	Q8WVM7
STAG2	20	NA	51	NA	3	NA	Q8N3U4
STXBP1	NA	NA	102	21	NA	NA	P61764
SYNCRIP	NA	NA	NA	NA	NA	NA	O60506
SYNGAP1	NA	NA	78	NA	NA	NA	Q96PV0
SYT1	1	NA	22	NA	NA	NA	P21579
TAF1	17	NA	78	NA	NA	NA	P21675
TAOK1	NA	NA	17	NA	NA	NA	Q7L7X3
TBL1XR1	18	19	47	3	NA	NA	Q9BZK7
TCF4	NA	NA	57	NA	NA	NA	P15884
TCF7L2	NA	NA	NA	NA	NA	NA	Q9NQB0
TFE3	NA	NA	NA	NA	NA	NA	P19532
TLK2	1	NA	46	NA	NA	NA	Q86UE8
TRAF7	NA	NA	17	NA	NA	NA	Q6Q0C0

TRIO	19	NA	NA	NA	NA	NA	O75962
TRIP12	NA	NA	35	NA	NA	NA	Q14669
U2AF2	NA	NA	NA	NA	NA	NA	P26368
UBE3A	24	NA	31	NA	NA	NA	Q05086
UBTF	NA	NA	NA	NA	NA	NA	P17480
UPF1	NA	NA	NA	NA	NA	NA	Q92900
USP7	27	NA	28	NA	18	NA	Q93009
USP9X	11	NA	59	NA	12	NA	Q93008
VCP	2	NA	3	NA	5	NA	P55072
WDR26	NA	NA	40	NA	NA	NA	Q9H7D7
WDR37	NA	NA	17	NA	NA	NA	Q9Y2I8
WDR45	NA	NA	NA	NA	NA	NA	Q5MNZ6
YY1	NA	NA	44	NA	NA	NA	P25490
ZBTB18	NA	NA	72	NA	NA	NA	Q99592
ZBTB20	NA	NA	83	NA	NA	NA	Q9HC78
ZEB2	NA	NA	61	NA	NA	NA	O60315
ZMIZ1	NA	NA	5	NA	NA	NA	Q9ULJ6
ZMYM2	NA	NA	NA	NA	NA	NA	Q9UBW7
ZMYND11	NA	NA	6	NA	4	NA	Q15326
ZNF148	NA	NA	NA	NA	NA	NA	Q9UQR1

Supplementary Table 7: Characteristic functional features associated with Essential3D sites. The table lists 12 functional features (position-specific functional feature annotations collected from the UniProt database) that are significantly enriched for Essentiial3D sites in proteins structures (191 NDD proteins, predicted structures in the monomeric-AF set), compared to the rest of the protein residues. The name and definition of the features are listed in the 1st column, the odds ratio (OR) enrichment and P values from two-tailed Fisher's Exact test are given in the 2nd column, and the 3rd column lists the genes that have an Esseetial3D site in the structure of the encoded proteins overlapped with the corresponding feature.

Feature name (definition according to UniProt)	OR enrichemnt (*P* value) of Essential3D sites on the feature	Genes with Essential3D sites identified in the encoded proteins' structures
Intramembrane region (the extent of a region that is buried within a membrane, but does not cross it)	14.95 (1.2e-227)	CACNA1C;CLCN4;GRIA2;GRIN1;GRIN2A;GRIN2B;KCNA2;KCNB1;KCND3;KCNH1;KCNK3;KCNQ2;KCNQ3;SCN1A;SCN2A;SCN8A
Nucleotide phosphate binding region (a region in the protein which binds nucleotide phosphates)	7.1 (1.7e-113)	ABL1;ARF1;ATRX;BMPR2;CDK13;CDK8;CDKL5;CHD2;CHD3;CHD4;CHD8;CSNK2A1;DDX3X;DNM1;DYNC1H1;DYRK1A;EEF1A2;GNAO1;GNAS;KIF11;KIF1A;KIF5C;KRAS;MAP2K1;MAP3K7;RAC1;RIT1;SCN8A;SMARCA2;SMARCA4;TAOK1;TLK2
Active site (residues that are directly involved in catalysis and activity of an enzyme)	7.4 (1.7e-05)	ABL1;ATP1A3;BMPR2;BRAF;CDK13;CDKL5;DYRK1A;HDAC8;KAT6A;PPP2CA;PPP3CA;PTEN;UBE3A
Binding site (residues that interact with another checmial enity, like physiological ligands, co-enzymes, prosthetic group, etc.)	6.9 (4.1e-14)	ABL1;BRAF;CDK13;CDK8;CDKL5;CLCN4;CREBBP;CSNK2A1;DYRK1A;EP300;GNAO1;GNAS;GRIN1;GRIN2A;GRIN2B;HDAC8;KDM5C;KMT5B;MAP3K7;NSD1;PTPN11;SETD1B;SETD2;SLC2A1
DNA binding domain (residues of a DNA binding domain)	7.2 (1.2e-262)	ERF;FOXG1;FOXP1;MEF2C;MEIS2;NFIA;NFIX;NR2F1;NR4A2;PBX1;POU3F3;SATB2;SOX10;SOX11;SOX5;TAF1
Transmembrane region (the extent of a membrane-spanning region of the protein)	6.3 (<1.0e-300)	ATP1A3;CACNA1A;CACNA1C;CACNA1E;CACNA1G;CLCN4;GABRA1;GABRB2;GABRB3;GRIA2;GRIN1;GRIN2A;GRIN2B;KCNA2;KCNB1;KCND3;KCNH1;KCNK3;KCNQ2;KCNQ3;KCNT1;SCN1A;SCN2A;SCN8A;SLC2A1;SLC35A2;SLC6A1
Metal binding site (residues that bind to a metal ion)	5.1 (1.1e-15)	ATP1A3;GNAO1;GNAS;HDAC8;KDM5C;KDM6A;KMT5B;NOTCH1;PPP2CA;PPP3CA;SETD2;SLC6A1;YY1

Domain (a segment in the protein with a specific combination of secondary structures organized into a characteristic three-dimensional structure or fold)	3.0 (<1.0e-300)	ABL1;ADCY5;ARHGEF9;ARID1A;ARID1B;ATRX;BMPR2;BRAF;CAMK2B;CAMTA1;CASK;CDK13;CDK8;CDKL5;CHD2;CHD3;CHD4;CHD8;CREBBP;CSNK2A1;DDX3X;DNM1;DNM1L;DNMT3A;DYRK1A;EEF1A2;EP300;FBN1;GNAO1;GNAS;HECW2;HK1;HNRNPK;HNRNPU;HUWE1;IQSEC2;KAT6A;KAT6B;KCNT1;KDM5C;KDM6A;KIF11;KIF1A;KIF5C;KMT2A;KMT2D;KMT5B;MAP2K1;MAP3K7;MEF2C;MTOR;NAA10;NACC1;NF1;NOTCH1;NSD1;PCDH19;PHIP;PIK3CA;PIK3R1;PIK3R2;POU3F3;PRPF8;PTEN;PTPN11;PUF60;RERE;RHOBTB2;RPS6KA3;SETD1B;SETD2;SMAD4;SMARCA2;SMARCA4;SMC1A;SMC3;SNAP25;SPOP;SPTAN1;STAG1;STAG2;SYNGAP1;SYT1;TAF1;TAOK1;TCF4;TLK2;TRIO;TRIP12;UBE3A;USP7;USP9X;WDR26
Motif (short, usually less than 20 amino acids, conserved sequence motifs of biological significance)	2.7 (2.3e-17)	ABL1;ARID1B;ATRX;CACNA1C;CHD2;CHD3;CHD4;CLCN4;DDX3X;KCNA2;KCNB1;KCND3;KCNQ2;KCNQ3;KRAS;NR4A2;PPP3CA;RAC1;SMARCA2;SMARCA4;SOX10
Zinc finger (residues of zinc fingers in prteins)	2.3 (1.8e-67)	ADNP;BCL11A;BCL11B;BRPF1;CREBBP;CTCF;DNMT3A;DPF2;EBF3;EP300;KAT6B;KDM5C;MYT1L;NR2F1;NR4A2;NSD1;PHF6;POGZ;YY1;ZBTB18;ZBTB20;ZEB2;ZMYND11
Repeat (repeated sequence motifs or repeated domains within the protein)	2.2 (8.0e-201)	CACNA1A;CACNA1C;CACNA1E;CACNA1G;CAMTA1;CLTC;CTNNB1;FBXO11;FBXW7;GNB1;MTOR;NAA15;NOTCH1;OGT;PHIP;PPP2R1A;SCN1A;SCN2A;SCN8A;SMARCB1;SPTAN1;TBL1XR1;TRAF7;WDR26
Topological domain (non-membrane regions of membrane-spanning proteins)	1.1 (0.0028)	ADCY5;ATP1A3;BMPR2;CACNA1A;CACNA1C;CACNA1E;CACNA1G;GABRA1;GABRB2;GABRB3;GRIA2;GRIN1;GRIN2A;GRIN2B;KCNA2;KCNB1;KCND3;KCNH1;KCNK3;KCNQ2;KCNQ3;KCNT1;NOTCH1;PCDH19;SCN1A;SCN2A;SCN8A;SLC2A1;SLC6A1;SYT1

7. Discussion

In the current landscape of clinical management for rare, potentially genetic diseases, verifying the genetic basis of these conditions is paramount. Uncovering disease-causing variants and understanding their pathogenicity and molecular functions are essential for accurate prognosis and the implementation of targeted therapies. Innovative sequencing techniques have made the exponential discovery of novel genetic variants possible, but the challenge to interpret these variants remains.

Paralogous pathogenic variants serve as a proxy for variant pathogenicity and thus increase the application of two criteria of the standard guidelines for variant pathogenicity interpretation four-fold. The standard guidelines proposed by the American College of Genetics and Genomics (ACMG) to interpret the variant's pathogenicity consider multiple criteria that include clinical, functional, and genetic information (Richards *et al.*, 2015). However, for only a few of them, there is available information to apply these criteria. Consequently, many missense variants are classified as variants of uncertain significance. Two ACMG criteria provide evidence for pathogenicity if an established pathogenic variant is observed at the same position as the novel variant and harbors the same (PS1) or a different (PM5) amino acid exchange. We demonstrated that these criteria can be extended by considering pathogenic variants observed in paralogous genes as residues overlapping with paralogous pathogenic variants are enriched for pathogenic vs control variants (OR = 6.7) (Brünger *et al.*, in preparation). By considering paralogous pathogenic variants, the number of residues where the criteria can be applied increases more than 4-fold. We showed that our approach extends and improves existing methods based on paralogous genes. Finally, we demonstrated in a set of >1700 clinically curated patients across voltage-gated sodium channels (VGSC) that the power of the extended criteria to provide evidence for variant pathogenicity can be further improved by integrating phenotype information, a pattern to be proven in additional gene families.

Functional effects identified in VGSCs are conserved across paralogous genes and guide precision medicine. We demonstrated that the functional redundancy across voltage-gated sodium channels could be utilized to infer the functional impact of genetic variants on the protein function (Brunklaus, Feng, *et al.*, 2022). To prove this, we collected 437 functionally tested variants from the literature, representing the most extensive set of functionally tested variants at publication across sodium channels. We observed across 38 variant pairs across paralogous genes that the functional effect on the protein function, specifically whether the variant causes a gain (GoF) or loss (LoF) of the protein function, is concordant in 92% of these cases. Knowledge of the functional impact of the variants in these genes is vital to enable

targeted treatments for patients with severe cardiac or neurodevelopmental disorders (NDDs). Currently, all functionally tested variants represent less than 1% of all potential missense variants in these genes. Thus, inferring the functional effect through functionally tested paralogous genes can be highly impactful. In addition to the functional concordance of conserved paralogous variants, we observed correlations between the location of the variants in the protein structure and the functional effect. For example, variants close to the pore region where the selectivity filter is located were causing a loss of the protein function (OR = 42.1, P<0.001). The collected data represent the first GoF versus LoF topological map of SCN proteins, indicating shared patterns of functional effects aiding variant analysis and guiding precision therapy. Finally, our study builds the foundation of a recently published *in-silico* approach to predict the function of variants across the VGSCs (Boßelmann *et al.*, 2022).

Comprehensive identification of structural regions conserved across ion channels inform variant pathogenicity and variant function. Ion channels are implicated in a broad range of mild to serve diseases (Hübner and Jentsch, 2002). Current medications are often not sufficient to relieve patients with severe symptoms. Several approaches are being tested to enable targeted treatment of patients. However, they do not only require a precise understanding of whether a variant is pathogenic but also of the underlying molecular disease mechanism. We developed a 3D framework to demonstrate that the pore is the most critical region across all ion channels and that variant pathogenicity is correlated with the distance from the pore (Brünger *et al.*, 2022). Variants within a 5 Å distance from the pore axis center had the highest enrichment for pathogenic versus control variants (OR>20). Within residues closely located at the pore, hydrophobicity played a key role in variant pathogenicity, which aligns with previous observations (Xenakis *et al.*, 2021). By demonstrating that our pore distance approach correlates with *in vitro* functional readouts and clinical phenotypes in patients with NDDs, we propose that our results could be integrated into clinical decision support algorithms.

Systematic identification of functionally essential protein regions across NDD-associated genes provides insights into variant pathogenicity and the molecular disease mechanism. It has been previously shown that 3D structural information of proteins helps to identify mutational hotspots and can unveil perturbed biological pathways (Kamburov *et al.*, 2015; Sivley *et al.*, 2018; Iqbal *et al.*, 2020; Pandurangan and Blundell, 2020). We identified essential sites in 189 NDD-associated genes (Iqbal *et al.*, 2022) by combining structural information with additional genomic indications. In addition, we expanded the analyzed set of proteins by including protein structure predictions that became recently accessible (Varadi *et al.*, 2022). All our annotations highlighting critical protein residues (Brünger *et al.*, 2022; Iqbal *et al.*, 2022) can be integrated into the guidelines following the PM1

criterion, which can be applied when a variant is located in a mutational hotspot and well-established functional domain. The PM1 criterion presents only moderate evidence for pathogenicity (Richards *et al.*, 2015). However, following recent studies (Pejaver *et al.*, 2022) that suggested the strength of a criterion (moderate for PM1) could be adjusted based on the observed effect sizes, for instance, for residues located at the pore (OR>20), the PM1 criterion could be annotated as strong evidence for pathogenicity.

The benefits and limitations of the novel approaches outlined in my studies to determine variant pathogenicity and the underlying functional changes of variants on protein function are discussed below.

7.1 Conserved structural regions inform variant pathogenicity and protein function

Pathogenic variants frequently cluster in functionally critical regions of protein structures. Our research (Brünger *et al.*, 2022; Iqbal *et al.*, 2022), along with studies conducted by others (Meyer *et al.*, 2016; Laskowski *et al.*, 2020), has shown that identifying these clusters can aid in variant pathogenicity assessments and provide insights into the molecular mechanisms underlying protein function, either individually or in combination with other evidence. However, the location of these hotspots is typically specific to individual proteins or protein classes, and the importance of features derived from protein structures may vary across different protein classes (Iqbal *et al.*, 2020). Due to their structural and functional variety, ion channels represent a particularly challenging class of proteins regarding their variant interpretation (Millar, Bruce and Brown, 2007). In our study, we demonstrated for the first time that, despite this variety, the pore section located inside the membrane is the most critical region for disease-causing mutations across the structurally diverse group of ion channels investigated (Brünger *et al.*, 2022). Our findings generalize previous molecular biology studies on individual ion channel-encoding genes, which have reported that ion channel residues at the pore are more susceptible to disease-causing mutations (Zuberi *et al.*, 2011; Hernandez *et al.*, 2017; Goto *et al.*, 2019; Li *et al.*, 2019).

The location of a variant within a protein structure can offer valuable insights into the effect a variant may have on protein function. We (Salles *et al.*, 2021; Brunklaus, Brünger, *et al.*, 2022; Brunklaus, Feng, *et al.*, 2022; Johannesen *et al.*, 2022) and others (Gerasimavicius, Livesey and Marsh, 2022) showed that the mutational effect of a variant could be associated with its position within the protein. Employing our structural framework

developed for the ion channel proteins (Brünger *et al.*, 2022), we demonstrated a significant correlation between the variant position and its functional effect. These observations align with reports in the literature; for example, LoF variants in NMDA receptors tend to cluster at the ligand-binding domains, while GoF variants are typically located at the pore inside the membrane (Strehlow *et al.*, 2019). A recent study by Thompson *et al.*, 2023 revealed that, based on multiple functional parameters such as ion flow magnitude, activation, and inactivation kinetics, channel function is often more complex than the binary classification of variants as causing either gain or loss of protein function. Although this study presents a new layer of complexity in understanding protein function using multiple electrophysiological parameters, variants at the same protein domain frequently exhibited similar mutational effects on protein function. For example, they observed a cluster of variants in the domain IV and the C-terminus that all exhibited depolarized shifts in the voltage dependence of channel inactivation. These results suggest that the functional similarities which we observed hold even beyond the GoF versus LoF classification including the level of individual electrophysiological parameters (Thompson *et al.*, 2023).

7.2 Knowledge transfer from established pathogenic and functionally tested variants to variants in paralogous genes informs variant pathogenicity and protein function

To date, only a limited number of studies have incorporated information from paralogous genes into variant classification and functional effect prediction. In previous research, individual pathogenic variants located at analogous protein positions in paralogous genes have been shown to enhance variant classification and resolve variants of unknown significance in *SCN5A*, *RYR1*, *KCNQ1*, and *KCNQ4* (Ware *et al.*, 2012; Walsh *et al.*, 2014; Pérez-Palma *et al.*, 2020). In our study, we present, for the first time, the feasibility of this approach on an exome-wide scale (Brünger *et al.*, in preparation). We identified over 500 gene families in which patient versus control variants demonstrated enrichment of paralogous pathogenic variants (OR = 6.3). Our findings align with other exome-wide studies focused on paralogs, which have shown that protein regions with significant clustering of pathogenic variants across paralogous genes are highly enriched for pathogenic versus control variants (Pérez-Palma *et al.*, 2020). A pathogenic missense variant with the same amino acid located in the same gene and at the same position as a novel variant is considered strong evidence for pathogenicity (Richards *et al.*, 2015). We demonstrated that this criterion could be extended to paralogous genes and thus enables the application of the criterion to more than >140,000

additional protein residues, which are otherwise not captured by the ACMG criteria. Our proposal to expand the framework lines up with previous studies that suggest modifications or extensions to the original guidelines (Brandt *et al.*, 2020; Patel *et al.*, 2021; Masson *et al.*, 2022). Furthermore, adapting the strength of different evidence criteria has been suggested (Brnich *et al.*, 2019; Pejaver *et al.*, 2022). Applied to our approach, the strength of the paralog-based criterion can be individually calculated for each gene family, resulting in pieces of evidence that range from supporting to strong evidence. Although the average strength of this criterion is lower than that of the PS1 criterion, moderate (our extended criteria) compared to strong (PS1), the proposed extension does not prevent the application of the PS1 criteria when applicable but rather provides evidence for pathogenicity when information needed to apply the current PS1 is not available.

Estimating the effect of genetic variants on protein function is particularly challenging for missense variants. Prior research has proposed that similar to the pathogenicity, the functional effects of variants might be conserved across paralogous variants (Brunklaus *et al.*, 2020). In our study (Brunklaus, Feng, *et al.*, 2022), we demonstrated that the functional effects of conserved paralogous variants are highly conserved (>90%) across VGSCs. This observation is in line with the evolutionarily conserved nature of VGSCs encoding genes and previous findings, which indicate that, for example, variants causing a loss of protein function are frequently located at the selectivity filter, while variants causing an increase in protein function are situated at the inactivation gate of the channels (Brunklaus *et al.*, 2020). Since functional testing is currently limited and less than 1% of potential missense variants have been characterized across VGSCs, transferring functional effects from one gene to another can substantially enhance the coverage of missense variants with known functional effects. As a functionally tested variant in one gene can potentially inform variants in the remaining ten genes encoding for VGSCs, this approach presents a powerful resource for improving our understanding of these variants. In addition, whenever a novel variant is functionally tested in one gene, our approach enables us to predict the function of all conserved variants in other VGSC encoding genes, thus in up to ten additional cases. Our method, in contrast to in-*silico* prediction tools designed exclusively to predict the functional impact of VGSCs, predicts the function for fewer variants (Heyne *et al.*, 2020; Boßelmann *et al.*, 2022). However, it achieves greater accuracy when applied. In addition, by aggregating the largest set of functionally tested variants we could demonstrate that the prediction performance of the *in-silico* prediction tool (Heyne *et al.*, 2020) differs depending on the protein region where the variant is located, thus providing a valuable dataset to test and validate existing and novel *in-silico* methods and identify potential weaknesses.

7.3 Essential sites identified through a combination of evolutionary and structural features inform variant pathogenicity and provide insights into molecular disease mechanisms

Structural information and evolutionary conservation across paralogous genes provide complementary information. In several studies within this thesis, we demonstrated that knowledge transfer from paralogous genes (Brunklaus, Feng, *et al.*, 2022; Brünger *et al.*, in preparation), as well as the location of variants in the protein, particularly in the protein structure(Salles *et al.*, 2021; Brünger *et al.*, 2022; Brunklaus, Brünger, *et al.*, 2022; Johannesen *et al.*, 2022; Zarate *et al.*, 2023), can inform variant pathogenicity and functional effects of proteins caused by genetic variants. Integrating multiple levels of information to inform the desired outcome is shared across various disciplines. For instance, complex in-*silico* prediction models designed to discriminate between pathogenic and benign variants often incorporate various types of information, including structural data and evolutionary conservation (Rentzsch *et al.*, 2019; Qi *et al.*, 2021). Another example are the ACMG guidelines, which aggregate information from 28 individual criteria to interpret a variant (Richards *et al.*, 2015). Building on this concept, we developed a consensus approach for 243 NDD-associated genes that considers paralog conservation and constraint for missense variants in the context of protein structures, as well as an enrichment of patient versus control variants within the structures, to identify essential protein regions for each protein (Iqbal *et al.*, 2022). In concordance with the principles of the variant interpretation frameworks, the overall likelihood that a residue is pathogenic or, in this case, essential increases when multiple lines of evidence support its critical nature. As a result, our annotation presents not only a powerful proxy for variant pathogenicity but also a tool to identify critical functional elements such as binding sites or protein-protein interaction sites or to highlight residues as essential, whose molecular functional roles need to be determined in future studies. For example, we validated previously described disease mechanisms, such as ATP binding across the GNAO1 protein (Kelly *et al.*, 2019), and developed hypotheses for yet unstudied disease mechanisms.

7.4 Web-based tools improve access to novel scores and annotations

Prior research has shown that web-based environments significantly improve access to novel scores and annotations (Afgan *et al.*, 2016). Due to this, we have made our annotations accessible through various web-based solutions (https://es-ndd.broadinstitute.org, http://SCN-viewer.broadinstitute.org). Additionally, integrating our annotations into

bioinformatics pipelines, such as Annovar (Wang, Li and Hakonarson, 2010) or the Variant Effect Predictor (VEP) (McLaren *et al.*, 2016) that are frequently used to obtain multiple annotations simultaneously, could further enhance their widespread use. Variant interpretation, typically facilitated by genetic counselors and clinical geneticists, must be conducted in a reasonable timeframe and incorporate different lines of evidence (Richards *et al.*, 2015). Consequently, in a clinical setting using a very large number of resources is not feasible. To address this issue, we developed gene and gene family-specific portals that provide one-stop access to functional, genetic, and phenotypic information for evaluating variants following ACMG guidelines (https://scn-portal.broadinstitute.org/, https://grin-portal.broadinstitute.org/, https://slc6a1-portal.broadinstitute.org/). Unlike more comprehensive web-based variant annotation tools (Chang and Wang, 2012; McLaren *et al.*, 2016), our web portals focus on a limited number of genes. However, they incorporate gene-specific annotations and scores and expert-curated functional and clinical information that is typically not accessible through other large platforms.

7.5 Summary of the conducted research outlined in this thesis

Our developed methods presented in this thesis offer improvements for two major limitations in clinical care and the development and application of targeted treatments that are determining the pathogenicity of variants and subsequently evaluate the impact of pathogenic variants on the protein function. Currently, many genetic variants are classified as variants of uncertain significance, because the number of criteria of the ACMG guidelines that can be applied for a variant is limited due to missing information, such as the inheritance of the genetic variants their functional consequences, or whether another pathogenic variant has been already observed at the same position. We demonstrated that the applicability of two of these criteria can be extended by considering paralogous pathogenic variants thereby increasing the number of residues where the criteria can be applied more than 4-fold. Especially for patients with severe diseases accurate clinical variant interpretation is urgently needed to enable optimized clinical care. Due to extensive collaboration due to which we got access to high-quality genetic, clinical, and functional datasets, we focus on three studies on the development of approaches that improve pathogenicity and functional variant interpretation in specific groups of genes associated with severe disease. We identified the most functionally essential sites across genes associated with mild to severe NDDs. Most NDD-associated genes were just recently discovered and are not well studied. Our annotations can highlight the essential sites in these genes and provide insights into the molecular disease mechanism. A second group of genes implicated in severe neurological and cardiac disorders encodes ion channels.

Based on the functionally important properties shared across all ion channels which we systematically identified, we developed a structural framework that showed that a close variant distance from the pore serves as a proxy for variant pathogenicity in all ion channels investigated. Pathogenic variants in ion channels are known to either increase or decrease the protein function, with complementary treatment opportunities. We could comprehensively demonstrate for a subset of ion channel encoding genes that encode for VGSCs that the position of the variants is associated with the localization in the protein structure. In addition, we showed this set of genes that knowledge of the functional effect of a variant in one gene can be transferred to conserved paralogous genes and thus enrich the number of variants whose function is informed by a single experimental test up to 10-fold. Our findings can inform the impact of variant function channel function where experimental testing is not available and pave the way toward advanced function predictions. Overall, our results improve the accurate interpretation of variant pathogenicity and function, especially for severely affected patients, which eventually can support clinical decision-making for targeted therapies.

8. References

Addis, L. *et al.* (2017) 'Epilepsy-associated GRIN2A mutations reduce NMDA receptor trafficking and agonist potency - molecular profiling and functional rescue', *Scientific Reports*, 7(1), p. 66. Available at: https://doi.org/10.1038/s41598-017-00115-w.

Afgan, E. *et al.* (2016) 'The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update', *Nucleic Acids Research*, 44(Web Server issue), pp. W3–W10. Available at: https://doi.org/10.1093/nar/gkw343.

Allen, N.M. *et al.* (2019) 'Genetic potassium channel-associated epilepsies: Clinical review of the Kv family', *European Journal of Paediatric Neurology* [Preprint]. Available at: https://doi.org/10.1016/j.ejpn.2019.12.002.

Araya, C.L. *et al.* (2016) 'Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations', *Nature Genetics*, 48(2), pp. 117–125. Available at: https://doi.org/10.1038/ng.3471.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. Available at: https://doi.org/10.1038/nature15393.

Backwell, L. and Marsh, J.A. (2022) 'Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm', *Annual Review of Genomics and Human Genetics*, 23, pp. 475–498. Available at: https://doi.org/10.1146/annurev-genom-111221-103208.

Berman, H.M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*, 28(1), pp. 235–242. Available at: https://doi.org/10.1093/nar/28.1.235.

Boßelmann, C.M. *et al.* (2022) 'Learning with phenotypic similarity improves the prediction of functional effects of missense variants in voltage-gated sodium channels'. bioRxiv, p. 2022.09.29.510111. Available at: https://doi.org/10.1101/2022.09.29.510111.

Brandt, T. *et al.* (2020) 'Adapting ACMG/AMP sequence variant classification guidelines for single-gene copy number variants', *Genetics in Medicine*, 22(2), pp. 336–344. Available at: https://doi.org/10.1038/s41436-019-0655-2.

Brnich, S.E. *et al.* (2019) 'Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework', *Genome Medicine*, 12(1), p. 3. Available at: https://doi.org/10.1186/s13073-019-0690-2.

Brünger, T. *et al.* (2022) 'Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes', *Brain: A Journal of Neurology*, p. awac305. Available at: https://doi.org/10.1093/brain/awac305.

Brünger, T. *et al.* (in preparation) 'Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria'.

Brunklaus, A. *et al.* (2012) 'Prognostic, clinical and demographic features in SCN1A mutation-positive Dravet syndrome', *Brain: A Journal of Neurology*, 135(Pt 8), pp. 2329–2336. Available at: https://doi.org/10.1093/brain/aws151.

Brunklaus, A. *et al.* (2020) 'Biological concepts in human sodium channel epilepsies and their relevance in clinical practice', *Epilepsia*, 61(3), pp. 387–399. Available at: https://doi.org/10.1111/epi.16438.

Brunklaus, A., Feng, T., *et al.* (2022) 'Gene variant effects across sodium channelopathies predict function and guide precision therapy', *Brain: A Journal of Neurology*, p. awac006. Available at: https://doi.org/10.1093/brain/awac006.

Brunklaus, A., Brünger, T., *et al.* (2022) 'The gain of function SCN1A disorder spectrum: novel epilepsy phenotypes and therapeutic implications', *Brain: A Journal of Neurology*, p. awac210. Available at: https://doi.org/10.1093/brain/awac210.

Catterall, W.A. (2012) 'Sodium Channel Mutations and Epilepsy', in J.L. Noebels et al. (eds) *Jasper's Basic Mechanisms of the Epilepsies*. 4th edn. Bethesda (MD): National Center for Biotechnology Information (US). Available at: http://www.ncbi.nlm.nih.gov/books/NBK98185/ (Accessed: 5 February 2022).

Chang, X. and Wang, K. (2012) 'wANNOVAR: annotating genetic variants for personal genomes via the web', *Journal of medical genetics*, 49(7), pp. 433–436. Available at: https://doi.org/10.1136/jmedgenet-2012-100918.

Chen, G.T. and Geschwind, D.H. (2022) 'Challenges and opportunities for precision medicine in neurodevelopmental disorders', *Advanced Drug Delivery Reviews*, 191, p. 114564. Available at: https://doi.org/10.1016/j.addr.2022.114564.

Chen, W.-H. *et al.* (2013) 'Human Monogenic Disease Genes Have Frequently Functionally Redundant Paralogs', *PLoS Computational Biology*, 9(5), p. e1003073. Available at: https://doi.org/10.1371/journal.pcbi.1003073.

Chiron, C. *et al.* (2000) 'Stiripentol in severe myoclonic epilepsy in infancy: a randomised placebo-controlled syndrome-dedicated trial. STICLO study group', *Lancet (London, England)*, 356(9242), pp. 1638–1642. Available at: https://doi.org/10.1016/s0140-6736(00)03157-3.

Chuang, S.-H. and Reddy, D.S. (2018) 'Genetic and Molecular Regulation of Extrasynaptic GABA-A Receptors in the Brain: Therapeutic Insights for Epilepsy', *The Journal of Pharmacology and Experimental Therapeutics*, 364(2), pp. 180–197. Available at: https://doi.org/10.1124/jpet.117.244673.

Claes, L. *et al.* (2001) 'De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy', *American Journal of Human Genetics*, 68(6), pp. 1327–1332. Available at: https://doi.org/10.1086/320609.

Claes, L.R.F. *et al.* (2009) 'The SCN1A variant database: a novel research and diagnostic tool', *Human Mutation*, 30(10), pp. E904-920. Available at: https://doi.org/10.1002/humu.21083.

David, A. and Sternberg, M.J.E. (2015) 'The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease', *Journal of Molecular Biology*, 427(17), pp. 2886–2898. Available at: https://doi.org/10.1016/j.jmb.2015.07.004.

Davydov, E.V. *et al.* (2010) 'Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++', *PLoS Computational Biology*, 6(12). Available at: https://doi.org/10.1371/journal.pcbi.1001025.

Dewey, F.E. *et al.* (2016) 'Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study', *Science (New York, N.Y.)*, 354(6319). Available at: https://doi.org/10.1126/science.aaf6814.

Dickerson, J.E. and Robertson, D.L. (2012) 'On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication', *Molecular Biology and Evolution*, 29(1), pp. 61–69. Available at: https://doi.org/10.1093/molbev/msr111.

Dobson, C.M. (2003) 'Protein folding and misfolding', *Nature*, 426(6968), pp. 884–890. Available at: https://doi.org/10.1038/nature02261.

Dufayard, J.-F. *et al.* (2005) 'Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases', *Bioinformatics*, 21(11), pp. 2596–2603. Available at: https://doi.org/10.1093/bioinformatics/bti325.

van El, C.G. *et al.* (2013) 'Whole-genome sequencing in health care', *European Journal of Human Genetics*, 21(6), pp. 580–584. Available at: https://doi.org/10.1038/ejhg.2013.46.

Emerson, E. (2012) 'Deprivation, ethnicity and the prevalence of intellectual and developmental disabilities', *Journal of Epidemiology and Community Health*, 66(3), pp. 218–224. Available at: https://doi.org/10.1136/jech.2010.111773.

Encinas, A.C. *et al.* (2020) 'Variable patterns of mutation density among NaV1.1, NaV1.2 and NaV1.6 point to channel-specific functional differences associated with childhood epilepsy', *PLoS ONE*, 15(8), p. e0238121. Available at: https://doi.org/10.1371/journal.pone.0238121.

ENCODE Project Consortium (2004) 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science (New York, N.Y.)*, 306(5696), pp. 636–640. Available at: https://doi.org/10.1126/science.1105136.

Epi25 Collaborative and Epi25 Collaborative (2021) 'Sub-genic intolerance, ClinVar, and the epilepsies: A whole-exome sequencing study of 29,165 individuals', *American Journal of Human Genetics*, 108(6), pp. 965–982. Available at: https://doi.org/10.1016/j.ajhg.2021.04.009.

Esposito, D. *et al.* (2019) 'MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect', *Genome Biology*, 20(1), p. 223. Available at: https://doi.org/10.1186/s13059-019-1845-6.

Fernandez-Marmiesse, A., Gouveia, S. and Couce, M.L. (2018) 'NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment', *Current Medicinal Chemistry*, 25(3), pp. 404–432. Available at: https://doi.org/10.2174/0929867324666170718101946.

Fowler, D.M. and Fields, S. (2014) 'Deep mutational scanning: a new style of protein science', *Nature Methods*, 11(8), pp. 801–807. Available at: https://doi.org/10.1038/nmeth.3027.

Frazer, J. *et al.* (2021) 'Disease variant prediction with deep generative models of evolutionary data', *Nature*, 599(7883), pp. 91–95. Available at: https://doi.org/10.1038/s41586-021-04043-8.

Gao, M., Zhou, H. and Skolnick, J. (2015) 'Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis', *Structure (London, England: 1993)*, 23(7), pp. 1362–1369. Available at: https://doi.org/10.1016/j.str.2015.03.028.

Gelb, B.D. *et al.* (2018) 'ClinGen's RASopathy Expert Panel consensus methods for variant interpretation', *Genetics in Medicine*, 20(11), pp. 1334–1345. Available at: https://doi.org/10.1038/gim.2018.3.

Gerasimavicius, L., Livesey, B.J. and Marsh, J.A. (2022) 'Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure', *Nature Communications*, 13(1), p. 3895. Available at: https://doi.org/10.1038/s41467-022-31686-6.

Glazer, A.M. *et al.* (2020) 'High-Throughput Reclassification of SCN5A Variants', *American Journal of Human Genetics* [Preprint]. Available at: https://doi.org/10.1016/j.ajhg.2020.05.015.

Goto, A. *et al.* (2019) 'Characteristics of KCNQ2 variants causing either benign neonatal epilepsy or developmental and epileptic encephalopathy', *Epilepsia*, 60(9), pp. 1870–1880. Available at: https://doi.org/10.1111/epi.16314.

Gudmundsson, S. *et al.* (2022) 'Variant interpretation using population databases: Lessons from gnomAD', *Human Mutation*, 43(8), pp. 1012–1030. Available at: https://doi.org/10.1002/humu.24309.

Hansen, K.B. *et al.* (2021) 'Structure, Function, and Pharmacology of Glutamate Receptor Ion Channels', *Pharmacological Reviews*, 73(4), pp. 298–487. Available at: https://doi.org/10.1124/pharmrev.120.000131.

Hernandez, C.C. *et al.* (2017) 'GABA A Receptor Coupling Junction and Pore GABRB3 Mutations are Linked to Early-Onset Epileptic Encephalopathy', *Scientific Reports*, 7(1), p. 15903. Available at: https://doi.org/10.1038/s41598-017-16010-3.

Heyne, H.O. *et al.* (2018) 'De novo variants in neurodevelopmental disorders with epilepsy', *Nature Genetics*, 50(7), pp. 1048–1053. Available at: https://doi.org/10.1038/s41588-018-0143-7.

Heyne, H.O. *et al.* (2020) 'Predicting functional effects of missense variants in voltage-gated sodium and calcium channels', *Science Translational Medicine*, 12(556). Available at: https://doi.org/10.1126/scitranslmed.aay6848.

Hille, B. (1989) 'The Sharpey-Schafer Lecture. Ionic channels: evolutionary origins and modern roles', *Quarterly Journal of Experimental Physiology (Cambridge, England)*, 74(6), pp. 785–804. Available at: https://doi.org/10.1113/expphysiol.1989.sp003349.

Hodgkin, A.L. and Huxley, A.F. (1952) 'A quantitative description of membrane current and its application to conduction and excitation in nerve', *The Journal of Physiology*, 117(4), pp. 500–544.

Hübner, C.A. and Jentsch, T.J. (2002) 'Ion channel diseases', *Human Molecular Genetics*, 11(20), pp. 2435–2445. Available at: https://doi.org/10.1093/hmg/11.20.2435.

Iqbal, S. *et al.* (2020) 'Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants', *Proceedings of the National Academy of Sciences of the United States of America*, 117(45), pp. 28201–28211. Available at: https://doi.org/10.1073/pnas.2002660117.

Iqbal, S. *et al.* (2022) 'Delineation of functionally essential protein regions for 242 neurodevelopmental disorders', *Brain: A Journal of Neurology*, p. awac381. Available at: https://doi.org/10.1093/brain/awac381.

Jen, J.C. *et al.* (2007) 'Primary episodic ataxias: diagnosis, pathogenesis and treatment', *Brain*, 130(10), pp. 2484–2493. Available at: https://doi.org/10.1093/brain/awm126.

Jia, X. *et al.* (2020) 'Massively parallel functional testing of MSH2 missense variants conferring Lynch syndrome risk', *American Journal of Human Genetics* [Preprint]. Available at: https://doi.org/10.1016/j.ajhg.2020.12.003.

Johannesen, K.M. *et al.* (2022) 'Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications', *Brain: A Journal of Neurology*, 145(9), pp. 2991–3009. Available at: https://doi.org/10.1093/brain/awab321.

Jurkat-Rott, K. and Lehmann-Horn, F. (2004) 'The Patch Clamp Technique in Ion Channel Research', *Current pharmaceutical biotechnology*, 5, pp. 387–95. Available at: https://doi.org/10.2174/1389201043376715.

Kamburov, A. *et al.* (2015) 'Comprehensive assessment of cancer missense mutation clustering in protein structures', *Proceedings of the National Academy of Sciences of the United States of America*, 112(40), pp. E5486–E5495. Available at: https://doi.org/10.1073/pnas.1516373112.

Kaplanis, J. *et al.* (2020) 'Evidence for 28 genetic disorders discovered by combining healthcare and research data', *Nature*, 586(7831), pp. 757–762. Available at: https://doi.org/10.1038/s41586-020-2832-5.

Karczewski, K.J. *et al.* (2020) 'The mutational constraint spectrum quantified from variation in 141,456 humans', *Nature*, 581(7809), pp. 434–443. Available at: https://doi.org/10.1038/s41586-020-2308-7.

Katsonis, P. *et al.* (2022) 'Genome interpretation using in silico predictors of variant impact', *Human Genetics*, 141(10), pp. 1549–1577. Available at: https://doi.org/10.1007/s00439-022-02457-6.

Kelly, M. *et al.* (2019) 'Spectrum of neurodevelopmental disease associated with the GNAO1 guanosine triphosphate-binding region', *Epilepsia*, 60(3), pp. 406–418. Available at: https://doi.org/10.1111/epi.14653.

Kelly, M.A. *et al.* (2018) 'Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel', *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 20(3), pp. 351–359. Available at: https://doi.org/10.1038/gim.2017.218.

Kessler, M.D. *et al.* (2020) 'De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the Amish founder population', *Proceedings of the National Academy of Sciences*, 117(5), pp. 2560–2569. Available at: https://doi.org/10.1073/pnas.1902766117.

Khanna, T. *et al.* (2021) 'Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants', *Human Genetics* [Preprint]. Available at: https://doi.org/10.1007/s00439-020-02246-z.

Kim, J.-B. (2014) 'Channelopathies', *Korean Journal of Pediatrics*, 57(1), pp. 1–18. Available at: https://doi.org/10.3345/kjp.2014.57.1.1.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511623486.

Koboldt, D.C. (2020) 'Best practices for variant calling in clinical sequencing', *Genome Medicine*, 12(1), p. 91. Available at: https://doi.org/10.1186/s13073-020-00791-w.

Koonin, E.V. (2005) 'Orthologs, paralogs, and evolutionary genomics', *Annual Review of Genetics*, 39, pp. 309–338. Available at: https://doi.org/10.1146/annurev.genet.39.073003.114725.

Lal, D. *et al.* (2020) 'Gene family information facilitates variant interpretation and identification of disease-associated genes in neurodevelopmental disorders', *Genome Medicine*, 12(1), p. 28. Available at: https://doi.org/10.1186/s13073-020-00725-6.

Landrum, M.J. *et al.* (2018) 'ClinVar: improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*, 46(D1), pp. D1062–D1067. Available at: https://doi.org/10.1093/nar/gkx1153.

Larsen, J. *et al.* (2015) 'The phenotypic spectrum of SCN8A encephalopathy', *Neurology*, 84(5), pp. 480–489. Available at: https://doi.org/10.1212/WNL.0000000000001211.

Laskowski, R.A. *et al.* (2020) 'VarSite: Disease variants and protein structure', *Protein Science: A Publication of the Protein Society*, 29(1), pp. 111–119. Available at: https://doi.org/10.1002/pro.3746.

Lek, M. *et al.* (2016) 'Analysis of protein-coding genetic variation in 60,706 humans', *Nature*, 536(7616), pp. 285–291. Available at: https://doi.org/10.1038/nature19057.

Lesk, A.M. and Chothia, C. (1980) 'How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins', *Journal of Molecular Biology*, 136(3), pp. 225–270. Available at: https://doi.org/10.1016/0022-2836(80)90373-3.

Li, J. *et al.* (2019) 'De novo GRIN variants in NMDA receptor M2 channel pore-forming loop are associated with neurological diseases', *Human Mutation*, 40(12), pp. 2393–2413. Available at: https://doi.org/10.1002/humu.23895.

Livesey, B.J. and Marsh, J.A. (2019) *Using deep mutational scanning data to benchmark computational phenotype predictors and identify pathogenic missense mutations*. preprint. Bioinformatics. Available at: https://doi.org/10.1101/855957.

Livesey, B.J. and Marsh, J.A. (2022) 'Interpreting protein variant effects with computational predictors and deep mutational scanning', *Disease Models & Mechanisms*, 15(6), p. dmm049510. Available at: https://doi.org/10.1242/dmm.049510.

Luo, X. *et al.* (2019) 'ClinGen Myeloid Malignancy Variant Curation Expert Panel recommendations for germline RUNX1 variants', *Blood Advances*, 3(20), pp. 2962–2979. Available at: https://doi.org/10.1182/bloodadvances.2019000644.

Macnee, M. *et al.* (2023) 'Data-driven historical characterization of epilepsy-associated genes', *European Journal of Paediatric Neurology*, 42, pp. 82–87. Available at: https://doi.org/10.1016/j.ejpn.2022.12.005.

Majewski, J. *et al.* (2011) 'What can exome sequencing do for you?', *Journal of Medical Genetics*, 48(9), pp. 580–589. Available at: https://doi.org/10.1136/jmedgenet-2011-100223.

Mantegazza, M. and Cestèle, S. (2018) 'Pathophysiological mechanisms of migraine and epilepsy: Similarities and differences', *Neuroscience Letters*, 667, pp. 92–102. Available at: https://doi.org/10.1016/j.neulet.2017.11.025.

Margulies, E.H. *et al.* (2007) 'Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome', *Genome Research*, 17(6), pp. 760–774. Available at: https://doi.org/10.1101/gr.6034307.

Marsh, J.A. and Teichmann, S.A. (2015) 'Structure, Dynamics, Assembly, and Evolution of Protein Complexes', *Annual Review of Biochemistry*, 84(1), pp. 551–575. Available at: https://doi.org/10.1146/annurev-biochem-060614-034142.

Masson, E. *et al.* (2022) 'Expanding ACMG variant classification guidelines into a general framework', *Human Genomics*, 16(1), p. 31. Available at: https://doi.org/10.1186/s40246-022-00407-x.

McCarthy, G.M. *et al.* (2021) 'Recessive, gain-of-function toxicity in an APOL1 BAC transgenic mouse model mirrors human APOL1 kidney disease', *Disease Models & Mechanisms*, 14(8), p. dmm048952. Available at: https://doi.org/10.1242/dmm.048952.

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122. Available at: https://doi.org/10.1186/s13059-016-0974-4.

Mersch, J. *et al.* (2018) 'Prevalence of Variant Reclassification Following Hereditary Cancer Genetic Testing', *JAMA*, 320(12), pp. 1266–1274. Available at: https://doi.org/10.1001/jama.2018.13152.

Mester, J.L. *et al.* (2018) 'Gene-specific criteria for PTEN variant curation: Recommendations from the ClinGen PTEN Expert Panel', *Human Mutation*, 39(11), pp. 1581–1592. Available at: https://doi.org/10.1002/humu.23636.

Meyer, M.J. *et al.* (2016) 'mutation3D: cancer gene prediction through atomic clustering of coding variants in the structural proteome', *Human mutation*, 37(5), pp. 447–456. Available at: https://doi.org/10.1002/humu.22963.

Millar, I.D., Bruce, J.I. and Brown, P.D. (2007) 'Ion channel diversity, channel expression and function in the choroid plexuses', *Cerebrospinal Fluid Research*, 4(1), p. 8. Available at: https://doi.org/10.1186/1743-8454-4-8.

Milman, A. *et al.* (2021) 'Genotype-Phenotype Correlation of SCN5A Genotype in Patients With Brugada Syndrome and Arrhythmic Events: Insights From the SABRUS in 392 Probands', *Circulation. Genomic and Precision Medicine*, 14(5), p. e003222. Available at: https://doi.org/10.1161/CIRCGEN.120.003222.

Ng, P.C. and Henikoff, S. (2003) 'SIFT: predicting amino acid changes that affect protein function', *Nucleic Acids Research*, 31(13), pp. 3812–3814.

Nguengang Wakap, S. *et al.* (2020) 'Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database', *European journal of human genetics: EJHG*, 28(2), pp. 165–173. Available at: https://doi.org/10.1038/s41431-019-0508-0.

Nickless, A., Bailis, J.M. and You, Z. (2017) 'Control of gene expression through the nonsense-mediated RNA decay pathway', *Cell & Bioscience*, 7(1), p. 26. Available at: https://doi.org/10.1186/s13578-017-0153-7.

Orphanet (2022) *Prevalence and incidence of rare diseases: Bibliographic data*. Available at: https://www.orpha.net/orphacom/cahiers/docs/GB/Prevalence_of_rare_diseases_by_alphabetical_list.pdf.

Pandurangan, A.P. and Blundell, T.L. (2020) 'Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning', *Protein Science*, 29(1), pp. 247–257. Available at: https://doi.org/10.1002/pro.3774.

Parenti, I. *et al.* (2020) 'Neurodevelopmental Disorders: From Genetics to Functional Pathways', *Trends in Neurosciences*, 43(8), pp. 608–621. Available at: https://doi.org/10.1016/j.tins.2020.05.004.

Patel, M.J. *et al.* (2021) 'Disease-specific ACMG/AMP guidelines improve sequence variant interpretation for hearing loss', *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 23(11), pp. 2208–2212. Available at: https://doi.org/10.1038/s41436-021-01254-2.

Paysan-Lafosse, T. *et al.* (2023) 'InterPro in 2022', *Nucleic acids research*, 51(D1), pp. D418–D427. Available at: https://doi.org/10.1093/nar/gkac993.

Pejaver, V. *et al.* (2022) 'Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria', *American Journal of Human Genetics*, 109(12), pp. 2163–2177. Available at: https://doi.org/10.1016/j.ajhg.2022.10.013.

Pereira, J. *et al.* (2021) 'High-accuracy protein structure prediction in CASP14', *Proteins: Structure, Function, and Bioinformatics*, 89(12), pp. 1687–1699. Available at: https://doi.org/10.1002/prot.26171.

Pérez-Palma, E. *et al.* (2020) 'Identification of pathogenic variant enriched regions across genes and gene families', *Genome Research*, 30(1), pp. 62–71. Available at: https://doi.org/10.1101/gr.252601.119.

Petukh, M., Kucukkal, T.G. and Alexov, E. (2015) 'On human disease-causing amino acid variants: statistical study of sequence and structural patterns', *Human Mutation*, 36(5), pp. 524–534. Available at: https://doi.org/10.1002/humu.22770.

Plon, S.E. *et al.* (2008) 'Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results', *Human mutation*, 29(11), pp. 1282–1291. Available at: https://doi.org/10.1002/humu.20880.

Qi, H. *et al.* (2021) 'MVP predicts the pathogenicity of missense variants by deep learning', *Nature Communications*, 12(1), p. 510. Available at: https://doi.org/10.1038/s41467-020-20847-0.

Quinodoz, M. *et al.* (2022) 'Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity', *American Journal of Human Genetics*, 109(3), pp. 457–470. Available at: https://doi.org/10.1016/j.ajhg.2022.01.006.

Rabbani, B. *et al.* (2012) 'Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders', *Journal of Human Genetics*, 57(10), pp. 621–632. Available at: https://doi.org/10.1038/jhg.2012.91.

Ragsdale, D.S. (2008) 'How do mutant Nav1.1 sodium channels cause epilepsy?', *Brain Research Reviews*, 58(1), pp. 149–159. Available at: https://doi.org/10.1016/j.brainresrev.2008.01.003.

Ramensky, V., Bork, P. and Sunyaev, S. (2002) 'Human non-synonymous SNPs: server and survey', *Nucleic Acids Research*, 30(17), pp. 3894–3900.

Rentzsch, P. *et al.* (2019) 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Research*, 47(D1), pp. D886–D894. Available at: https://doi.org/10.1093/nar/gky1016.

Richards, S. *et al.* (2015) 'Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology', *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 17(5), pp. 405–424. Available at: https://doi.org/10.1038/gim.2015.30.

Ripke, S. *et al.* (2014) 'Biological insights from 108 schizophrenia-associated genetic loci', *Nature*, 511(7510), pp. 421–427. Available at: https://doi.org/10.1038/nature13595.

Rivaud, M.R., Delmar, M. and Remme, C.A. (2020) 'Heritable arrhythmia syndromes associated with abnormal cardiac sodium channel function: ionic and non-ionic mechanisms', *Cardiovascular Research*, 116(9), pp. 1557–1570. Available at: https://doi.org/10.1093/cvr/cvaa082.

Royer-Bertrand, B. and Rivolta, C. (2015) 'Whole genome sequencing as a means to assess pathogenic mutations in medical genetics and cancer', *Cellular and Molecular Life Sciences*, 72(8), pp. 1463–1471. Available at: https://doi.org/10.1007/s00018-014-1807-9.

Rudin, C. (2019) 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', *Nature machine intelligence*, 1(5), pp. 206–215. Available at: https://doi.org/10.1038/s42256-019-0048-x.

Salles, P.A. *et al.* (2021) 'ATP1A3-Related Disorders: An Ever-Expanding Clinical Spectrum', *Frontiers in Neurology*, 12, p. 637890. Available at: https://doi.org/10.3389/fneur.2021.637890.

Samocha, K.E. *et al.* (2014) 'A framework for the interpretation of de novo mutation in human disease', *Nature Genetics*, 46(9), pp. 944–950. Available at: https://doi.org/10.1038/ng.3050.

Samocha, K.E. *et al.* (2017) 'Regional missense constraint improves variant deleteriousness prediction', *bioRxiv*, p. 148353. Available at: https://doi.org/10.1101/148353.

Sharo, A.G. *et al.* (2022) 'ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden'. medRxiv, p. 2022.10.26.22281567. Available at: https://doi.org/10.1101/2022.10.26.22281567.

Sivley, R.M. *et al.* (2018) 'Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures', *American Journal of Human Genetics*, 102(3), pp. 415–426. Available at: https://doi.org/10.1016/j.ajhg.2018.01.017.

Śledź, P. and Caflisch, A. (2018) 'Protein structure-based drug design: from docking to molecular dynamics', *Current Opinion in Structural Biology*, 48, pp. 93–102. Available at: https://doi.org/10.1016/j.sbi.2017.10.010.

Stefl, S. *et al.* (2013) 'Molecular mechanisms of disease-causing missense mutations', *Journal of Molecular Biology*, 425(21), pp. 3919–3936. Available at: https://doi.org/10.1016/j.jmb.2013.07.014.

Stenson, P.D. *et al.* (2003) 'Human Gene Mutation Database (HGMD): 2003 update', *Human Mutation*, 21(6), pp. 577–581. Available at: https://doi.org/10.1002/humu.10212.

Strehlow, V. *et al.* (2019) 'GRIN2A-related disorders: genotype and functional consequence predict phenotype', *Brain: A Journal of Neurology*, 142(1), pp. 80–92. Available at: https://doi.org/10.1093/brain/awy304.

Striano, P. and Minassian, B.A. (2020) 'From Genetic Testing to Precision Medicine in Epilepsy', *Neurotherapeutics*, 17(2), pp. 609–615. Available at: https://doi.org/10.1007/s13311-020-00835-4.

Sudlow, C. *et al.* (2015) 'UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age', *PLoS medicine*, 12(3), p. e1001779. Available at: https://doi.org/10.1371/journal.pmed.1001779.

Taliun, D. *et al.* (2021) 'Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program', *Nature*, 590(7845), pp. 290–299. Available at: https://doi.org/10.1038/s41586-021-03205-y.

Tandy-Connor, S. *et al.* (2018) 'False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care', *Genetics in Medicine*, 20(12), pp. 1515–1521. Available at: https://doi.org/10.1038/gim.2018.38.

Teixeira, J.C. and Huber, C.D. (2021) 'The inflated significance of neutral genetic diversity in conservation genetics', *Proceedings of the National Academy of Sciences of the United States of America*, 118(10), p. e2015096118. Available at: https://doi.org/10.1073/pnas.2015096118.

Tennessen, J.A. *et al.* (2012) 'Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes', *Science*, 337(6090), pp. 64–69. Available at: https://doi.org/10.1126/science.1219240.

Thapar, A., Cooper, M. and Rutter, M. (2017) 'Neurodevelopmental disorders', *The Lancet. Psychiatry*, 4(4), pp. 339–346. Available at: https://doi.org/10.1016/S2215-0366(16)30376-5.

Thompson, C.H. *et al.* (2023) 'Epilepsy-associated SCN2A (NaV1.2) Variants Exhibit Diverse and Complex Functional Properties'. bioRxiv, p. 2023.02.23.529757. Available at: https://doi.org/10.1101/2023.02.23.529757.

Tokheim, C. *et al.* (2016) 'Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure', *Cancer research*, 76(13), pp. 3719–3731. Available at: https://doi.org/10.1158/0008-5472.CAN-15-3190.

del-Toro, N. *et al.* (2019) 'Capturing variation impact on molecular interactions in the IMEx Consortium mutations data set', *Nature Communications*, 10(1), p. 10. Available at: https://doi.org/10.1038/s41467-018-07709-6.

Traynelis, J. *et al.* (2017) 'Optimizing genomic medicine in epilepsy through a gene-customized approach to missense variant interpretation', *Genome Research*, 27(10), pp. 1715–1729. Available at: https://doi.org/10.1101/gr.226589.117.

Tsai, G.J. *et al.* (2019) 'Outcomes of 92 patient-driven family studies for reclassification of variants of uncertain significance', *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 21(6), pp. 1435–1442. Available at: https://doi.org/10.1038/s41436-018-0335-7.

Tunyasuvunakool, K. *et al.* (2021) 'Highly accurate protein structure prediction for the human proteome', *Nature*, 596(7873), pp. 590–596. Available at: https://doi.org/10.1038/s41586-021-03828-1.

Varadi, M. *et al.* (2022) 'AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models', *Nucleic Acids Research*, 50(D1), pp. D439–D444. Available at: https://doi.org/10.1093/nar/gkab1061.

Walsh, R. *et al.* (2014) 'Paralogue annotation identifies novel pathogenic variants in patients with Brugada syndrome and catecholaminergic polymorphic ventricular tachycardia', *Journal of Medical Genetics*, 51(1), pp. 35–44. Available at: https://doi.org/10.1136/jmedgenet-2013-101917.

Wang, K., Li, M. and Hakonarson, H. (2010) 'ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data', *Nucleic Acids Research*, 38(16), p. e164. Available at: https://doi.org/10.1093/nar/gkq603.

Ware, J.S. *et al.* (2012) 'Paralogous annotation of disease-causing variants in long QT syndrome genes', *Human Mutation*, 33(8), pp. 1188–1191. Available at: https://doi.org/10.1002/humu.22114.

Wells, G. *et al.* (2018) 'The GluN2B-Glu413Gly NMDA receptor variant arising from a de novo GRIN2B mutation promotes ligand-unbinding and domain opening', *Proteins*, 86(12), pp. 1265–1276. Available at: https://doi.org/10.1002/prot.25595.

Wilkie, A.O. (1994) 'The molecular basis of genetic dominance', *Journal of Medical Genetics*, 31(2), pp. 89–98. Available at: https://doi.org/10.1136/jmg.31.2.89.

Wolff, M. *et al.* (2017) 'Genetic and phenotypic heterogeneity suggest therapeutic implications in SCN2A-related disorders', *Brain: A Journal of Neurology*, 140(5), pp. 1316–1336. Available at: https://doi.org/10.1093/brain/awx054.

Won, J. *et al.* (2019) 'Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning', *Proteins*, 87(12), pp. 1351–1360. Available at: https://doi.org/10.1002/prot.25804.

Xenakis, M.N. *et al.* (2021) 'Hydropathicity-based prediction of pain-causing NaV1.7 variants', *BMC Bioinformatics*, 22. Available at: https://doi.org/10.1186/s12859-021-04119-2.

Yates, B. *et al.* (2021) 'Updates to HCOP: the HGNC comparison of orthology predictions tool', *Briefings in Bioinformatics* [Preprint], (bbab155). Available at: https://doi.org/10.1093/bib/bbab155.

Zarate, Y.A. *et al.* (2023) 'Quantitative Phenotype Morbidity Description of *SATB2*-Associated Syndrome', *Human Mutation*, 2023, p. e8200176. Available at: https://doi.org/10.1155/2023/8200176.

Zuberi, S.M. *et al.* (2011) 'Genotype-phenotype associations in SCN1A-related epilepsies', *Neurology*, 76(7), pp. 594–600. Available at: https://doi.org/10.1212/WNL.0b013e31820c309b.

9. Appendix

9.1 Statutory declaration and statement

„Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht."

Teilpublikationen sind unter „List of publications" aufgeführt.

01.05.2023, Tobias Brünger T. Brü

Datum, Name, Unterschrift

9.2 List of publications

First and shared first authorships (as indicated by ⁺):

Tobias Brünger⁺, Alina Ivaniuk⁺, Eduardo Pérez-Palma, Katrine M Johannesen, Ingo Helbig, Sameer Zuberi, Michael Nothnagel, Rikke S Møller, Patrick May, Andreas Brunklaus, Dennis Lal. Enhancing clinical genetic variant interpretation by extension of the PS1/PM5 ACMG criteria. 2023. (in preparation for submission to **Genetics in Medicine**)

Andreas Brunklaus⁺, Tony Feng⁺, **Tobias Brünger⁺**, Eduardo Perez-Palma, Henrike Heyne, Emma Matthews, Christopher Semsarian, Joseph D Symonds, Sameer M Zuberi, Dennis Lal, Stephanie Schorge. Gene variant effects across sodium channelopathies predict function and guide precision therapy. 2022. **Brain**. DOI:10.1093/brain/awac006. (published)

Tobias Brünger, Eduardo Pérez-Palma, Ludovica Montanucci, Michael Nothnagel, Rikke S Møller, Stephanie Schorge, Sameer Zuberi, Joseph Symonds, Johannes R Lemke, Andreas Brunklaus, Stephen F Traynelis, Patrick May, Dennis Lal. Conserved patterns across ion channels correlate with variant pathogenicity and clinical phenotypes. 2022. **Brain**. DOI: 10.1093/brain/awac305. (published)

Sumaiya Iqbal⁺, **Tobias Brünger⁺**, Eduardo Pérez-Palma, Marie Macnee, Andreas Brunklaus, Mark J Daly, Arthur J Campbell, David Hoksza, Patrick May, Dennis Lal. Delineation of functionally essential protein regions for 242 neurodevelopmental genes. 2023. **Brain**. 10.1093/brain/awac381. (published)

Ludovica Montanucci⁺, **Tobias Brünger⁺**, Yuchen Xu, Nisha Bhattarai, Rui Song, Wenshu XiangWei, Sukhan Kim, James P. Allen, Daniel Teuscher, Lingling Xie, Ruth K. Mizu, Wenjuan Chen, Chiara Klöckner, Piero Fariselli, Patrick May, Johannes R. Lemke, Scott J. Myers, Hongjie Yuan, Stephen F. Traynelis, Dennis Lal. Predicting variant pathogenicity and functional effects in NMDA receptor encoding GRIN genes. 2023. **American Journal of Human Genetics**. (submitted)

Co-authorships:

Andreas Brunklaus, **Tobias Brünger**, Tony Feng, Carmen Fons, Anni Lehikoinen, Eleni Panagiotakaki, Mihaela-Adela Vintan, Joseph Symonds, James Andrew, Alexis Arzimanoglou, Sarah Delima, Julie Gallois, Donncha Hanrahan, Gaetan Lesca, Stewart MacLeod, Dragan Marjanovic, Amy McTague, Noemi Nuñez-Enamorado, Eduardo Perez-Palma, M Scott Perry, Karen Pysden, Sophie J Russ-Hall, Ingrid E Scheffer, Krystal Sully, Steffen Syrbe, Ulvi Vaher, Murugan Velayutham, Julie Vogt, Shelly Weiss, Elaine Wirrell, Sameer M Zuberi, Dennis Lal, Rikke S Møller, Massimo Mantegazza, Sandrine Cestèle. The gain of function SCN1A disorder spectrum: novel epilepsy phenotypes and therapeutic implications. 2022. **Brain**. DOI: 10.1093/brain/awac210. (published)

Marie Macnee, Eduardo Pérez-Palma, **Tobias Brünger**, Chiara Klöckner, Konrad Platzer, Arthur Stefanski, Ludovica Montanucci, Allan Bayat, Maximilian Radtke, Ryan L Collins, Michael Talkowski, Daniel Blankenberg, Rikke S Møller, Johannes R Lemke, Michael Nothnagel, Patrick May, Dennis Lal. CNV-ClinViewer: Enhancing the clinical interpretation of large copy-number variants online. 2023. **Bioinformatics**. DOI: 10.1093/bioinformatics/btad290. (published)

Arthur Stefanski, Eduardo Pérez-Palma, **Tobias Brünger**, Ludovica Montanucci, Cornelius Gati, Chiara Klöckner, Katrine M Johannesen, Kimberly Goodspeed, Marie Macnee, Alexander T Deng, Ángel Aledo-Serrano, Artem Borovikov, Maina Kava, Arjan M Bouman, MJ Hajianpour, Deb K Pal, Marc Engelen, Eveline E O Hagebeuk, Marwan Shinawi, Alexis R Heidlebaugh, Kathryn Oetjens, Trevor L Hoffman, Pasquale Striano, Amanda S Freed, Line Futtrup, Thomas Balslev, Anna Abulí, Leslie Danvoye, Damien Lederer, Tugce Balci, Maryam Nabavi Nouri, Elizabeth Butler, Sarah Drewes, Kalene van Engelen, Katherine B Howell, Jean Khoury, Patrick May, Marena Trinidad, Steven Froelich, Johannes R Lemke, Jacob Tiller, Amber N Freed, Jing-Qiong Kang, and Arthur Wuster, Rikke S Møller, Dennis Lal. SLC6A1 variant pathogenicity, molecular function, and phenotype: a genetic and clinical analysis. 2023. **Brain**. (in review)

Ludovica Montanucci[+], David Lewis-Smith[+], Ryan L. Collins[+], Lisa-Marie Niestroj[+], Shridhar Parthasarathy, Julie Xian, Shiva Ganesan, Marie Macnee, **Tobias Brünger**, Rhys H. Thomas, Michael Talkowski, Epi25 Collaborative, Ingo Helbig, Costin Leu*, and Dennis Lal. Genome-wide identification and phenotypic characterization of seizure-associated copy number variants in 741,075 individuals. 2023. **Nature Communications.** (accepted)

Yuri A. Zarate, Katherine Bosanko, Amrit Kannan, Ashlen Thomason, Beth Nutt, Nihit Kumar, Kirt Simmons, Aaron Hiegert, Larry Hartzell, Adam Johnson, Tabitha Prater, Eduardo Pérez-Palma, **Tobias Brünger**, Arthur Stefanski, Dennis Lal, and Aisling R. Caffrey. Quantitative Phenotype Morbidity Description of SATB2-Associated Syndrome. 2023. **Human Mutation**. DOI: 10.1155/2023/8200176. (published)

Juan Lorenzo B. Pablo[+], Savannah L. Cornett[+], Lei A. Wang, Sooyeon Jo, **Tobias Brünger**, Nikita Budnik, Mudra Hegde, John Doench, Dennis Lal, Al George, Jen Pan. Base editing scanning of SCN2A reveals both gain and loss of function mutations in voltage-gated sodium channel NaV1.2. 2023. **Cell Reports**. (in revision)

Philippe A Salles, Ignacio F Mata, **Tobias Brünger**, Dennis Lal, Hubert H Fernandez. ATP1A3-Related Disorders: An Ever-Expanding Clinical Spectrum. 2021. **Frontiers in Neurology**. DOI: 10.3389/fneur.2021.637890. (published)

Niklas Schwarz, Simone Seiffert, Manuela Pendziwiat, Annika Verena Rademacher, **Tobias Brünger**, Ulrike B S Hedrich, Paul B Augustijn, Hartmut Baier, Allan Bayat, Francesca Bisulli, Russell J Buono, Ben Zeev Bruria, Michael G Doyle, Renzo Guerrini, Gali Heimer, Michele Iacomino, Hugh Kearney, Karl Martin Klein, Ioanna Kousiappa, Wolfram S Kunz, Holger Lerche, Laura Licchetta, Ebba Lohmann, Raffaella Minardi, Marie McDonald, Sarah Montgomery, Lejla Mulahasanovic, Renske Oegema, Barel Ortal, Savvas S Papacostas, Francesca Ragona, Tiziana Granata, Phillip S Reif, Felix Rosenow, Annick Rothschild, Paolo Scudieri, Pasquale Striano, Paolo Tinuper, George A Tanteles, Annalisa Vetro, Felix Zahnert, Ethan M Goldberg, Federico Zara, Dennis Lal, Patrick May, Hiltrud Muhle, Ingo Helbig, Yvonne Weber. Spectrum of Phenotypic, Genetic, and Functional Characteristics in Patients With Epilepsy With KCNC2 Pathogenic Variants. 2022. **Neurology**. DOI: 10.1212/WNL.0000000000200660. (published)

Allan Bayat, Alexej Knaus, Manuela Pendziwiat, Alexandra Afenjar, Tahsin Stefan Barakat, Friedrich Bosch, Bert Callewaert, Patrick Calvas, Berten Ceulemans, Nicolas Chassaing, Christel Depienne, Milda Endziniene, Carlos R Ferreira, Carolina Fischinger Moura de Souza, Cécile Freihuber, Shiva Ganesan, Svetlana Gataullina, Renzo Guerrini, Anne-Marie Guerrot, Lars Hansen, Aleksandra Jezela-Stanek, Caroline Karsenty, Anneke Kievit, Frank R Kooy, Christian M Korff, Johanne Kragh Hansen, Martin Larsen, Valérie Layet, Gaetan Lesca, Kim L

McBride, Marije Meuwissen, Cyril Mignot, Martino Montomoli, Hannah Moore, Sophie Naudion, Caroline Nava, Marie-Christine Nougues, Elena Parrini, Matthew Pastore, Jurgen H Schelhaas, Steven Skinner, Krzysztoł Szczałuba, Ashley Thomas, Mads Thomassen, Lisbeth Tranebjaerg, Marjon van Slegtenhorst, Lynne A Wolfe, Dennis Lal, Elena Gardella, Lilian Bomme Ousager, **Tobias Brünger**, Ingo Helbig, Peter Krawitz, Rikke S Møller. Lessons learned from 40 novel PIGA patients and a review of the literature. 2020. **Epilepsia**. DOI: 10.1111/epi.16545. (published)

Katrine M Johannesen, Yuanyuan Liu, Mahmoud Koko, Cathrine E Gjerulfsen, Lukas Sonnenberg, Julian Schubert, Christina D Fenger, Ahmed Eltokhi, Maert Rannap, Nils A Koch, Stephan Lauxmann, Johanna Krüger, Josua Kegele, Laura Canafoglia, Silvana Franceschetti, Thomas Mayer, Johannes Rebstock, Pia Zacher, Susanne Ruf, Michael Alber, Katalin Sterbova, Petra Lassuthová, Marketa Vlckova, Johannes R Lemke, Konrad Platzer, Ilona Krey, Constanze Heine, Dagmar Wieczorek, Judith Kroell-Seger, Caroline Lund, Karl Martin Klein, P Y Billie Au, Jong M Rho, Alice W Ho, Silvia Masnada, Pierangelo Veggiotti, Lucio Giordano, Patrizia Accorsi, Christina E Hoei-Hansen, Pasquale Striano, Federico Zara, Helene Verhelst, Judith S Verhoeven, Hilde M H Braakman, Bert van der Zwaag, Aster V E Harder, Eva Brilstra, Manuela Pendziwiat, Sebastian Lebon, Maria Vaccarezza, Ngoc Minh Le, Jakob Christensen, Sabine Grønborg, Stephen W Scherer, Jennifer Howe, Walid Fazeli, Katherine B Howell, Richard Leventer, Chloe Stutterd, Sonja Walsh, Marion Gerard, Bénédicte Gerard, Sara Matricardi, Claudia M Bonardi, Stefano Sartori, Andrea Berger, Dorota Hoffman-Zacharska, Massimo Mastrangelo, Francesca Darra, Arve Vøllo, M Mahdi Motazacker, Phillis Lakeman, Mathilde Nizon, Cornelia Betzler, Cecilia Altuzarra, Roseline Caume, Agathe Roubertie, Philippe Gélisse, Carla Marini, Renzo Guerrini, Frederic Bilan, Daniel Tibussek, Margarete Koch-Hogrebe, M Scott Perry, Shoji Ichikawa, Elena Dadali, Artem Sharkov, Irina Mishina, Mikhail Abramov, Ilya Kanivets, Sergey Korostelev, Sergey Kutsev, Karen E Wain, Nancy Eisenhauer, Monisa Wagner, Juliann M Savatt, Karen Müller-Schlüter, Haim Bassan, Artem Borovikov, Marie Cecile Nassogne, Anne Destrée, An Sofie Schoonjans, Marije Meuwissen, Marga Buzatu, Anna Jansen, Emmanuel Scalais, Siddharth Srivastava, Wen Hann Tan, Heather E Olson, Tobias Loddenkemper, Annapurna Poduri, Katherine L Helbig, Ingo Helbig, Mark P Fitzgerald, Ethan M Goldberg, Timo Roser, Ingo Borggraefe, **Tobias Brünger**, Patrick May, Dennis Lal, Damien Lederer, Guido Rubboli, Henrike O Heyne, Gaetan Lesca, Ulrike B S Hedrich, Jan Benda, Elena Gardella, Holger Lerche, Rikke S Møller. Genotype-phenotype correlations in SCN8A-related disorders reveal prognostic and therapeutic implications. 2022. **Brain**. DOI: 10.1093/brain/awab321. (published)

Katrine M Johannesen, Elena Gardella, Alejandra C Encinas, Anna-Elina Lehesjoki, Tarja Linnankivi, Michael B Petersen, Ida Charlotte Bay Lund, Susanne Blichfeldt, Maria J Miranda, Deb K Pal, Karine Lascelles, Peter Procopis, Alessandro Orsini, Alice Bonuccelli, Thea Giacomini, Ingo Helbig, Christina D Fenger, Sanjay M Sisodiya, Laura Hernandez-Hernandez, Sundararaman Krithika, Melissa Rumple, Silvia Masnada, Marialuisa Valente, Cristina Cereda, Lucio Giordano, Patrizia Accorsi, Sarah E Bürki, Margherita Mancardi, Christian Korff, Renzo Guerrini, Sarah von Spiczak, Dorota Hoffman-Zacharska, Tomasz Mazurczak, Antonietta Coppola, Salvatore Buono, Marilena Vecchi, Michael F Hammer, Costanza Varesio, Pierangelo Veggiotti, Dennis Lal, **Tobias Brünger**, Federico Zara, Pasquale Striano, Guido Rubboli, Rikke S Møller. The spectrum of intermediate SCN8A-related epilepsy. 2019. **Epilepsia**. DOI: 10.1111/epi.14705. (published, based on work before entering the Ph.D. program)