PLOS Computational Biology

SOFTWARE

VUStruct: A compute pipeline for high throughput and personalized structural biology

Christopher W. Moth[1], Jonathan H. Sheehan[2], Abdullah Al Mamun[3], R. Michael Sivley[4], Alican Gulsevin[5], David C. Rinker[6], Zenab F. Mchaourab[7], Undiagnosed Diseases Network[¶], John A. Capra[8], Jens Meiler[1,9]*

1 Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology and Institute of Chemical Biology; Vanderbilt University, Nashville, Tennessee, United States of America, **2** Division of Infectious Diseases, Department of Internal Medicine, Washington University School of Medicine, St. Louis, Missouri, United States of America, **3** Department of Biomedical Data Science, School of Applied Computational Sciences, Meharry Medical College, Nashville, Tennessee, United States of America, **4** Biomedical Informatics at 5Prime Sciences, Montreal, Quebec, Canada, **5** Department of Pharmaceutical Sciences, College of Pharmacy and Health Sciences, Butler University, Indianapolis, Indiana, United States of America, **6** Department of Biological Sciences, Evolutionary Studies Initiative; Vanderbilt University, Nashville, Tennessee, United States of America, **7** Meharry Medical College, Nashville, Tennessee, United States of America, **8** Bakar Computational Health Science Institute and Department of Epidemiology and Biostatistics, University of California, San Francisco, California, United States of America, **9** Leipzig University Medical School, Institute for Drug Discovery, Leipzig, Germany

* jens.meiler@vanderbilt.edu

¶ Membership of the Undiagnosed Diseases Network is listed in the supporting information (S5 File).

OPEN ACCESS

Citation: Moth CW, Sheehan JH, Al Mamun A, Sivley RM, Gulsevin A, Rinker DC, et al. (2026) VUStruct: A compute pipeline for high throughput and personalized structural biology. PLoS Comput Biol 22(5): e1014183. https://doi.org/10.1371/journal.pcbi.1014183

Editor: Jianhan Chen, University of Massachusetts Amherst, UNITED STATES OF AMERICA

Received: April 10, 2025

Accepted: March 31, 2026

Published: May 4, 2026

Data availability statement: For this software article and your journal requirements, all computational code is available without restriction at https://github.com/meilerlab/VUStruct, https://github.com/CapraLab/pdbmap, and https://github.com/CapraLab/

Abstract

Effective diagnosis and treatment of rare genetic disorders requires the interpretation of a patient's genetic variants of unknown significance (VUSs). Today, clinical decision-making is primarily guided by gene-phenotype association databases and DNA-based scoring methods. Our web-accessible variant analysis pipeline, VUStruct, supplements these established approaches by deeply analyzing the downstream molecular impact of variation in context of 3D protein structure. VUStruct's growing impact is fueled by the co-proliferation of protein 3D structural models, gene sequencing, compute power, and artificial intelligence. Contextualizing VUSs in protein 3D structural models also illuminates longitudinal genomics studies and biochemical bench research focused on VUS, and we created VUStruct for clinicians and researchers alike. We now introduce VUStruct to the broad scientific community as a mature, web-facing, extensible, High-Performance Computing (HPC) software pipeline. VUStruct maps missense variants onto automatically selected protein structures and launches a broad range of analyses. These include energy-based assessments of protein folding and stability, pathogenicity prediction through spatial clustering analysis, and machine learning (ML) predictors of binding surface disruptions and nearby post-translational modification sites. The pipeline also considers the entire input set of VUS and identifies genes potentially involved in digenic disease. VUStruct's utility in clinical rare disease genome interpretation has been demonstrated through its analysis of over 175 Undiagnosed Disease Network (UDN) Patient

pipeline_download_scripts. These codes include build scripts to make all needed Apptainer/singularity containers. The large container image files are also made available for download from https://meilerlab.org/VUStruct/containers/ (with the exception of the Rosetta ΔΔGfolding Cartesian and Monomer images, as these require a free academic or paid commercial license from rosettacommons.org).

Funding: This work was supported by the National Institutes of Health (U01HG010215 to JS; U01HG007674 to JS, CM, DR; R01LM013434 to JM, DR, CM, AC; U01NS134354 to JM, CM; U01NS134349 to JM, CM, ZM) JM acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1052 (209933838), and SPP 2363 (460865652). JM is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. JM is supported by BMBF (Federal Ministry of Education and Research) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI). This work is partly supported by the Federal Ministry of Education and Research (BMBF) through DAAD project 57616814 (SECAI, School of Embedded Composite AI). Work in the Meiler laboratory is further supported through the NIH (R01 HL122010, R01 DA046138, R01 AG068623, U01 AI150739, R01 CA227833, R01 LM013434, S10 OD016216, S10 OD020154, S10 OD032234). This work was supported by the BMBF-funded German Network for Bioinformatics Infrastructure (de.NBI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

cases. VUStruct-leveraged hypotheses have often informed clinicians in their consideration of additional patient testing, and we report here details from two cases where VUStruct was key to their solution. We also note successes with academic research collaborators, for whom VUStruct has informed research directions in both computational genomics and wet lab studies.

Author summary

For patients suffering from rare genetic disorders, whole genome DNA sequencing offers a promise of understanding that was previously unimaginable. However, advances in sequencing have magnified the need for improved approaches that can not only identify disease-causing genetic variants with confidence but also elucidate the molecular mechanisms by which these genetic changes lead to disease. For DNA variants that alter a protein's amino acid sequences, software developed by computational structural biologists has become increasingly valuable in this pursuit, due to ongoing explosions in computational power, 3D protein structures and models with atomistic detail, and sophisticated software integrating artificial intelligence-based technologies. We developed the VUStruct pipeline to automate a range of computational structural biology analyses relevant to interpreting genomes from undiagnosed rare genetic disorders into a unified system. Starting from a single list of genetic variants as input, VUStruct selects 3D structures and models, plans calculations, and then launches those calculations on a high-performance computing cluster. VUStruct updates a case website as results arrive. In this paper, we describe the mechanics of the VUStruct system and highlight two patient cases in which the disease hypotheses suggested by VUStruct contributed to resolving the cases. In Supplemental Materials online, we share additional insights we have gained from our use of VUStruct reports to support physicians in weekly patient case discussions.

Introduction

Clinical diagnosis of the genetic causes of rare diseases is primarily guided by databases of known gene-phenotype associations [1] and computational methods for quantifying the effects of genetic variants. Examples of these methods include GERP, which analyzes evolutionary constraint [2,3]; SIFT [4], which performs protein sequence homology analysis; and Polyphen [5], which is additionally trained on observed and predicted protein 3D structural features. While variant effect prediction algorithms have demonstrated utility in distinguishing known pathogenic variants from benign variants across large variant sets, these algorithms suffer from low specificity. Thus, computational methods are often of limited utility for the small sets of prefiltered variants [6] that are typically analyzed in clinical cases and other applications involving small sets of variants (bench studies of proteins and metabolic pathways,

deep mutational scans, etc.) [7]. The recent development of more sophisticated ML techniques and larger training data sets has increased the predictive accuracy of scoring algorithms [8,9]. Nonetheless, even AlphaMissense's scores lack reliability in cases of specific variants [8] and can exhibit high false positive rates [9]. I.e., the longitudinal statistical significance of these algorithms cannot diagnose an individual patient's disease, nor reliably identify disruption points in a single protein or metabolic pathway.

Moreover, computational variant effect prediction approaches reveal neither molecular nor biological mechanistic hypotheses. Instead, these tools are focused on the broad binary classification of mutations into pathogenic vs. benign. The critical biology of life unfolds in 3D space and time. Whereas scores compress this complex biology into a single number, supplementing scores with VUStruct analyses can point to hypotheses around the functional consequences of VUSs and their mechanisms of disease progression.

Mechanistically, variants in protein coding regions can disrupt protein function and cause diseases in various ways. As examples, amino acid substitutions can compromise the subtle energetics of protein folding and thermodynamic stability. Protein-protein interactions can be disrupted, post-translational modifications can be impeded, and metabolic networks can be broken [10].

Recently, computational protein structural analyses have demonstrated the power of mechanistic modeling of variants' effects to reveal causes of rare disease. For example, structural modeling suggested that a de-novo VUS in KCNC2 (V469L) could block the ion channel pore, impacting the stability of the protein [11]. This provided a rational and foundational hypothesis for the mechanism by which V469L causes developmental and epileptic encephalopathies (DEE) symptoms. Structure-based calculations also revealed that a missense variant in MSH2 could destabilize the protein, leading to cellular protein degradation and Lynch-syndrome disorder [12]. In these cases, structure-based calculations outperformed the traditionally used genetic disease predictors. The success of the MSH2 study, as well as numerous other single-gene focused analyses, informed the creation of a generalized structure-based workflow for variant classification in the clinic [13]. While this workflow provides guidance for 3D structural model selection and curation, the prescribed processes require significant human input. Once selected, structures must be manually forwarded to various external webservers which perform specific calculations, the results of which must still be integrated into final reports and explored with external visualization tools.

We created VUStruct considering these successes and analytical challenges. We hypothesized that an automated pipeline of structure-based calculations could reveal clues of variant structural and functional impacts which, in turn, could lead to the plausible identification of the root causes of rare genetic disorders in many patient cases.

The goal of VUStruct is not to predict pathogenic variants per se, but rather to provide robust context on the effects of a VUS on protein structure and function - context that enables the development of mechanistic clinical hypotheses about the causes of disease. VUStruct's automated contextualization of VUSs in protein 3D structural models can also illuminate longitudinal genomics studies and biochemical bench research focused on VUS. The pipeline automatically selects structures, integrates a broad spectrum of established computational approaches, and caps the calculation with holistic case-wide reporting. In contrast to other webservers which display variants on protein structures alongside precomputed and pre-aggregated scores [14,15], VUStruct performs fresh calculations based on queries to current genomic and protein model databases. Many servers require upload of a (single) protein structure file [16] or default to Alphafold-2 models [17] covering only Uniprot [18] canonical transcripts. VUStruct expands the scope of previous methods by integrating analyses of multiple 3D structures per protein and non-canonical transcripts when available. The final computed product is a website that enables drilling-down from a top-level case report, to each transcript, to 3D structure visualization. For each 3D structure, NGLviewer [19] sessions afford not only 3D manipulation of a variant's spatial environment, but also visualization of the proximity of known pathogenic and benign variants and evolutionary constraint within and between species (PathProx [20,21], ConSurf [22,23] and COSMIS [24]). VUStruct also investigates the potential for combinations of the VUSs to cause disease with DiGePred [25] and DIEP [26] ML algorithms trained to detect digenic disease.

Taken together, these automated and parallel calculations inform the clinic or laboratory at a 3D structural and molecular mechanistic level. VUStruct provides a compelling supplement to the insights gained from conventional genome-based scoring analysis alone, and we report the pipeline's contribution to two UDN [27,28] patient cases.

Design and implementation

The VUStruct computational pipeline is primarily implemented as Python codes which query and filter a wide range of pre-downloaded databases. Additional code launches and monitors the calculations which run inside Singularity [29] containers.

Conceptually, the pipeline runs in five discrete phases following upload of a variant set via the initial web form:

1) Optional pre-processing of human genomic coordinates
2) 3D structure selection and compute job planning
3) Launch of job arrays (for each variant) on the HPC
4) Progress monitoring
5) Report generation

Except for progress monitoring, these phases are depicted in Fig 1 and detailed below.

1 Variant upload and (optional) genomic preprocessing. VUStruct supports several input formats, which are first converted into a "vustruct.csv" pipeline-ready, comma-delimited flat file. This "VUStruct CSV" file contains gene names, transcript identifiers, amino acid variants, and parental inheritance when known. When users already know precise transcript identifiers and amino acid changes of interest, then "VUStruct CSV" format may be selected from the outset, and the pipeline will proceed directly to phase 2.

When starting from patient genetic variants (the underlying changes to alleles at chromosome positions), VUStruct converts these data to proteomic impacts. For variants loaded in VCF [30] format, parsed genomic coordinates are fed through the ENSEMBL [31] Variant Effect Predictor (VEP) [32] and missense variants are retained from the VEP output for subsequent structure selection and calculation scheduling. Since the VEP often reports impacts to many predicted transcripts which lack experimental validation, VUstruct restricts its calculations to the subset of returned genomic transcript IDs which cross-reference to Swiss-Prot curated Uniprot [18] protein IDs. Non-canonical transcripts are increasingly found *in-vivo* as proteomics methods evolve [33], and VUstruct includes all of a gene's curated splice variants. I.e., we incorporate the curated but non-canonical sequences identified in Uniprot with additional "-N" suffixed identifiers.

A challenge in our field is that annotations to the human reference genome [34] are not static. There are relatively frequent amino acid sequence discrepancies between ENSEMBL transcript records and Swiss-Prot curated Uniprot sequences (as cross-referenced by UniParc identifiers in Uniprot's "Id Mapping" resources). As a practical example, in early 2023, while 48,308 ENSEMBL transcripts cross-referenced to Uniprot sequences perfectly, we also found that 8,500 curated Uniprot IDs had no cross references to any GRCh38 Ensembl Transcript identifiers. 1,166 transcripts had cross-references and the transcript lengths were the same in both databases. However, the amino acid sequences were different. 370 transcripts had varying transcript lengths between ENSEMBL and Uniprot. Uniprot is constantly working to improve cross-references, and the MANE [35] collaboration is also informing the field. Today, for variants that cannot be immediately processed due to these disconnects, the pipeline reports these problems to the user and the pipeline stops. This provides users with the opportunity to either rework the genomic coordinates input data, or manually download and patch the preprocessor-generated vustruct.csv file for input to phase 2.

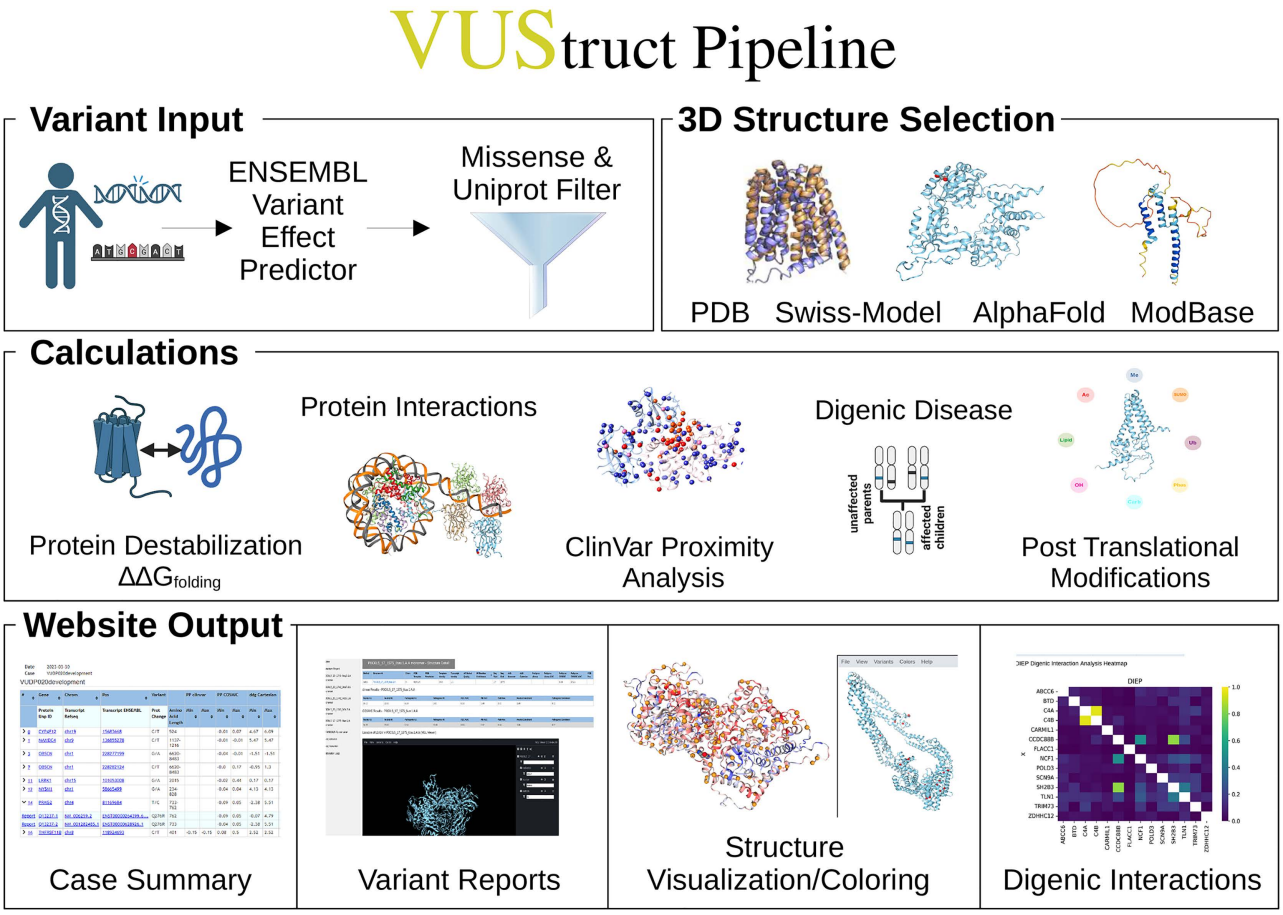

Fig 1. Starting from user-provided variant genomic coordinates (top left), VUStruct identifies missense variants and maps them onto protein structures which VUStruct automatically curates from experimental depositions and model databases. Various parallel calculations are then launched on the HPC, as enumerated in the text.

https://doi.org/10.1371/journal.pcbi.1014183.g001

2 Structure selection and compute job planning. From the Uniprot IDs in the "vustruct.csv" file, the pipeline "plans" the set of calculations by gathering structural information for target proteins. Available experimental structures are mined from the PDB and aligned to current transcripts via the SIFTS database [36,37]. SwissModel and ModBase models are integrated [38,39]. For mutations on canonical transcripts, AlphaFold [40] models are added to the set of representative structure. The final structure selections minimize redundancy and maximize diversity of experimental techniques, variant-coverage, model confidence and experimental quality metrics. Multimeric complexes are also prioritized in this process. See S2 Text for details.

Below the single directory for the user-provided Case ID, a subdirectory is created for each variant. For each retained structure, calculations are planned, and command line parameters are set for each job. These details are recorded in the workplan.csv of the variant subdirectory. Importantly, planning is entirely independent of the HPC architecture. To ensure that no job conflicts with any other, each user-input Case ID is appended to a Globally Unique Identifier (GUID) [41] and assigned a work directory in the hierarchy of VUStruct/CaseID and GUID/Transcript ID/3D Structure Type and ID/Calculation Type/Work Directory/. A sibling/Status Directory/ is used by each running job in VUStruct to uniformly communicate progress, competition, or failure to the VUStruct monitor application (described under phase 4).

3 Job launch. To launch the hundred(s) of jobs typically planned for a set of VUS (e.g., from a UDN case or from a list of variants from genome sequencing), the pipeline writes submission scripts to the SLURM [42] cluster environment on the back end. Each launched job runs out of a Singularity [29] container. From the container, the bound filesystem of the HPC environment is accessible but the application is otherwise blind to the surrounding HPC API. A single short script, external to the container architecture, launches all the HPC jobs, and records assigned job numbers for downstream monitoring.

The currently launched calculations include:

1) Rosetta $\Delta\Delta G_{folding}$ [43–45] estimates the energetic impact of each amino acid substitution on the free energy of protein folding. These two-part calculations are stored in a repository to avoid redundant "relax" steps and save compute time.
2) PathProx [20,21] predicts pathogenic variants when they better fit with clusters of "known pathogenic" sites (mined from ClinVar) [46] vs. randomly placed vs. benign variant sites found in Gnomad [47].
3) ScanNet [48] estimates the likelihood that a variant disrupts a protein-protein interaction, via an ML algorithm. We are also evaluating a newer method, PeSTo [49]
4) MusiteDeep [50] predicts protein post-translational (PTM) site modification through a deep-learning framework. We are also evaluating a newer method, PTMGP2 [51].
5) Digenic disease interactions are predicted with DigePred [25] and DIEP [26].

4 Job monitoring. Over the course of a VUStruct run, the case report is refreshed as jobs complete, to reflect the latest calculated data. The stdout, stderr, and.log files for each individual job are also updated.

The pipeline also informs the user of both overall and individual job progress on the cluster. In a large shared HPC environment, launched jobs are assigned unique job numbers, but do not immediately run. Traditionally, HPC users monitor job progress with a suite of HPC-provided command line tools. Through its web interface, VUStruct interfaces to these back-end tools, and dynamically reports on job prioritization, submission delays, remaining run time, and resource allocation. These technical status updates are presented to the user via a JavaScript monitor running in the case landing page. This page receives updates from an HPC node via middleware on the web server host. Expected run times are described in S3 Text.

5 Reporting. The pipeline generates a case-wide report as a landing page that combines calculated results for each transcript. As shown in Fig 2, the report also integrates queried scores from AlphaMissense [52], ConSurf [22,23] and COSMIS [24] for all the individual variants. This is followed by digenic analysis outputs.

From the case-wide report, the user may click into specific transcript reports (Fig 3). Clicking into a transcript report presents the user with a PFAM domain graphic [53], followed by a tabular summary of calculation results for the associated structures which were selected in step 2. The "navbar" at top left allows the user to hop to individual 3D structures, where NGL WebViewer [19] sessions are available to inspect the atomic environment of variants (Fig 4). The customized viewer also allows backbone coloring of the various calculated constraint scores, and model confidence. Outputs from newer methods, such as PeSTo [49] and PTMGPT2 [51], are only shown on the transcript reports during our evaluation period.

The downstream audience for VUStruct case reports is broader than the structural biologists trained to interpret the pipeline's detailed outputs. Typically, that final audience includes clinicians and geneticists who are primarily interested in whether VUStruct identifies a candidate gene for ongoing consideration, and how pipeline outputs, at high level, inform that recommendation. To communicate the high-level findings of VUStruct succinctly, VUStruct drafts a case summary spreadsheet (S1A Fig in S1 File). The S1 Text also suggests approaches to communicating with clinical partners and includes advice on calculation interpretation.

Date	2025-07-11 19:35 USA Central Time (GMT - 6)
Case	SAMPLE_CASE_01
Log Files	Click for log files
Supplemental Report Files	Click for supplemental report files

Maximum Cluster Compute End 7/15/2025 10:29:39 PM

This page will refresh every 15.0 minutes

3 case jobs are still active on the cluster

Case Preparation

VUStruct Phase	Application	Start Time	End Time	Input File	Log
plan	vustruct_plan.py	2025-07-11T10:09:16.382201	2025-07-11T10:10:09.626497	Input File	Log File
launch	vustruct_launch.py	2025-07-11T10:11:02.449669	2025-07-11T10:11:04.553337		Log File
monitor	vustruct_monitor.py	2025-07-11T19.34.52.748747	2025-07-11T19:34:55.023790		Log File

#	Gene	Chrom	Pos	Variant		PP clinvar		PP COSMIC		ddg Monomer		ddg Cartesian			Alpha	ScanNet	MusiteDeep	
	Protein Unp ID	Transcript Refseq	Transcript ENSEMBL	Prot Change	Amino Acid Length	Min	Max	Min	Max	Min	Max	Min	Max	COSMIS	Missense	PPI	PTM	Notes
0	TCF3	chr19	1632083	G/A	651-654			0.02	0.24			-0.34	-0.34		0.13 - 0.15	26%		
2	AP3D1	chr19	2115592	C/T	1153-1215			-0.07	0.02			-0.08	0.26		0.21 - 0.31	67%	67%	
4	AP3D1	chr19	2123841	C/T	1153-1215			0.02	0.04	0.59	0.6	0.17	0.59	-1.2	0.38 - 0.56	8%	85%	
6	GUCA1C	chr3	108908148	A/C	209			0.01	0.31			3.7	7.48	1.9	0.07	6%		
7	GUCA1C	chr3	108953665	C/T	209			0.18	0.69			1.55	4.22	0.8	0.63	41%		
8	BMP2K	chr4	78910693	C/T	1161			-0.01	-0.01			-11.2	-11.2		0.07	21%	85%	
9	CDH10	chr5	24537430	G/A	788			-0.17	0.22	1.12	3.02	-0.03	2.47	-0.9	0.07	7%	54% - 72%	
10	CR2	chr1	207474301	C/A	1033-1092			-0.01	0.05			6.85	6.85	0.6	0.75 - 0.82	56%		
12	DES	chr2	219420154	C/T	470	-0.24	-0.05	0.07	0.3	0.11	0.11	-0.06	0.07	-1.3	0.26	5%		
13	ESR1	chr6	151808139	C/G	595	-0.02	-0.02	-0.06	-0.06			-11.02	-11.02		0.06	64%		
14	HFE	chr6	26090939	G/A	242-348	-0.01	0.42	-0.01	0.43	5.66	17.84	1.35	10.39	-0.4	0.36 - 0.52	2%		
Report	Q30201-1	NM_000410.3	ENST00000357618.10	V59M	348	-0.01	0.39	-0.01	0.4	17.84	17.84	1.35	7.74	-0.4	0.41	2%		
Report	Q30201-10	NM_139003.2	ENST00000336625.12	V59M	242	0.23	0.42	0.16	0.43	8.5	8.5	3.34	7.74		0.46			
Report	Q30201-3	NM_139006.2	ENST00000461397.5	V59M	334	0.25	0.39	0.23	0.33	13.8	13.8	7.74	10.39		0.52			
Report	Q30201-5	NM_139009.2	ENST00000397022.7	V36M	325	0.2	0.41	0.32	0.4	11.16	11.16	7.74	9.38		0.44			
Report	Q30201-7	NM_139004.2	ENST00000317896.11	V59M	256	0.25	0.39	0.23	0.33	5.66	5.66	6.17	7.74		0.36			
19	LIMK1	chr7	74121197	C/T	613-647			-0.11	0.07	-0.32	-0.32	-1.39	1.08	1.2	0.09	29%		
21	MBD5	chr2	148468836	C/T	851-1494	-0.09	-0.09	-0.1	-0.0			-24.06	-24.06		0.59 - 0.90	40%	88%	
24	SERPINA6	chr14	94309814	T/C	405			-0.03	0.18			2.46	5.0	0.5	0.16	8%		
25	WDR41	chr5	77438279	T/C	404-459	-0.11	-0.11	-0.04	0.35	41.63	41.63	3.61	19.23	2.3	0.96 - 0.97	1%	72%	

DiGePred Analysis

Top Scoring DiGePred Gene Pairs

DiGePred Score	Gene A	Gene B
0.5	CR2	TCF3

Fig 2. Screenshot from the VUStruct-generated case report landing web page. In the table, each row summarizes the range of calculated values for each variant. Ranges arise in some calculations because multiple structures are considered for each transcript. In the case of input genomic coordinates, multiple transcript isoforms are often impacted for each variant. The “Refresh Cluster Jobs Information” box allows detailed monitoring and

troubleshooting. The drawn red box shows how summary row 14 (a change to gene HFE on Chrom 6) has been expanded to display five rows for different impacted transcripts. The first row corresponds to the canonical UniProt isoform. Clicking the "Report" link for that line will display detailed calculations for this variant in the context of that transcript (see next Fig).

https://doi.org/10.1371/journal.pcbi.1014183.g002

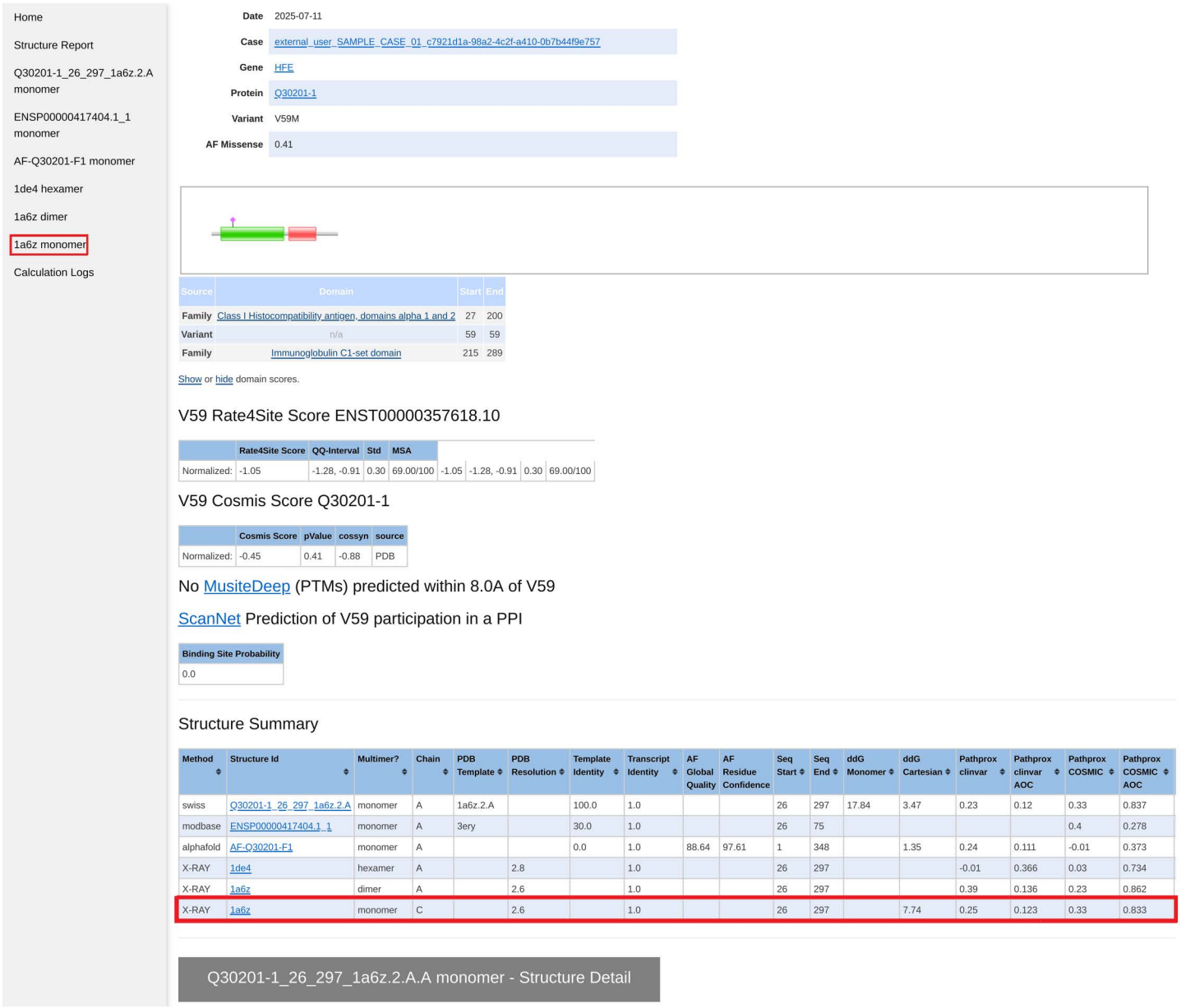

Fig 3. Screenshot of the VUStruct-generated transcript variant report, showing the variant location as a pink diamond in the context of the protein's PFAM domain [53] annotation. This is followed by results for Rate4Site, COSMIS, MusiteDeep, and ScanNet calculations. A key table is the Structure Summary, which lists all the structures (from the PDB, MODBASE, SWISS-MODEL database, and AlphaFold database) on which calculations were performed. For example, the highlighted row summarizes all the calculations performed on X-Ray crystal structure 1a6b.pdb, and the highlighted shortcut in the left column leads to the section of the page detailing those results (see next Fig).

https://doi.org/10.1371/journal.pcbi.1014183.g003

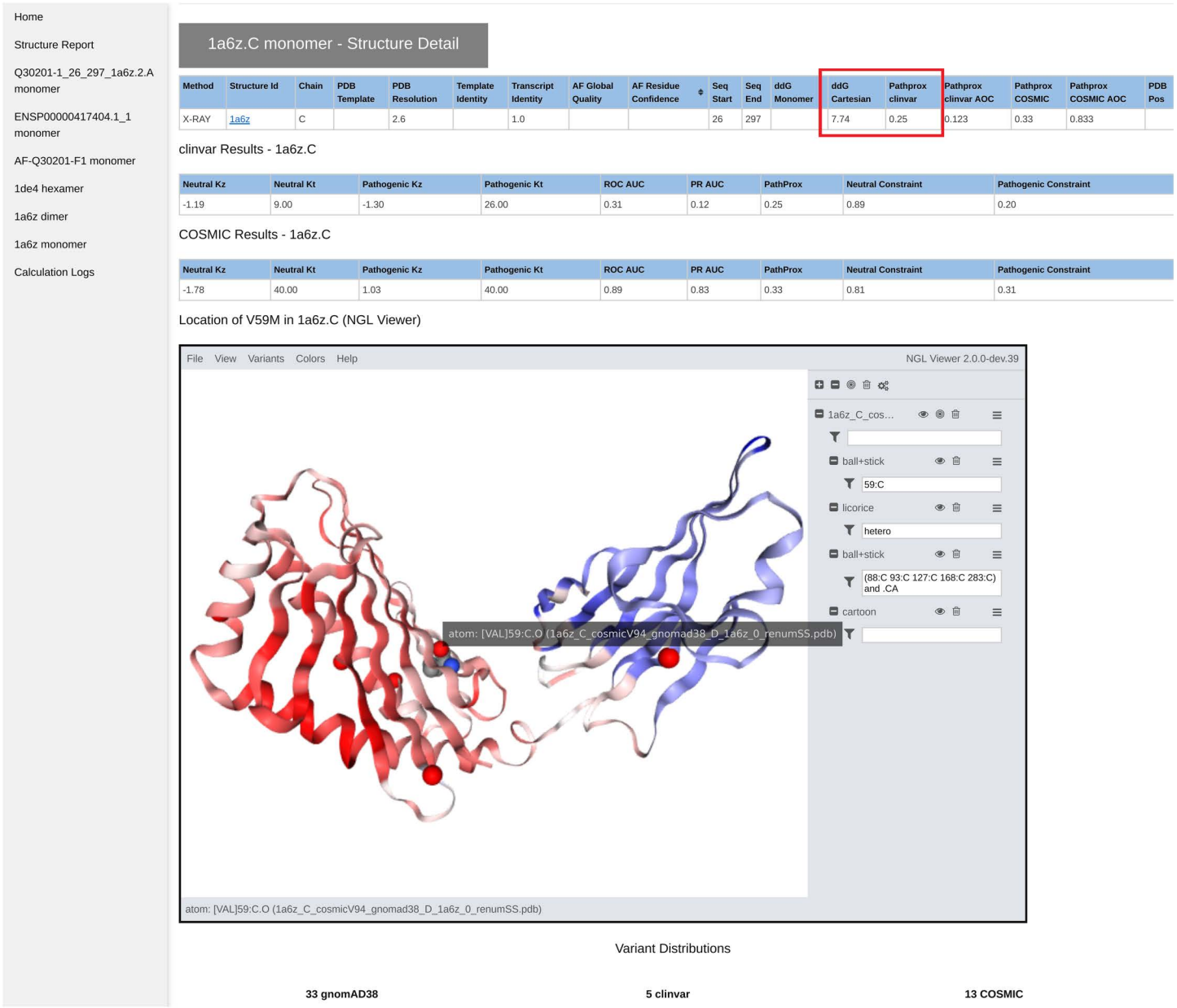

Fig 4. Screenshot of a section of the VUStruct results page, which for each structural model shows the results of the ΔΔG and PathProx calculations, highlighted in red, along with associated statistics to judge reliability. Each structure is displayed in a customized and interactive NGL Viewer [19] session. This view can be used to understand the structural context of the variant (highlighted here with a gray text pop-up. For example, one can display all pathogenic and likely pathogenic ClinVar variants (shown as red spheres), or color the model according to AlphaFold confidence or by PathProx score (from low- blue, to high - red - as shown above). Figures to illustrate a structural mechanistic hypothesis can be generated quickly from these images.

https://doi.org/10.1371/journal.pcbi.1014183.g004

Dependencies

VUStruct integrates several externally sourced databases. So that the pipeline can run responsively, and avoid vulnerability to external outages, the supporting databases are locally downloaded, installed, and maintained. The two support

pillars of VUStruct are the ENSEMBL GRCh38 PERL API and the UniProt ID mapping file. We locally import ENSEMBL's SQL database and additionally load UniProt [18] cross-references into SQL tables to speed sequence cross-references between genome and proteome. BASH scripts are additionally provided to aid download of ClinVar [46], COSMIC [54], and gnomAD [47] databases which are mined for PathProx's mathematical spatial analysis and for web-based visualizations. Several of our predictive calculations integrate sequence constraint, gleaned from both multi-species sequence alignments [22,23] and human population sequences [24]. These calculations, along with AlphaMissense [52] predictions, are downloaded as transcriptome-wide precomputations, and are integrated into final reports without the need for cluster launches.

Cited calculations are deployed inside Singularity Containers. Deployment of Rosetta $\Delta\Delta G_{folding}$ [43–45] Cartesian and Monomer calculations requires a free academic or paid commercial license from rosettacommons.org.

Results

We have demonstrated the VUStruct pipeline's utility in the interpretation of genetic VUS in collaboration with colleagues from the Vanderbilt UDN. The containerized VUStruct software pipeline has been applied to over 150 UDN Vanderbilt UDN patients and 25 Washington University patients. The pipeline provides researchers and clinician geneticists with insights into candidate missense variants in the context of 3D protein atomic structure. In contrast to the many algorithms and websites that perform a single calculation on a single protein variant on a single protein structure, VUStruct is holistic and automated. Our pipeline analyzes a set of patient genetic VUSs and unifies the results under a case-wide report page. VUStruct is also noteworthy for its principled selection of appropriate structures among the growing wealth of available experimental and computational structural models, automated calculation setup and launch, and progress monitoring.

As one illustration of VUStruct's potential to aid hypothesis generation, we highlight a patient with PASNA syndrome caused by a heterozygous variant in the CACNA1D gene that encodes a Human L-type voltage-gated calcium channel (Cav). Several candidate variants were selected from the patient genome sequencing (GS) data based on phenotype analysis. These variants were submitted to the pipeline and the 3D structure of the corresponding protein was analyzed by different computational methods including Rosetta ΔΔG [43–45], protein-protein interaction (PPI), post-translational modifications (PTM) and digenic predictions (DiGePred) analysis. VUStruct reported that the F767L variant in CACNA1D results in structural destabilization as evidenced by ΔΔG score in Rosetta. Starting from the VUStruct report, we hypothesized that the variant may contribute to the PASNA syndrome and conducted additional Rosetta simulations on the Cav structural model. In follow-up, two different variants F767L and F767S for Cav were used to calculate the ΔΔG in Rosetta using closed state conformation (PDB id: 7UHG [55]). F767S is a known pathogenic variant that causes a gain of function mechanism, and it was used as a positive control for this study. The higher calculated ΔΔG of F767L (~5.3 Rosetta Energy Units) vs. F767S (~3.9 R.E.U.) suggested that F767L could contribute to at least as much structural disruption as known pathogenic variant F767S for the closed state conformation. Thus we hypothesized that these variants destabilize the closed state, and push conformational equilibrium towards the channel opening state. The search for this crucial finding began with VUStruct analysis and led to the further confirmative analysis to diagnose the possible cause of the PASNA syndrome [56].

A second demonstration of VUStruct's utility was aiding a diagnosis of Diamond Blackfin anemia (DBA) in a case which could not be explained by simple Mendelian inheritance. The VUStruct report suggested that a missense variant in the RPS19 gene results in a slight stabilization, based on Rosetta ΔΔG. In addition, the proband carried another variant in the RPL27 gene, which DiGePred [25] and DIEP [26] analysis predicted to have a strong digenic interaction with RPS19. These clues helped to focus further structural analysis. We investigated different 80S ribosome structures available in the protein data bank. Although RPS19 and RPL27 are on opposite sides of the complex, it is plausible that T55M in RPS19 changes allosteric interactions between the two proteins, disrupting the 80S ribosome function. These structural analyses

inspired further co-segregation and RNA sequencing analysis of the proband. Further analysis of these suggested the proband's DBA is caused by the digenic interactions between RPS19 and RPL27 [57].

These two examples represent the sort of contribution that VUStruct analyses can make in favorable cases. Across a subset of Vanderbilt UDN cases annotated as "solved" with a single VUS, 74% of VUStruct runs yielded a prediction for structural destabilization, 21% yielded a statistically significant prediction for spatial clustering with pathogenic variants, and 13% yielded a prediction that the variant was involved in protein-protein interaction, among other calculation results. (See S4 Text)

Availability and future directions

The website is made available to all, without condition. For those wishing to setup their own pipeline environment, all our code and containers (with one exception), are licensed under the MIT License and can be downloaded from https://github.com/meilerlab/VUStruct. The one exception is the Rosetta ΔΔG module containers, which require Rosetta Commons licensing, available at no charge to academic users.

VUStruct development is continuously fueled by ongoing explosions in available protein 3D structures, genome sequencing, computer power, and artificial intelligence. We are committed to the pipeline's flexibility and continuous improvement.

One current limitation is VUStruct does not include strategies for generating multiple protein conformations. We are monitoring research in this area and plan to benchmark emerging methods soon [58]. We also are exploring enhancements to our structure selection algorithms to identify and retain, as example, *apo vs. holo depositions.* We hope to soon mine the application of AlphaFold to CHESS non-canonical transcripts [59]. Pending its public opening, we hope to mine the AlphaFold 3 [60] repository for its updated structural coverage that includes multimeric complexes [61,62], and build Boltz-2 [61,62] models as an integrated new stage of our pipeline. Predictions of digenic interactions should benefit from model retraining, given the emergence of new ground truth data sets [63].

Supporting information

S1 Text. VUStruct clinical case support and variant interpretation.
(PDF)

S2 Text. VUStruct structure selection algorithm.
(PDF)

S3 Text. VUStruct runtime expectations.
(PDF)

S4 Text. VUStruct towards quantitative assessment of clinical utility.
(PDF)

S1 File. Members of the undiagnosed diseases network.
(PDF)

S5 Text. S5 Members of the Undiagnosed Diseases Network.
(PDF)

Acknowledgement

This work leveraged the resources provided by the Vanderbilt Advanced Computing Center for Research and Education (ACCRE), a collaboratory operated by and for Vanderbilt faculty. ACCRE is comprised of over 3,000 researchers from more than 40 campus departments.

The pipeline would not have been possible without the energetic and helpful support from staff at Uniprot, ENSEMBL, SwissModel, and Modbase.

Author contributions

Conceptualization: Christopher W Moth, Jonathan H Sheehan, R Michael Sivley, John A Capra, Jens Meiler.

Data curation: Christopher W Moth, Jonathan H Sheehan, R Michael Sivley, Alican Gulsevin.

Formal analysis: Christopher W Moth, Jonathan H Sheehan, R Michael Sivley, Alican Gulsevin.

Funding acquisition: John A Capra, Jens Meiler.

Investigation: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, R Michael Sivley, Alican Gulsevin, David C Rinker, Zenab H Mchaourab.

Methodology: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, R Michael Sivley, Alican Gulsevin, Zenab H Mchaourab.

Project administration: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, Alican Gulsevin, David C Rinker, John A Capra, Jens Meiler.

Software: Christopher W Moth, R Michael Sivley, Alican Gulsevin.

Supervision: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, Alican Gulsevin, David C Rinker, John A Capra, Jens Meiler.

Validation: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, R Michael Sivley, Alican Gulsevin, David C Rinker, Zenab H Mchaourab.

Visualization: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, R Michael Sivley, Alican Gulsevin, David C Rinker, Zenab H Mchaourab.

Writing – original draft: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun.

Writing – review & editing: Christopher W Moth, Jonathan H Sheehan, Abdullah Al Mamun, Zenab H Mchaourab, John A Capra, Jens Meiler.

References

1. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47(D1):D1038–43. https://doi.org/10.1093/nar/gky1151 PMID: 30445645
2. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP. PLoS Computational Biology. 2010;6(12):e1001025. https://doi.org/10.1371/journal.pcbi.1001025
3. Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901–13. https://doi.org/10.1101/gr.3577405 PMID: 15965027
4. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4. https://doi.org/10.1093/nar/gkg509 PMID: 12824425
5. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002;30(17):3894–900. https://doi.org/10.1093/nar/gkf493 PMID: 12202775
6. Kobren SN, Baldridge D, Velinder M, Krier JB, LeBlanc K, Esteves C. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. Genetics in Medicine. 2021;23(6):1075–85. https://doi.org/10.1038/s41436-020-01084-8
7. Flanagan SE, Patch A-M, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomarkers. 2010;14(4):533–7. https://doi.org/10.1089/gtmb.2010.0036 PMID: 20642364
8. McDonald EF, Oliver KE, Schlebach JP, Meiler J, Plate L. Benchmarking AlphaMissense pathogenicity predictions against cystic fibrosis variants. PLoS One. 2024;19(1):e0297560. https://doi.org/10.1371/journal.pone.0297560 PMID: 38271453
9. Murali H, Wang P, Liao EC, Wang K. Genetic variant classification by predicted protein structure: a case study on IRF6. Comput Struct Biotechnol J. 2024;23:892–904. https://doi.org/10.1016/j.csbj.2024.01.019 PMID: 38370976

10. Taipale M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms 1. Biochem Cell Biol. 2019;97(1):46–57. https://doi.org/10.1139/bcb-2018-0007 PMID: 29693415
11. Mukherjee S, Cassini TA, Hu N, Yang T, Li B, Shen W, et al. Personalized structural biology reveals the molecular mechanisms underlying heterogeneous epileptic phenotypes caused by de novo KCNC2 variants. HGG Adv. 2022;3(4):100131. https://doi.org/10.1016/j.xhgg.2022.100131 PMID: 36035247
12. Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genet. 2017;13(4):e1006739. https://doi.org/10.1371/journal.pgen.1006739 PMID: 28422960
13. Caswell RC, Gunning AC, Owens MM, Ellard S, Wright CF. Assessing the clinical utility of protein structural analysis in genomic variant classification: experiences from a diagnostic laboratory. Genome Med. 2022;14(1):77. https://doi.org/10.1186/s13073-022-01082-2 PMID: 35869530
14. Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: disease variants and protein structure. Protein Sci. 2020;29(1):111–9. https://doi.org/10.1002/pro.3746 PMID: 31606900
15. Stephenson JD, Totoo P, Burke DF, Jänes J, Beltrao P, Martin MJ. ProtVar: mapping and contextualizing human missense variation. Nucleic Acids Res. 2024;52(W1):W140–7. https://doi.org/10.1093/nar/gkae413 PMID: 38769064
16. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can predicted protein 3D structures provide reliable insights into whether missense variants are disease associated?. J Mol Biol. 2019;431(11):2197–212. https://doi.org/10.1016/j.jmb.2019.04.009
17. Philipp M, Moth CW, Ristic N, Tiemann JKS, Seufert F, Panfilova A, et al. MutationExplorer: a webserver for mutation of proteins and 3D visualization of energetic impacts. Nucleic Acids Res. 2024;52(W1):W132–9. https://doi.org/10.1093/nar/gkae301 PMID: 38647044
18. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, et al. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res. 2023 Jan 6;51(D1):D523–31.https://doi.org/10.1093/nar/gkac1052
19. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlic A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018;34(21):3755–8. https://doi.org/10.1093/bioinformatics/bty419 PMID: 29850778
20. 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(3):415–26. https://doi.org/10.1016/j.ajhg.2018.01.017 PMID: 29455857
21. Sivley RM, Sheehan JH, Kropski JA, Cogan J, Blackwell TS, Phillips JA, et al. Three-dimensional spatial analysis of missense variants in RTEL1 identifies pathogenic variants in patients with Familial Interstitial Pneumonia. BMC Bioinformatics. 2018;19(1):18. https://doi.org/10.1186/s12859-018-2010-z PMID: 29361909
22. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Research. 2016;44(W1):W344-50. https://doi.org/10.1093/nar/gkw408
23. Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics. 2002;18 Suppl 1:S71-7. https://doi.org/10.1093/bioinformatics/18.suppl_1.s71 PMID: 12169533
24. Li B, Roden DM, Capra JA. The 3D mutational constraint on amino acid sites in the human proteome. Nat Commun. 2022;13(1):3273. https://doi.org/10.1038/s41467-022-30936-x PMID: 35672414
25. Mukherjee S, Cogan JD, Newman JH, Phillips JA 3rd, Hamid R, Undiagnosed Diseases Network, et al. Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. Am J Hum Genet. 2021;108(10):1946–63. https://doi.org/10.1016/j.ajhg.2021.08.010 PMID: 34529933
26. Yuan Y, Zhang L, Long Q, Jiang H, Li M. An accurate prediction model of digenic interaction for estimating pathogenic gene pairs of human diseases. Comput Struct Biotechnol J. 2022;20:3639–52. https://doi.org/10.1016/j.csbj.2022.07.011 PMID: 35891796
27. Gahl WA, Wise AL, Ashley EA. The undiagnosed diseases network of the national institutes of health. JAMA. 2015;314(17):1797. https://doi.org/10.1001/jama.2015.12249
28. Ramoni RB, Mulvihill JJ, Adams DR, Allard P, Ashley EA, Bernstein JA. The undiagnosed diseases network: accelerating discovery about health and disease. Am J Human Gene. 2017;100(2):185–92. https://doi.org/10.1016/j.ajhg.2017.01.006
29. Kurtzer GM, Sochat V, Bauer MW. Singularity: scientific containers for mobility of compute. PLoS One. 2017;12(5):e0177459. https://doi.org/10.1371/journal.pone.0177459 PMID: 28494014
30. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. https://doi.org/10.1093/bioinformatics/btr330
31. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988-95. https://doi.org/10.1093/nar/gkab1049
32. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A. The ensembl variant effect predictor. Genome Biology. 2016;17(1):122. https://doi.org/10.1186/s13059-016-0974-4
33. Sinitcyn P, Richards AL, Weatheritt RJ, Brademan DR, Marx H, Shishkova E, et al. Global detection of human variants and isoforms by deep proteome sequencing. Nat Biotechnol. 2023;41(12):1776–86. https://doi.org/10.1038/s41587-023-01714-x PMID: 36959352

34. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017;27(5):849–64. https://doi.org/10.1101/gr.213611.116 PMID: 28396521
35. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310–5. https://doi.org/10.1038/s41586-022-04558-8 PMID: 35388217
36. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res. 2019;47(D1):D482–9. https://doi.org/10.1093/nar/gky1114 PMID: 30445541
37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235–42. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235
38. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 2014;42(Database issue):D336-46. https://doi.org/10.1093/nar/gkt1144 PMID: 24271400
39. Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, et al. The SWISS-MODEL repository-new features and functionality. Nucleic Acids Res. 2017;45(D1):D313–9. https://doi.org/10.1093/nar/gkw1132 PMID: 27899672
40. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
41. Leach P, Mealling M, Salz R. A Universally Unique IDentifier (UUID) URN Namespace. 2005. https://doi.org/10.17487/rfc4122
42. Yoo AB, Jette MA, Grondona M. SLURM: simple linux utility for resource management. 2003. 44–60. https://doi.org/10.1007/10968987_3
43. Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. J Chem Theory Comput. 2016;12(12):6201–12. https://doi.org/10.1021/acs.jctc.6b00819 PMID: 27766851
44. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins. 2011;79(3):830–8. https://doi.org/10.1002/prot.22921 PMID: 21287615
45. Frenz B, Lewis SM, King I, DiMaio F, Park H, Song Y. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. Front Bioeng Biotechnol. 2020;8:558247. https://doi.org/10.3389/fbioe.2020.558247 PMID: 33134287
46. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7. https://doi.org/10.1093/nar/gkx1153 PMID: 29165669
47. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–43. https://doi.org/10.1038/s41586-020-2308-7 PMID: 32461654
48. Tubiana J, Schneidman-Duhovny D, Wolfson HJ. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods. 2022;19(6):730–9. https://doi.org/10.1038/s41592-022-01490-7 PMID: 35637310
49. Krapp LF, Abriata LA, Cortés Rodriguez F, Dal Peraro M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. Nat Commun. 2023;14(1):2175. https://doi.org/10.1038/s41467-023-37701-8 PMID: 37072397
50. Wang D, Zeng S, Xu C, Qiu W, Liang Y, Joshi T, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics. 2017;33(24):3909–16. https://doi.org/10.1093/bioinformatics/btx496 PMID: 29036382
51. Shrestha P, Kandel J, Tayara H, Chong KT. Post-translational modification prediction via prompt-based fine-tuning of a GPT-2 model. Nat Commun. 2024;15(1):6699. https://doi.org/10.1038/s41467-024-51071-9 PMID: 39107330
52. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381(6664). https://doi.org/10.1126/science.adg7492
53. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9. https://doi.org/10.1093/nar/gkaa913 PMID: 33125078
54. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. 2019;47(D1):D941-7. https://doi.org/10.1093/nar/gky1015
55. Yao X, Gao S, Yan N. Structural basis for pore blockade of human voltage-gated calcium channel Cav1.3 by motion sickness drug cinnarizine. Cell Res. 2022;32(10):946–8. https://doi.org/10.1038/s41422-022-00663-5 PMID: 35477996
56. Ezell KM, Tinker RJ, Furuta Y, Gulsevin A, Bastarache L, Hamid R. Undiagnosed disease network collaborative approach in diagnosing rare disease in a patient with a mosaic CACNA1D variant. Am J Med Genet A. 2024;194(7). https://doi.org/10.1002/ajmg.a.63597
57. Furuta Y, Tinker RJ, Gulsevin A, Neumann SM, Hamid R, Cogan JD. Probable digenic inheritance of Diamond–Blackfan anemia. Am J Med Genet A. 2024;194(3). https://doi.org/10.1002/ajmg.a.63454
58. Brown BP, Stein RA, Meiler J, Mchaourab HS. Approximating projections of conformational boltzmann distributions with alphafold2 predictions: opportunities and limitations. J Chem Theory Comput. 2024;20(3):1434–47. https://doi.org/10.1021/acs.jctc.3c01081 PMID: 38215214
59. Sommer MJ, Cha S, Varabyou A, Rincon N, Park S, Minkin I. Structure-guided isoform identification for the human transcriptome. eLife. 2022;11. https://doi.org/10.7554/eLife.82556
60. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024. https://doi.org/10.1038/s41586-024-07487-w

61. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler S, Somnath VR. Boltz-2: towards accurate and efficient binding affinity prediction. 2025. https://doi.org/10.1101/2025.06.14.659707

62. Wohlwend J, Corso G, Passaro S, Getz N, Reveiz M, Leidal K. Boltz-1 democratizing biomolecular interaction modeling. 2024. https://doi.org/10.1101/2024.11.19.624167

63. Nachtegael C, Gravel B, Dillen A, Smits G, Nowé A, Papadimitriou S, et al. Scaling up oligogenic diseases research with OLIDA: the oligogenic diseases database. Database (Oxford). 2022;2022:baac023. https://doi.org/10.1093/database/baac023 PMID: 35411390