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

Christopher W. Moth[1], Jonathan H. Sheehan[2], Abdullah Al Mamun[1], R. Michael Sivley[3], Alican Gulsevin[4], David Rinker[5], John A. Capra[6], Jens Meiler[17*]

[1] Departments of Chemistry, Pharmacology, and Biomedical Informatics; Center for Structural Biology and Institute of Chemical Biology; Vanderbilt Univ., Nashville, TN 37232, USA

[2] Division of Infection Diseases, Milliken Dept. of Internal Medicine, Washington Univ. of Medicine in St. Louis, MO 63110, USA

[3] Biomedical Informatics at 5Prime Sciences, Montreal, Canada

[4] Department of Pharmaceutical Sciences, College of Pharmacy and Health Sciences, Butler University, Indianapolis, IN 46208, USA.

[5] Department of Biological Sciences, Evolutionary Studies Initiative; Vanderbilt Univ., Nashville, TN 37232, USA

[6] Bakar Computational Health Science Institute and Department of Epidemiology and Biostatistics, Univ. of California San Francisco, CA 94143, USA

[7] Leipzig University Medical School, Institute for Drug Discovery, Brüderstraße 34, 04103 Leipzig, Germany

* Corresponding author

E-mail: jens@meilerlab.org

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 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.

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 pre-filtered 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., with such a high false positive rate, the touted 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.

Compounding the above caveats, computational variant effect prediction approaches reveal neither molecular nor biological mechanistic hypotheses. Instead, these tools are focused on the broad classification of mutations into pathogenic or benign, a vague partitioning with limited clinical utility. The critical biology of life unfolds in 3D space and time. Yet, scores compress this complex biology into a single number which obscures 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 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 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(27) 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 Figure 1 and detailed below.

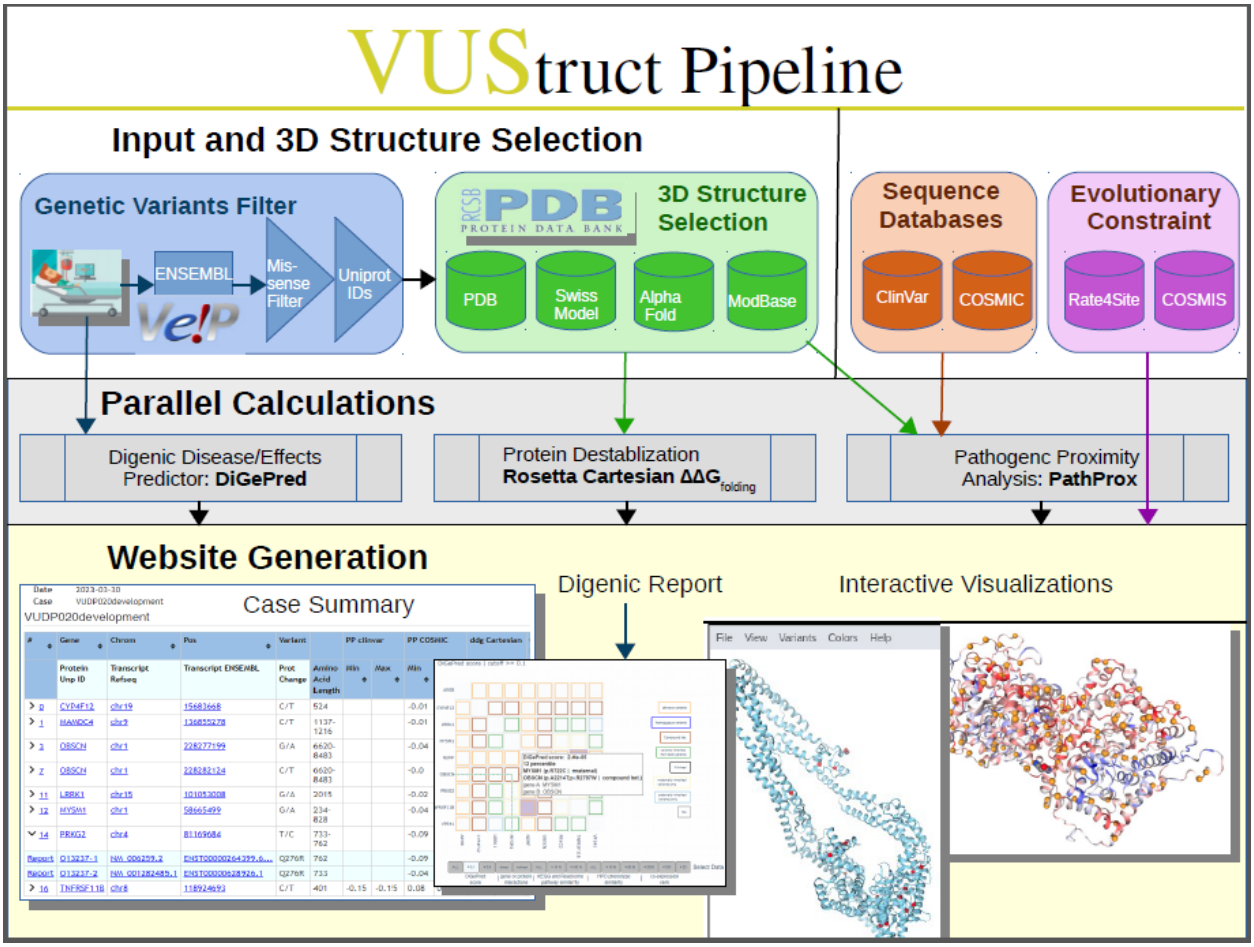


Figure 1. Starting from user-provided variant genomic coordinates (top left), VUStruct identifies missense variants and maps them onto protein structures which VUStruct automatically curates from experimental depositions and model databases. Various parallel calculations are then launched on the HPC,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

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(28) format, parsed genomic coordinates are fed through the ENSEMBL(29) Variant Effect Predictor (VEP)(30) 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(31), 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(32) 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 between ENSEMBL and Uniprot. Uniprot is constantly working to improve cross-references, and the MANE(33) 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.

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(34,35). SwissModel and ModBase models are integrated (36,37). For mutations on canonical transcripts, AlphaFold(38) 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.

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)(39) 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 list of variants from genome sequencing), the pipeline writes submission scripts for the supported cluster environments on the back end (either SLURM(40) or IBM LSF™). Each launched job runs out of a Singularity(27) 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 $\Delta\Delta G_{folding}$(41–43) 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) (44) vs. randomly placed vs. benign variant sites found in Gnomad (45).
3) ScanNet (46) estimates the likelihood that a variant to disrupt a protein-protein interaction, via an ML algorithm.
4) MusiteDeep (47) predicts protein post-translational (PTM) site modification through a deep-learning framework.
5) Digenic disease interactions are predicted with DigePred(25) and DIEP(26).

Additional suggestions for the interpretation of these outputs are provided in the Supplemental Information.

4 Job Monitoring

Over the course of a VUStruct run, the case report is refreshed at 30 minute intervals, 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 tools on the back end, 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.

5 Reporting

The pipeline generates a case-wide report as a landing page that combines calculated results for each transcript. As shown in Figure 2, the report also integrates queried scores from AlphaMissense(48), 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 (Figure 3). Clicking into a transcript report presents the user with a PFAM domain graphic(49) followed by a tabular summary of calculation results for the associated structures that 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 (Figure 4). The customized viewer also allows backbone coloring of the various calculated constraint scores, and model confidence.

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 (Supplemental Information

Figure 1). The Supplemental Information also suggests approaches to communicating with clinical partners and includes advice on calculation interpretation.

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(44), COSMIC(50), and gnomAD(45) 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(48) 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 $\Delta\Delta G_{folding}$ (41–43) Cartesian and Monomer calculations requires a free academic or paid commercial license from rosettacommons.org.

PROGRAM IN
Personalized Structural Biology

Date	2024-08-05 14:35 USA Central Time (GMT - 6)
Case	SAMPLE_CASE_01
Log Files	Click for log files

Maximum Cluster Compute End 8/8/2024 4:47:58 AM

This page will refresh every 5.0 minutes

Case Preparation

VUStruct Phase	Application	Start Time	End Time	Input File	Log
plan	vustruct_plan.py	2024-08-02T16:12:12.116534	2024-08-02T16:13:00.296837	Input File	Log File
launch	vustruct_launch.py	2024-08-02T16:13:24.589454	2024-08-02T16:13:39.279161		Log File

Refresh Cluster Jobs Information (slurm squeue)

Job Name	Job Id	Start Time	Slurm State	Job Info	ExitCode

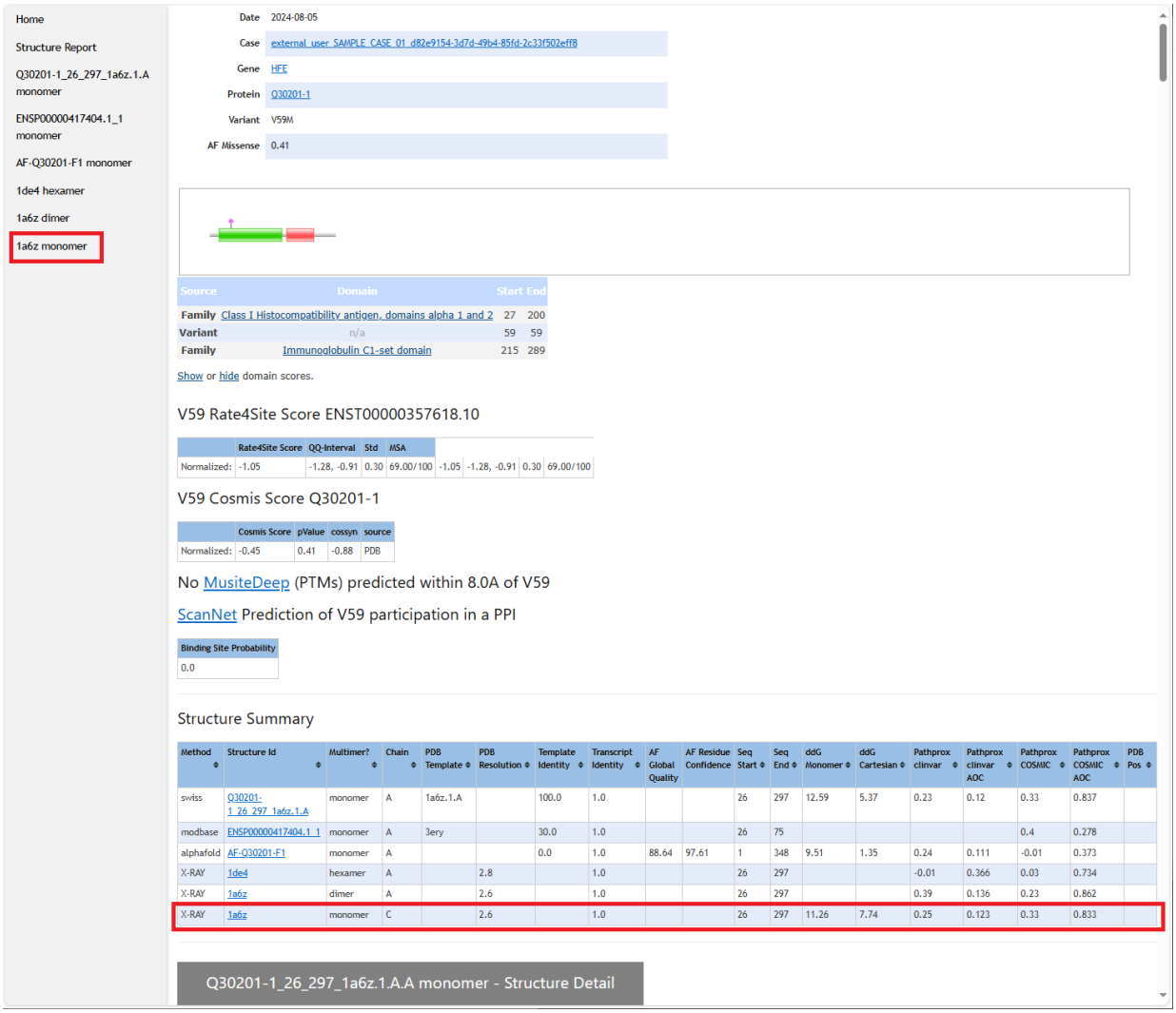
#	Gene	Chrom	Pos	Variant		PP clinvar		PP COSMIC		ddg Monomer		ddg Cartesian		Alpha	ScanNet	MusiteDeep	
	Protein Unp ID	Transcript Refseq	Transcript ENSEMBL	Prot Change	Amino Acid Length	Min	Max	Min	Max	Min	Max	Min	Max	Missense	PPI	PTM	Notes
0	TCF3	chr19	1632083	G/A	651-654			0.01	0.24	0.46	0.46	-0.34	-0.34	0.13 - 0.15	26%		
2	AP3D1	chr19	2115592	C/T	1153-1215			-0.07	0.02	-0.95	-0.5	-0.08	0.26	0.21 - 0.31	67%	67%	
4	AP3D1	chr19	2123841	C/T	1153-1215			0.02	0.06	-0.72	0.68	0.07	0.76	0.38 - 0.56	8%	85%	
6	GUCA1C	chr3	108908148	A/C	209			0.01	0.31	-2.02	1.68	3.7	7.48	0.07	6%		
7	GUCA1C	chr3	108953665	C/T	209			0.18	0.69	7.7	16.78	1.55	4.22	0.63	41%		
8	BMP2K	chr4	78910693	C/T	1161			-0.01	-0.01	3.93	3.93	-11.2	-11.2	0.07	21%	85%	
9	CDH10	chr5	24537430	G/A	788			-0.17	0.22	-1.05	3.81	0.23	2.47	0.07	7%	54% - 72%	
10	CR2	chr1	207474301	C/A	1033-1092			-0.01	0.05	14.66	14.66	6.85	6.85	0.75 - 0.82	56%		
12	DES	chr2	219420154	C/T	470	-0.36	-0.05	0.07	0.29	0.18	1.83	-0.06	0.07	0.26	5%		
13	ESR1	chr6	151808139	C/G	595	-0.02	-0.02	-0.06	-0.06	3.44	3.44	-11.02	-11.02	0.06	64%		
14	HFE	chr6	26090939	G/A	242-348	-0.01	0.42	-0.01	0.43	4.68	14.79	0.45	10.93	0.36 - 0.52	2%		
Report	Q30201-1	NM_000410.3	ENST00000357618.10	V59M	348	-0.01	0.39	-0.01	0.4	9.51	12.59	1.35	7.74	0.41	2%		
Report	Q30201-10	NM_139003.2	ENST00000336625.12	V59M	242	0.25	0.42	0.23	0.43	5.93	11.26	3.34	9.05	0.46			
Report	Q30201-3	NM_139006.2	ENST00000461397.5	V59M	334	0.2	0.39	0.23	0.33	11.26	14.79	7.74	10.93	0.52			
Report	Q30201-5	NM_139009.2	ENST00000397022.7	V36M	325	0.2	0.41	0.32	0.4	5.44	11.26	0.45	7.74	0.44			
Report	Q30201-7	NM_139004.2	ENST00000317896.11	V59M	256	0.25	0.39	0.23	0.33	4.68	11.26	0.46	7.74	0.36			
19	LIMK1	chr7	74121197	C/T	613-647			-0.11	0.07	-2.41	0.71	-1.39	1.07	0.09	29%	71% - 87%	
21	MBD5	chr2	148468836	C/T	851-1494	-0.09	-0.09	-0.1	-0.0	18.42	18.42	-24.06	-24.06	0.59 - 0.90	40%	88%	
24	SERPINA6	chr14	94309814	T/C	405			-0.03	0.19	-1.27	3.14	2.46	5.0	0.16	8%		
25	WDR41	chr5	77438279	T/C	404-459	-0.11	-0.11	-0.04	0.35	-4.33	63.26	3.61	19.23	0.96 - 0.97	1%	72%	

DiGePred Analysis

Top Scoring DiGePred Gene Pairs

DiGePred Score	Gene A	Gene B
0.5	CR2	TCF3

Figure 2 The case report landing page contains a table, in which 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 of 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 figure).



Home
Structure Report
Q30201-1_26_297_1a6z.1.A monomer
ENSP00000417404.1_1 monomer
AF-Q30201-F1 monomer
1de4 hexamer
1a6z dimer
1a6z monomer

Date	2024-08-05
Case	external_user_SAMPLE_CASE_01_d82e9154-3d7d-49b4-85fd-2c33f502eff8
Gene	HFE
Protein	Q30201-1
Variant	V59M
AF Missense	0.41

Source	Domain	Start	End
Family	Class I Histocompatibility antigen, domains alpha 1 and 2	27	200
Variant	n/a	59	59
Family	Immunoglobulin C1-set domain	215	289

Show or hide domain scores.

V59 Rate4Site Score ENST00000357618.10

	Rate4Site Score	QQ-Interval	Std	MSA				
Normalized:	-1.05	-1.28, -0.91	0.30	69.00/100	-1.05	-1.28, -0.91	0.30	69.00/100

V59 Cosmis Score Q30201-1

	Cosmis Score	pValue	cossyn	source
Normalized:	-0.45	0.41	-0.88	PDB

No MusiteDeep (PTMs) predicted within 8.0A of V59

ScanNet Prediction of V59 participation in a PPI

Binding Site Probability
0.0

Structure Summary

Method	Structure Id	Multimer?	Chain	PDB Template	PDB Resolution	Template Identity	Transcript Identity	AF Global Quality	AF Residue Confidence	Seq Start	Seq End	ddG Monomer	ddG Cartesian	Pathprox clinvar	Pathprox clinvar AOC	Pathprox COSMIC	Pathprox COSMIC AOC	PDB Pos
swiss	Q30201-1_26_297_1a6z.1.A	monomer	A	1a6z.1.A		100.0	1.0			26	297	12.59	5.37	0.23	0.12	0.33	0.837	
modbase	ENSP00000417404.1_1	monomer	A	3ery		30.0	1.0			26	75					0.4	0.278	
alphafold	AF-Q30201-F1	monomer	A			0.0	1.0	88.64	97.61	1	348	9.51	1.35	0.24	0.111	-0.01	0.373	
X-RAY	1de4	hexamer	A		2.8		1.0			26	297			-0.01	0.366	0.03	0.734	
X-RAY	1a6z	dimer	A		2.6		1.0			26	297			0.39	0.136	0.23	0.862	
X-RAY	1a6z	monomer	C		2.6		1.0			26	297	11.26	7.74	0.25	0.123	0.33	0.833	

Q30201-1_26_297_1a6z.1.A.A monomer - Structure Detail

Figure 3 The top of each transcript variant report shows the variant location as a pink diamond in the context of the protein's PFAM domain(49) annotation. This is followed by results for Rate4Site, COSMIS, MusiteDeep, and ScanNet calculations. A key table is the Structure Summary, which lists all of 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 figure).

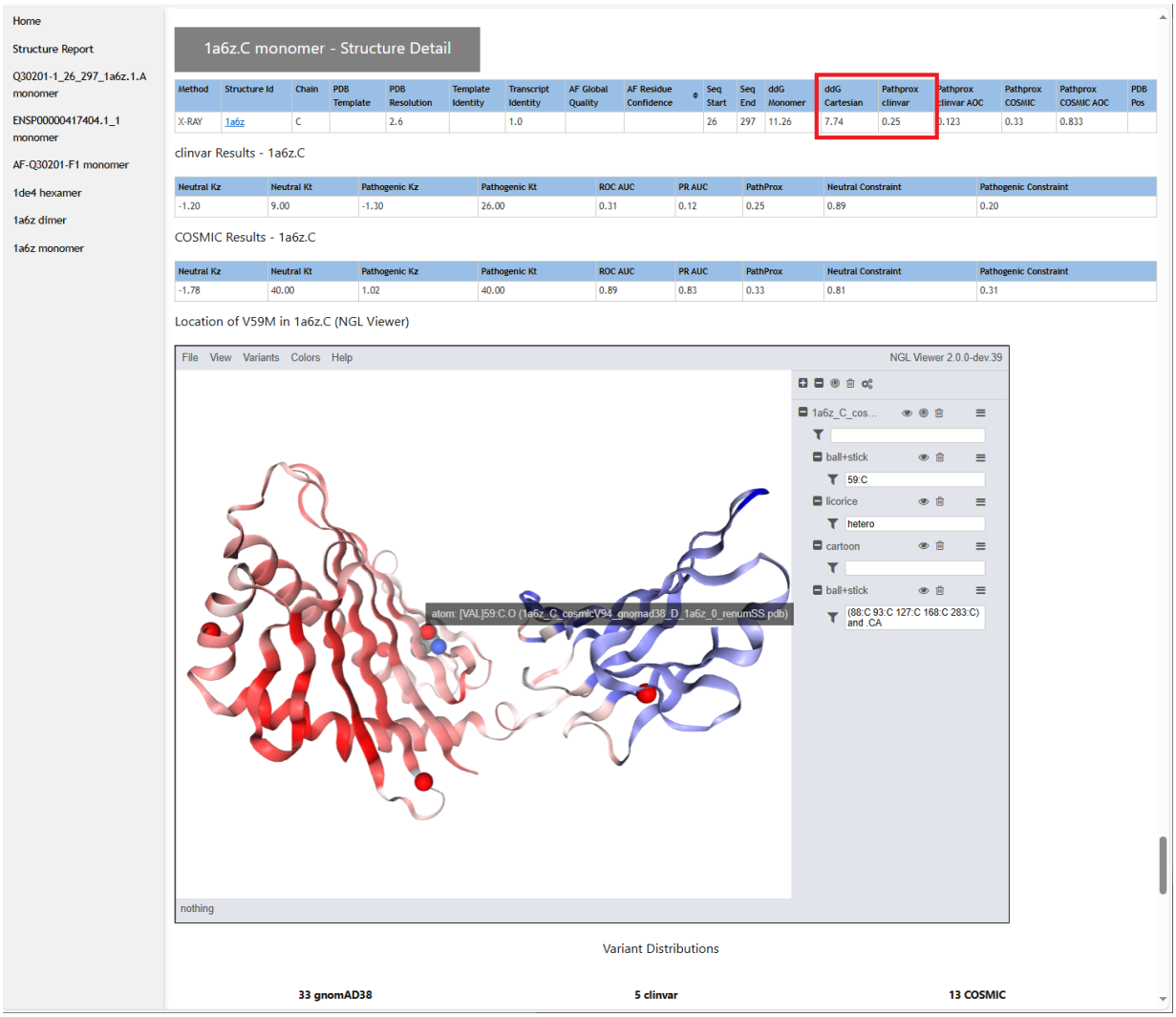


Figure 4. The section of the results page devoted to each structural model shows the results of the ΔΔG and PathProx calculations, highlighted in red, along with associated statistics to judge reliability. Each structure is displayed in a customized and interactive NGL Viewer(19) session. This view can 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.

Results/Application of VUStruct in the interpretation of clinical data:

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 (41–43), 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 (51)). 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 (52).

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(53).

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 pipeline limitation is that all calculations are based on sets of single structural models, and the implications of dynamics are not presently considered. Multi-conformer generation is an active area of research(54). We plan to integrate that work into the pipeline, so that more conformational states are sampled. Additionally, we are integrating AlphaFold models for non-canonical transcript sequences(55). Pending its public opening, we hope to mine the AlphaFold 3 repository for its updated structural coverage that includes multimeric complexes(56). Predictions of digenic interactions should benefit from model retraining, given the emergence of new ground truth data sets(57).

Support and Thanks

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.

Grants

This work has been supported by NIH grant R01 LM013434-04.

J.M. acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1052 (209933838), and SPP 2363 (460865652). J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. 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).

References

1. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res. 2019 Jan 8;47(D1):D1038–43.

2. Davydov E V., 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 Comput Biol. 2010 Dec 2;6(12):e1001025.

3. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005 Jul;15(7):901–13.

4. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

5. Ramensky V. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002 Sep 1;30(17):3894–900.

6. Kobren SN, Baldridge D, Velinder M, Krier JB, LeBlanc K, Esteves C, et al. Commonalities across computational workflows for uncovering explanatory variants in undiagnosed cases. Genetics in Medicine. 2021 Jun;23(6):1075–85.

7. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to Predict Loss-of-Function and Gain-of-Function Mutations. Genet Test Mol Biomarkers. 2010 Aug;14(4):533–7.

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.

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 Dec;23:892–904.

10. Taipale M. Disruption of protein function by pathogenic mutations: common and uncommon mechanisms. Biochemistry and Cell Biology. 2019 Feb;97(1):46–57.

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. Human Genetics and Genomics Advances. 2022 Oct;3(4):100131.

12. Nielsen S V., 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 Apr 19;13(4):e1006739.

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 Dec 22;14(1):77.

14. Laskowski RA, Stephenson JD, Sillitoe I, Orengo CA, Thornton JM. VarSite: Disease variants and protein structure. Protein Science. 2020 Jan 27;29(1):111–9.

15. Stephenson JD, Totoo P, Burke DF, Jänes J, Beltrao P, Martin MJ. ProtVar: mapping and contextualizing human missense variation. Nucleic Acids Res. 2024 May 20;

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 May;431(11):2197–212.

17. Philipp M, Moth CW, Ristic N, Tiemann JKS, Seufert F, Panfilova A, et al. <scp>MutationExplorer</scp> : a webserver for mutation of proteins and 3D visualization of energetic impacts. Nucleic Acids Res. 2024 Apr 22;

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.

19. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics. 2018 Nov 1;34(21):3755–8.

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. The American Journal of Human Genetics. 2018 Mar;102(3):415–26.

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 Dec 23;19(1):18.

22. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 2016 Jul 8;44(W1):W344–50.

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 Jul 1;18(suppl_1):S71–7.

24. Li B, Roden DM, Capra JA. The 3D mutational constraint on amino acid sites in the human proteome. Nat Commun. 2022 Jun 7;13(1):3273.

25. Mukherjee S, Cogan JD, Newman JH, Phillips JA, Hamid R, Meiler J, et al. Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. The American Journal of Human Genetics. 2021 Oct;108(10):1946–63.

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.

27. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017 May 11;12(5):e0177459.

28. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

29. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D988–95.

30. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016 Dec 6;17(1):122.

31. 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 Dec 23;41(12):1776–86.

32. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017 May;27(5):849–64.

33. 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 Apr 14;604(7905):310–5.

34. 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 Jan 8;47(D1):D482–9.

35. Berman HM. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

36. 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 Jan;42(D1):D336–46.

37. 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 Jan 4;45(D1):D313–9.

38. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

39. Leach P, Mealling M, Salz R. A Universally Unique IDentifier (UUID) URN Namespace. 2005 Jul.

40. Yoo AB, Jette MA, Grondona M. SLURM: Simple Linux Utility for Resource Management. In 2003. p. 44–60.

41. Park H, Bradley P, Greisen P, 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 Dec 13;12(12):6201–12.

42. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins: Structure, Function, and Bioinformatics. 2011 Mar 3;79(3):830–8.

43. 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 Oct 8;8.

44. 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 Jan 4;46(D1):D1062–7.

45. 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 May 28;581(7809):434–43.

46. Tubiana J, Schneidman-Duhovny D, Wolfson HJ. ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods. 2022 Jun 30;19(6):730–9.

47. 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 Dec 15;33(24):3909–16.

48. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (1979). 2023 Sep 22;381(6664).

49. 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 Jan 8;49(D1):D412–9.

50. 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 Jan 8;47(D1):D941–7.

51. 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 Apr 27;32(10):946–8.

52. Ezell KM, Tinker RJ, Furuta Y, Gulsevin A, Bastarache L, Hamid R, et al. Undiagnosed Disease Network collaborative approach in diagnosing rare disease in a patient with a mosaic <scp> *CACNA1D* </scp> variant. Am J Med Genet A. 2024 Jul 21;194(7).

53. Furuta Y, Tinker RJ, Gulsevin A, Neumann SM, Hamid R, Cogan JD, et al. Probable digenic inheritance of <scp>Diamond–Blackfan</scp> anemia. Am J Med Genet A. 2024 Mar 27;194(3).

54. 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 Jan 12;

55. Sommer MJ, Cha S, Varabyou A, Rincon N, Park S, Minkin I, et al. Structure-guided isoform identification for the human transcriptome. Elife. 2022 Dec 15;11.

56. 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 May 8;

57. 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 Apr 12;2022.