Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

VarMeter2: An enhanced structure-based method for predicting pathogenic missense variants through Mahalanobis distance

Shiho Ohno[a], Chika Ogura[b], Akane Yabuki[c], Kazuyoshi Itoh[d], Noriyoshi Manabe[a], Kiyohiko Angata[d], Akira Togayachi[d], Kiyoko Aoki-Kinoshita[c,d,e], Jun-ichi Furukawa[e], Kei-ichiro Inamori[f], Jin-Ichi Inokuchi[g], Tadashi Kaname[h,*], Shoko Nishihara[c,d,**], Yoshiki Yamaguchi[a,***]

[a] *Division of Structural Glycobiology, Institute of Molecular Biomembrane and Glycobiology, Tohoku Medical and Pharmaceutical University, Sendai, Miyagi 981-8558, Japan*
[b] *Department of Science and Engineering for Sustainable Innovation, Faculty of Science and Engineering, Soka University, Japan*
[c] *Department of Biosciences, Graduate School of Science and Engineering, Soka University, Japan*
[d] *Glycan and Life Systems Integration Center (GaLSIC), Soka University, Hachioji 192-8577, Japan*
[e] *Institute for Glyco-Core Research (iGCORE), Nagoya University, Nagoya 466-8601, Japan*
[f] *Division of Glycopathology, Institute of Molecular Biomembrane and Glycobiology, Tohoku Medical and Pharmaceutical University, Sendai, Miyagi 981-8558, Japan*
[g] *Forefront Research Center, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan*
[h] *Department of Genome Medicine, National Center for Child Health and Development, Tokyo 157-0074, Japan*

ARTICLE INFO

Keywords:
Missense variant
Pathogenicity
N-sulphoglucosamine sulphohydrolase
Solvent accessible surface area
Mutation energy
pLDDT

ABSTRACT

Various computational methods have been developed to predict the pathogenicity of missense variants, which is crucial for diagnosing rare diseases. Recently, we introduced VarMeter, a diagnostic tool for predicting variant pathogenicity based on normalized solvent-accessible surface area (nSASA) and mutation energy calculated from AlphaFold 3D models, and validated it on arylsulfatase L. To evaluate the broader applicability of VarMeter and enhance its predictive accuracy, here we analyzed 296 pathogenic and 240 benign variants extracted from the ClinVar database. By comparing structural features including nSASA, mutation energy, and predicted local distance difference test (pLDDT) score, we identified distinct characteristics between pathogenic and benign variants. These features were used to develop VarMeter2, which classifies variants based on Mahalanobis distance. VarMeter2 achieved a prediction accuracy of 82 % for the ClinVar dataset, a marked improvement over the original VarMeter (74 %), and 84 % for published missense variants of *N*-sulphoglucosamine sulphohydrolase (SGSH), an enzyme associated with Sanfillippo syndrome A. Application of VarMeter 2 to SGSH variants in our clinical database identified a novel SGSH variant, Q365P, as pathogenic. The recombinant Q365P protein lacked enzymatic activity as compared with wild-type SGSH. Furthermore, it was largely retained in the endoplasmic reticulum and failed to reach the Golgi, probably due to misfolding. Protein stability assays confirmed reduced stability of the variant, further explaining its loss of function. Consistently, the patient homozygous for this variant was diagnosed with Sanfilippo syndrome A. These results underscore the predictive power and versatility of VarMeter2 in assessing the pathogenicity of missense variants.

1. Introduction

Advancements in genome sequencing technologies have led to the accumulation of vast data on human genetic variants. Due to the challenges in distinguishing pathogenic variants from benign ones, however, many rare genetic diseases remain without identified causative variants. One approach to addressing this issue has been the development of algorithms that assess missense variants based on 3D

* Correspondence to: Department of Genome Medicine, National Center for Child Health and Development, 2-10-1 Okura, Setagaya-ku, Tokyo 157-0074, Japan.
** Correspondence to: Glycan and Life Systems Integration Center (GaLSIC), Soka University, 1-236 Tangi-machi, Hachioji 192-8577, Japan.
*** Correspondence to: Division of Structural Biology, Institute of Molecular Biomembrane and Glycobiology, Tohoku Medical and Pharmaceutical University, 4-4-1 Komatsushima, Aoba-ku, Sendai, Miyagi 981-8558, Japan.
E-mail addresses: kaname-t@ncchd.go.jp (T. Kaname), shoko@soka.ac.jp (S. Nishihara), yyoshiki@tohoku-mpu.ac.jp (Y. Yamaguchi).

https://doi.org/10.1016/j.csbj.2025.02.008
Received 31 October 2024; Received in revised form 14 February 2025; Accepted 14 February 2025
Available online 1 March 2025

protein structures [1,2].

In this context, we recently developed a novel prediction method called VarMeter (VARiant impact predicting MEthod combining muTation Energy and solvent-accessible surface aRea) [3]. Unlike many other pathogenicity prediction tools, which frequently incorporate sequence conservation and human variation data, VarMeter focuses solely on physical parameters derived from AlphaFold2 structural models [4,5]. The two key parameters are (i) normalized solvent-accessible surface area (nSASA) of the amino acid residue in the wild-type protein; and (ii) mutation energy, which reflects the difference in Gibbs free energy ($\Delta\Delta G$) of protein folding between wild-type and variant proteins. This unique approach enables VarMeter to make predictions independent of evolutionary data, facilitating its application to proteins with limited sequence conservation or functional annotations. The initial version of VarMeter was developed using 70 pathogenic and 16 benign missense variants from three proteins: arylsulfatase L (ARSL), chitobiosyldiphosphodolichol β-mannosyltransferase (ALG1), and mannose-6-phosphate isomerase (MPI). This method was successfully applied to predict the pathogenicity of a newly identified ARSL variant in a patient with undiagnosed disease [3].

A crucial factor in improving the accuracy of missense variant predictions is determining the optimal thresholds for nSASA and mutation energy to discriminate between pathogenic and benign variants. This can be achieved by increasing the number of variants analyzed across a wider variety of proteins. Additionally, the confidence level of AlphaFold2 models, reflected in the predicted local distance difference test (pLDDT) score, is important as it directly affects prediction reliability. Because VarMeter relies on AlphaFold2 models, careful consideration of residue-level confidence is essential for accurate predictions.

To improve the precision and generalizability of VarMeter, here we analyzed key structural parameters — nSASA, mutation energy, and pLDDT scores — using missense variants from the ClinVar database [6], a widely used resource for training and validating pathogenicity prediction models [7,8]. By incorporating Mahalanobis distance into the analysis of these three parameters, we developed an enhanced version of our tool, termed VarMeter2. To evaluate its effectiveness, we applied VarMeter2 to *N*-sulphoglucosamine sulphohydrolase (SGSH), a protein associated with numerous missense variants linked to Sanfilippo syndrome A (mucopolysaccharidosis type III) [9]. We compared our results with those from VarMeter [3], AlphaMissense [10], and CADD [11]. Furthermore, the pathogenicity of an SGSH variant (Q365P), newly identified by VarMeter2 was experimentally confirmed through enzymatic activity, immunostaining, and protein stability assays.

2. Materials and methods

2.1. Missense variants from the ClinVar database and related parameters

To update VarMeter, missense variants were extracted from the ClinVar database as of July 6, 2023. The dataset includes 296 pathogenic missense variants of 24 proteins and 240 benign variants of 23 proteins, with a review status of "Expert panel" (three stars). The amino acid sequences of each protein were retrieved from the UniProt database [12]. Allele frequencies of the missense variants were extracted from the Genome Aggregation Database (gnomAD) v4.1 (December 3, 2024) [13], and data were obtained from the TogoVar ftp site [14]. AlphaMissense scores were obtained from Zendo (https://zenodo.org/records/8208688; June 2024), and CADD v1.7 scores were downloaded from the CADD homepage (https://cadd.bihealth.org/; December 19, 2024) [15]. The ClinVar dataset analyzed in this study is included in Supplemental Table S1.

2.2. Preparation of 3D models of variants

The 3D models of the reference (wild-type) proteins were obtained from the AlphaFold2 Protein Structure Database [4]. The 3D models of each missense variant were generated using the "Calculate Mutation Energy/Stability" module in Discovery Studio 2021 (BIOVIA, Dassault Systèmes, San Diego, CA, USA).

2.3. Calculation of mutation energy and normalized solvent-accessible surface area

Mutation energy ($\Delta\Delta G$), reflecting the impact of the mutation on protein stability in kcal/mol, was calculated using the "Calculate Mutation Energy/Stability" module in Discovery Studio 2021, as described previously [3]. The solvent-accessible surface area (SASA) of each residue was calculated based on the AlphaFold2 model of the wild-type protein using Discovery Studio 2021. Each SASA was normalized (nSASA) by using reference SASA values [16].

2.4. Calculation of Mahalanobis distance from mutation energy, nSASA and pLDDT

The squared Mahalanobis distance (D_i^2) was calculated for the 296 pathogenic variants and 240 benign variants in the ClinVar dataset using mutation energy (x_i), nSASA (y_i) and pLDDT (z_i). The following equation, which incorporates the inverse covariance matrix (3 × 3), was used:

$$D_i^2 = \begin{pmatrix} x_i - \bar{x} & y_i - \bar{y} & z_i - \bar{z} \end{pmatrix} \begin{pmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{pmatrix}^{-1} \begin{pmatrix} x_i - \bar{x} \\ y_i - \bar{y} \\ z_i - \bar{z} \end{pmatrix}$$

where $\bar{x}$, $\bar{y}$ and $\bar{z}$ represent the mean values of mutation energy, nSASA and pLDDT, respectively; s_x^2, s_y^2 and s_z^2 are the variances of mutation energy, nSASA and pLDDT, respectively; and s_{xy}, s_{xz} and s_{yz} are the covariances between mutation energy and nSASA, mutation energy and pLDDT, and nSASA and pLDDT, respectively. The squared Mahalanobis distances (D_P^2 for pathogenic and D_B^2 for benign variants) were calculated and used to predict the pathogenicity of the variants.

2.5. Missense variants of SGSH

Missense variants of SGSH were collected from published studies using the Human Gene Mutation Database (HGMD). Pathogenic variants of SGSH were identified from patients clinically diagnosed with Sanfillippo syndrome A and were experimentally validated to have reduced enzymatic activity. In-house whole-exome sequencing data from patients with rare or undiagnosed disease were screened for the presence of SGSH gene variants. The studies for in-house database were approved by the ethical committee of the National Center for Child Health and Development (Approval No. 2020-326). SGSH variants with a ClinVar clinical significance of "benign" and a homozygote count of 2 or more were extracted from the gnomAD database (v2.1.1) (https://gnomad.broadinstitute.org/) as benign variants. The amino acid sequence of SGSH was obtained from the UniProt database (UniProt ID: P51688). Allele frequencies and prediction scores (AlphaMissense and CADD) were obtained using the same approach applied to the ClinVar dataset, as described in Section 2.1.

2.6. Plasmid construction

All expression vectors were generated using the Gateway cloning system (Thermo Fisher Scientific, Waltham, MA, USA). In brief, the gene encoding *SGSH* (NCBI reference sequence NM_000199) was amplified from cDNA prepared from HEK293 cells (Human Embryonic Kidney cells 293) using attB adaptor primers, and recombined into pDONR201 (Thermo Fisher Scientific, Waltham, MA, USA) to generate an entry clone. Nucleotide substitutions (A1094 to C: corresponding to amino

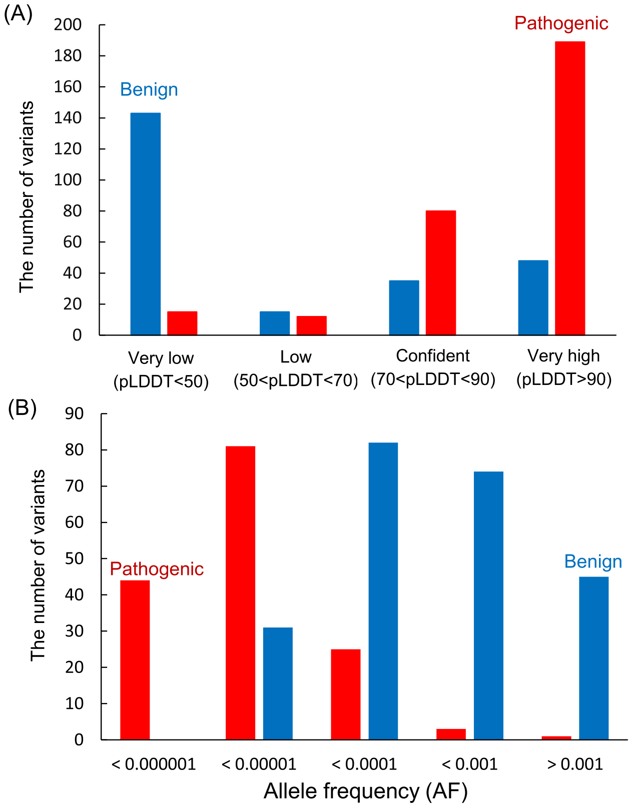


Fig. 1. Distribution of benign and pathogenic variants by pLDDT score and by allele frequency. (A) Benign and pathogenic variants were grouped into four confidence levels: very low (pLDDT<50), low (50 <pLDDT<70), confident (70 <pLDDT<90), and very high (pLDDT>90). (B) Benign and pathogenic variants were grouped into five levels of allele frequency (AF): AF< 0.000001, 0.000001 <AF< 0.00001, 0.00001 <AF< 0.0001, 0.0001 <AF< 0.001, and AF> 0.001. In (A) and (B), red bars indicate pathogenic variants; blue bars indicate benign variants.

Table 1
Statistical parameters of mutation energy, nSASA and pLDDT for the pathogenic and benign variants from the ClinVar dataset.

	Pathogenic (n = 296)			Benign (n = 240)		
	Mutation energy (kcal/ mol)	nSASA	pLDDT	Mutation energy (kcal/ mol)	nSASA	pLDDT
Mean ± SD	3.9 ± 9.6	0.21 ± 0.28	88.2 ± 14.2	0.1 ± 2.0	0.64 ± 0.29	53.1 ± 27.9
Variance	91.3	0.076	202.0	3.9	0.086	775.5
Covariance	–0.4 (mutation energy–nSASA) 18.0 (mutation energy–pLDDT) –2.7 (nSASA–pLDDT)			–0.1 (mutation energy–nSASA) 9.8 (mutation energy–pLDDT) –6.1 (nSASA–pLDDT)		

Abbreviations: nSASA, normalized surface-accessible area; pLDDT, predicted local distance difference test.

acid substitution Q365 to P) were introduced by site-directed mutagenesis using a KOD-Plus-Mutagenesis Kit (TOYOBO, Osaka, Japan). The entry clone encoding wild-type or variant *SGSH* genes without signal peptide (M1 to A20) was recombined via LR reaction into plasmid pFLAG-CMV-3-DEST-IRES-puro (a kind gift of Dr. Takashi Sato) [17], which contains the secretory signal peptide of Preprotrypsin, a FLAG tag at the N-terminus with a linker (LAAANSSIDLISTSLYKK), and the IRES-puro fragment from pIRESpuro3 (Takara Bio USA, Inc., Mountain View, CA, USA).

2.7. HEK293 cell transfection and selection of stable transformants

HEK293 cells were maintained in DMEM (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 10 % FBS (BioWest, Nuaillé, France) and 1 × Penicillin-Streptomycin Solution (FUJIFILM Wako Pure Chemical Corporation, Osaka, Japan) at 37℃ and 5 % CO_2. Cells were seeded in a 10-cm dish at 6 × 10^6 cells and incubated for 24 h. Cultured cells were transfected with 30 μg of expression vector (wild-type or Q365P SGSH-FLAG fusion protein) using 30 μL of Lipofectamine 2000 (Thermo Fisher Scientific, Waltham, MA, USA). Cells were grown for an additional 48 h, and then one-tenth of the cells were transferred to medium containing 2 μg/mL of puromycin (Merck, Darmstadt, Germany) to select transformants: HEK293-WT cells and HEK293-Q365P cells, stably expressing wild-type and Q365P SGSH-FLAG fusion proteins, respectively. The transformants were harvested, detached by treatment with 0.02 % EDTA (Kanto Chemical Co., Inc. Tokyo, Japan), washed three times with PBS, and stored at −80°C until use.

2.8. Protein expression and purification

The stored cell pellet was thawed on ice, resuspended in 50 mM Tris-HCl (pH 7.5), 150 mM NaCl, 1 % Triton X-100, and 1 × Protease inhibitor cocktail (Nacalai Tesque, Kyoto, Japan), and incubated on ice with occasional mixing for 10 min. The supernatant containing solubilized protein was obtained by centrifugation at 20,600 ×g for 5 min at 4℃, diluted five times with 50 mM Tris-HCl (pH 7.5) and 150 mM NaCl, and incubated with anti-FLAG magnetic beads (Anti-DYKDDDDK tag Antibody Magnetic Beads; FUJIFILM Wako Pure Chemical Corporation, Osaka, Japan), overnight at 4℃. After removing the supernatant, the beads were washed with 5 mM Tris-HCl (pH 7.5), 13.8 mM NaCl and 0.27 mM KCl. The beads with purified wild-type or Q365P SGSH-FLAG fusion protein were used for SGSH activity assay.

2.9. Western blot analysis

Samples were solubilized with 62.5 mM Tris-HCl (pH 6.8) containing

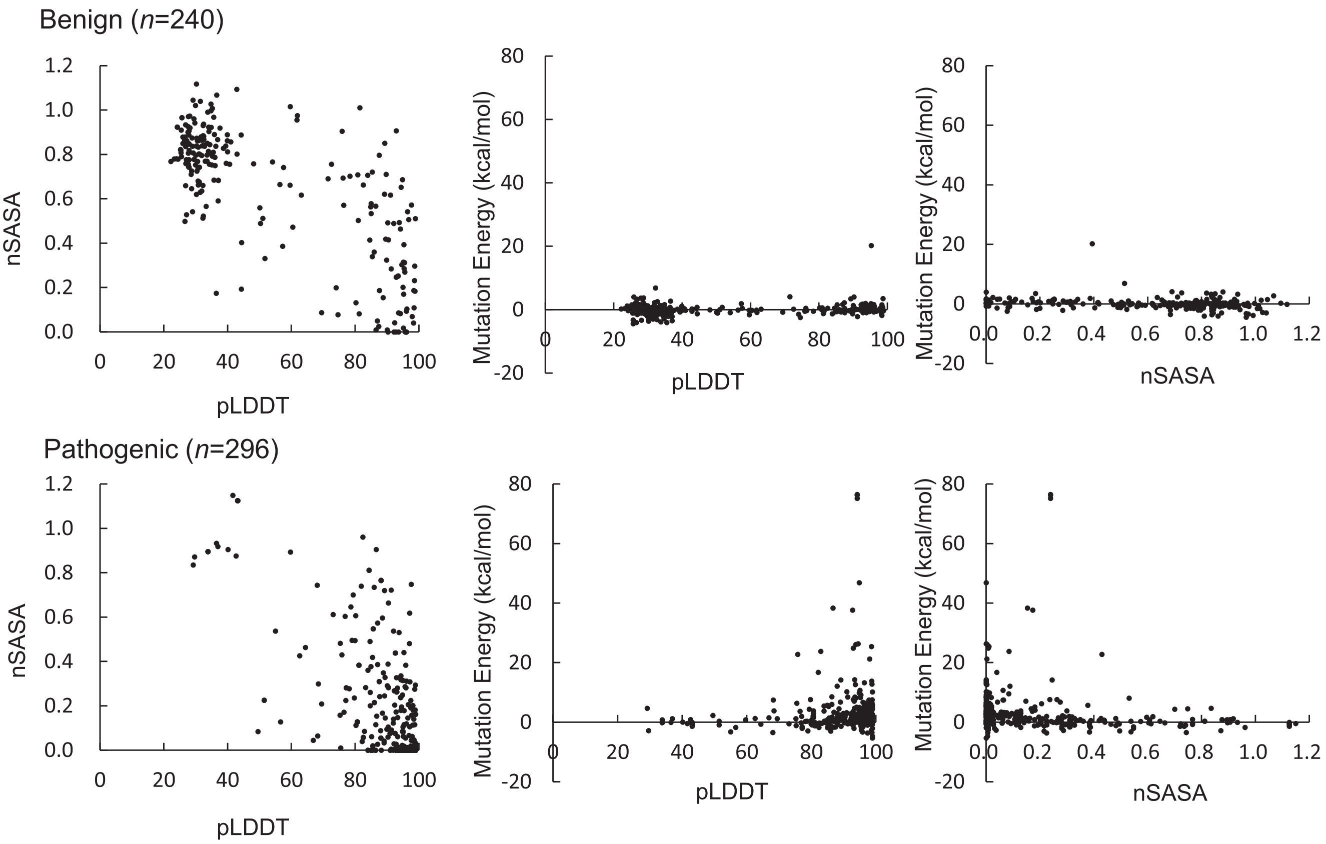


Fig. 2. Scatter plots of nSASA, mutation energy, and pLDDT for variants from the ClinVar dataset. Upper panels, benign variants ($n = 240$); lower panels, pathogenic variants ($n = 296$). Each plot illustrates the relationships between these three parameters, highlighting differences between benign and pathogenic variant behavior.

Table 2
Square of Mahalanobis distance (D_{P}^2 and D_{B}^2) for the pathogenic and benign variants.

	Pathogenic ($n = 296$)	Benign ($n = 240$)
D_{P}^2 (mean ± SD)	3.0 ± 6.8	11.6 ± 7.5
D_{B}^2 (mean ± SD)	30.5 ± 157.9	3.0 ± 7.0

10 % glycerol, 2 % SDS, 1 % 2-mercaptoethanol and 0.01 % bromophenol blue, denatured at 99℃, separated on an SDS–PAGE, and transferred to a PVDF membrane (Millipore, Burlington, MA, USA). Membranes were blocked with 1 % BSA and incubated with HRP-conjugated anti-FLAG mouse monoclonal antibody (Sigma, St. Louis, MO, USA) at 4 °C overnight. Detection was performed with ECL Prime (Cytiva, Tokyo, Japan).

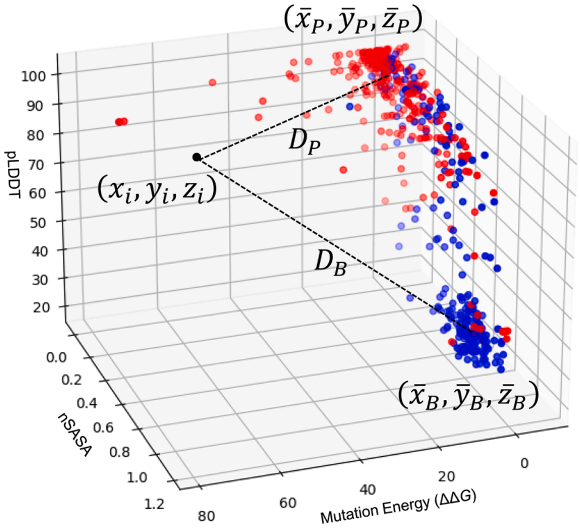


Fig. 3. Conceptual diagram of variant classification using Mahalanobis distance. The means of three variables (mutation energy, nSASA and pLDDT) are calculated for the pathogenic (red) and benign (blue) groups, denoted as ($\bar{x}_P, \bar{y}_P, \bar{z}_P$) and ($\bar{x}_B, \bar{y}_B, \bar{z}_B$), respectively. The Mahalanobis distances (D_P and D_B) are calculated for each data point (x_i, y_i, z_i) using the inverse covariance matrix (see Materials and Methods). A variant is classified as pathogenic if $D_P < D_B$ and as benign if $D_P > D_B$.

2.10. Immunostaining of HEK293 stable transformants expressing wild-type or Q365P SGSH protein

Cells were fixed with 4 % paraformaldehyde in PBS at room temperature for 10 min and then washed three times with PBS. Fixed cells were blocked with 1 % BSA/0.1 % Triton X-100 in PBS. For primary labeling, cells were incubated with anti-FLAG mouse monoclonal antibody (Sigma-Aldrich, St. Louis, MO, USA) and anti-calnexin (CANX) rabbit polyclonal antibody (GeneTex, Irvine, CA, USA) or anti-GOLPH2 rabbit polyclonal antibody (GeneTex, Irvine, CA, USA) at 4℃ overnight. After three washes with PBS, the cells were stained with Alexa Fluor 647-conjugated anti-mouse IgG1 (Thermo Fisher Scientific, Waltham, MA, USA) and Alexa Fluor 555-conjugated anti-rabbit IgG (Thermo Fisher Scientific, Waltham, MA, USA). Hoechst 33258 (FUJIFILM Wako Pure Chemical Corporation, Osaka, Japan) was used as a nuclear counterstain. Images were obtained using an LSM 700 confocal laser microscope (Carl Zeiss, Oberkochen, Baden-Württemberg, Germany). Three-dimensional images were constructed by IMARIS software

Table 3
Prediction accuracy of VarMeter, VarMeter2, AlphaMissense and CADD for the ClinVar dataset.

Prediction tool	Pathogenic (%)	Benign (%)	Overall accuracy (%)
VarMeter	72	75	74
VarMeter2	85	78	82
AlphaMissense[a]	91	92	91
CADD[a]	96	85	91

[a] Ambiguous variants were not included in the calculation of prediction accuracy.

(Bitplane, Belfast, United Kingdom). Colocalization analysis was performed by ImarisColoc software (Bitplane AG, Zurich, Switzerland).

2.11. Quantitative analysis of SGSH colocalization in confocal microscopy images

Quantitative analysis of SGSH colocalization with the Golgi marker GOLPH2 or the ER marker calnexin (CANX) was evaluated in two-dimensional images using ZEN 2012 software (Black Edition, Carl Zeiss). Quantitative results were calculated as the ratio of the total intensity of SGSH colocalizing with GOLPH2 or CANX to the total intensity of SGSH (Figs. 6F and 6G), or the ratio of the total number of pixels representing GOLPH2 or CANX colocalizing with SGSH to the total number of pixels representing GOLPH2 or CANX (Supplemental Fig. S2A and S2B). At least 12 cells were analyzed for each colocalization measurement.

2.12. Real-time PCR

Total RNA was isolated from cells using TRI Reagent (Molecular Research Center, Cincinnati, OH, USA) and reverse transcribed using Oligo dT primer (Thermo Fisher Scientific, Waltham, MA, USA) and Super Script II Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA, USA). Real-time PCR of *SGSH* mRNA in each HEK293 transformant was performed by Quant Studio 12 K Flex (Applied Biosystems, Waltham, MA, USA) using primers for *SGSH* (forward, 5’-GCATCAGAATGGGATGTACGG-3’; reverse, 5’-GAAGAAAGGCCGGTCATCC-3’) and *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*; forward, 5’-CAAAGTTGTCATGGATGACC-3’; reverse, 5’- CCATGGAGAAGGCTGGGG-3’). The amount of each *SGSH* mRNA was normalized to that of *GAPDH* mRNA in the same sample.

2.13. Enzymatic assay of SGSH variant

The activity of the purified wild-type and Q365P SGSH proteins was measured by a fluorometric two-step enzyme assay using 4-methylumbelliferyl 2-deoxy-2-sulfamino-α-D-glucopyranoside (4-MU-GlcNS) (Biosynth, Zurich, Switzerland) as substrate, as described previously [18,19] with a slight modification. In this assay, 4MU-GlcNS is hydrolyzed by SGSH to 4-methylumbelliferyl 2-deoxy-2-amino-α-D-glucopyranoside (4-MU-GlcNH$_2$), which is then hydrolyzed to the fluorescent product, 4-methylumbelliferone (4-MU), by yeast α-glucosidase (Sigma-Aldrich, St. Louis, MO, USA). The purified protein on beads was incubated in 30 μL of McIlvaine’s buffer containing 3.3 mM substrate for 16 h at 37°C; 30 μL of McIlvaine’s buffer containing 0.1 units of α-glucosidase was then added, followed by incubation for 24 h at 37°C.

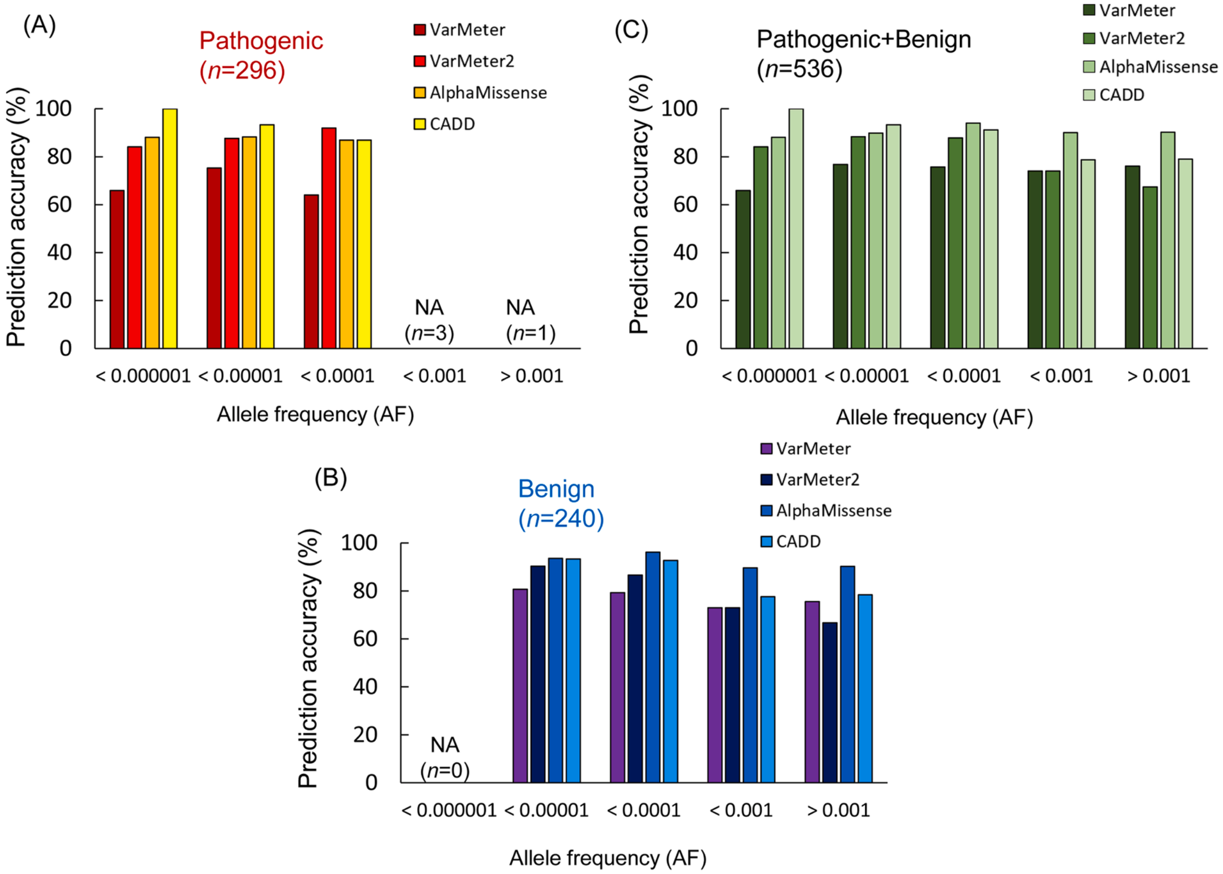


Fig. 4. Prediction accuracy of VarMeter, VarMeter2, AlphaMissense, and CADD methods for the ClinVar dataset based on allele frequency. (A) Prediction accuracy for pathogenic variants ($n = 296$). (B) Prediction accuracy for benign variants ($n = 240$). (C) overall prediction accuracy for all variants (pathogenic and benign, $n = 536$).

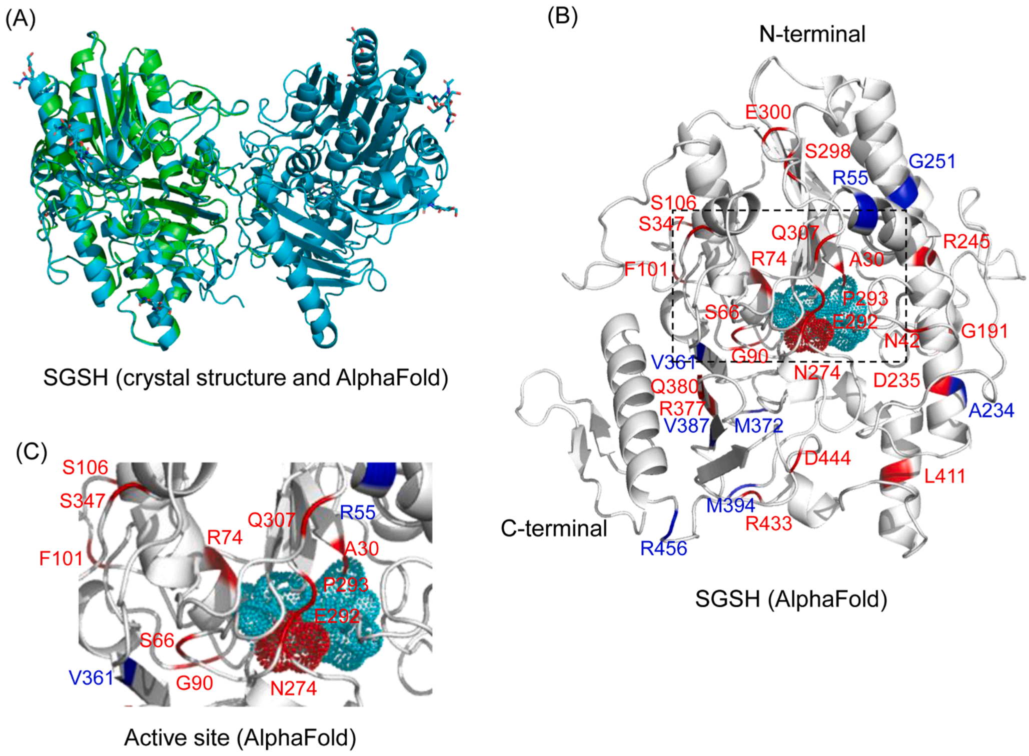


Fig. 5. Structural comparison and variant mapping of human SGSH. (A) Crystal structure of human SGSH dimer resolved at 2 Å resolution (PDB ID: 4MHX, cyan) and the 3D AlphaFold model of monomeric human SGSH (green). The AlphaFold model is superimposed on the A chain of the crystal structure for comparison. (B) AlphaFold model of monomeric wild-type SGSH with missense variants mapped onto the structure. The positions of pathogenic and benign variants are highlighted in red and blue, respectively. The active site residues of SGSH (D31, D32, C70 and D273) are shown as blue mesh, while N274 is shown as red mesh. (C) Detailed view of the SGSH active site from the AlphaFold model, corresponding to the dotted box in (B). The structure is depicted in cartoon representation; the figure was generated using PyMOL software.

The reaction was terminated by the addition of 200 μL stop buffer (0.5 M Na_2CO_3, pH 10.7, 0.025 % Triton X-100), and 50 μL was transferred to a 384-well plate. Fluorescence of 4-MU was measured by a Varioskan LUX instrument (Thermo Fisher Scientific, Waltham, MA, USA) at room temperature with excitation at 360 nm (bandwidth 12 nm), emission at 445 nm, and a measurement time of 100 ms.

2.14. Cycloheximide chase assay

To assess protein stability, a cycloheximide chase assay was performed on HEK293-WT and HEK293-Q365P cells. Cells were collected at 2, 4, and 8 hours after the addition of 100 μg/mL of cycloheximide (CHX) to the cell culture. Western blot analysis was subsequently conducted using 10 μg of cell lysate proteins, probed with an HRP-conjugated anti-FLAG mouse monoclonal antibody.

2.15. Trypsin sensitivity assay

Wild-type and variant SGSH proteins, whose concentrations were quantified using DYKDDDDK-BAP (FUJIFILM Wako Pure Chemical Corporation, Osaka, Japan) as a standard, were diluted to 3 μg/μL in TBS (pH 7.5) (25 mM Tris-HCl, 2.7 mM KCl, 137 mM NaCl), and mixed with 10 ng/μL of sequencing-grade modified trypsin (Promega, Madison, WI, USA) in an 8:1 ratio. Control samples were prepared using the same buffer without trypsin. The reaction was incubated at 37°C for 1 h and stopped by adding 10 mM PMSF to a final concentration of 1 mM. After SDS–PAGE, the gels were stained using Lumitein Protein Gel Stain (Biotium, Fremont, CA, USA), and the bands were quantified using ImageJ.

2.16. Statistical analysis

Statistical evaluation of differences between the groups was carried out by Student’s *t*-test or Tukey-Kramer test, implemented in Microsoft Excel. Differences were considered to be significant at a *p*-value of less than 0.05.

3. Results and discussion

3.1. Comparison of pathogenic and benign variants from the ClinVar dataset

Our previous prediction tool was based on the mutation energy and nSASA of 70 pathogenic and 16 benign missense variants of ARSL, ALG1 and MPI [3]. To improve the accuracy of the prediction, we expanded the dataset to 536 missense variants extracted from the ClinVar database, comprising 296 pathogenic and 240 benign variants (Supplemental Table S1), all of which had a review status classified as “Expert panel” (three stars). We chose this dataset to ensure both the reliability and feasibility of our analyses. It should be noted that the trained dataset predominantly includes variants of BRCA1 and MLH1 (Supplemental Fig. S1), which together comprise nearly half of the dataset, limiting overall gene representation. Further studies will be needed to expand and diversify the dataset for future development and variation.

We first investigated the correlation between pathogenicity and the predicted local distance difference test (pLDDT) score, which reflects the per-residue structural accuracy provided by the AlphaFold structure database [4]. The pLDDT scores of the 296 pathogenic and 240 benign variants were categorized into four confidence levels: very low ($pLDDT<50$), low ($50 <pLDDT<70$), confident ($70 <pLDDT<90$), and

Table 4
Prediction outcomes for published SGSH variants using VarMeter, VarMeter2, AlphaMissense and CADD [a].

SGSH variants (AA change)	Var Meter	Var Meter2	Alpha Missense	CADD	Reference
Pathogenic ($n = 24$)					
A30P	P	P	P	P	[38]
N42K	P	P	P	B	[39]
S66W	PD	P	P	P	[40]
R74C	P	P	P	P	[41]
G90R	PD	P	P	P	[42]
F101S[b]	PD	P	P	P	[43]
S106R	P	P	P	B	[44]
G191R	P	P	P	P	[44]
D235N	PD	P	P	P	[45]
R245H	P	P	P	P	[40]
N274D	P	P	P	A	[46]
E292K	PD	P	P	P	[47]
P293S	P	P	P	A	[39]
S298P	PD	P	P	P	[42]
E300V	B	B	B	B	[48]
Q307P	PD	P	A	P	[48]
S347F	PD	P	P	P	[49]
R377H	P	P	P	P	[41]
R377L	P	P	P	P	[38]
R377C	P	P	P	P	[50]
Q380R	PD	P	B	B	[41]
L411R	P	P	P	P	[51]
R433W	PD	P	P	P	[45]
D444G	PD	P	P	P	[49]
Benign ($n = 8$)					
R55C	PD	B	B	A	
A234G	PD	B	B	B	[41]
G251A	PD	P	A	P	[52]
V361I	P	P	B	B	
M372I	P	P	P	P	
V387M	PD	P	A	B	[53]
M394I[b]	B	B	A	A	
R456H	B	B	B	B	

[a] The classifications are denoted as follows: P (pathogenic), PD (possibly damaging), B (benign), and A (ambiguous). 3D structural parameters, allele frequencies, and prediction scores are shown in Supplemental Table S2.

[b] Located at dimer interface.

Table 5
Comparison of prediction accuracy (%) for SGSH variants by VarMeter, VarMeter2, AlphaMissense and CADD [a].

	Pathogenic (P + PD)	Benign (B)	Total
VarMeter	96	25	78
VarMeter2	96	50	84
AlphaMissense[b]	91	80	89
CADD[b]	82	67	79

[a] Abbreviations: P, pathogenic (damaging); PD, possibly damaging; B, benign.

[b] Ambiguous variants were not included in the calculation of prediction accuracy.

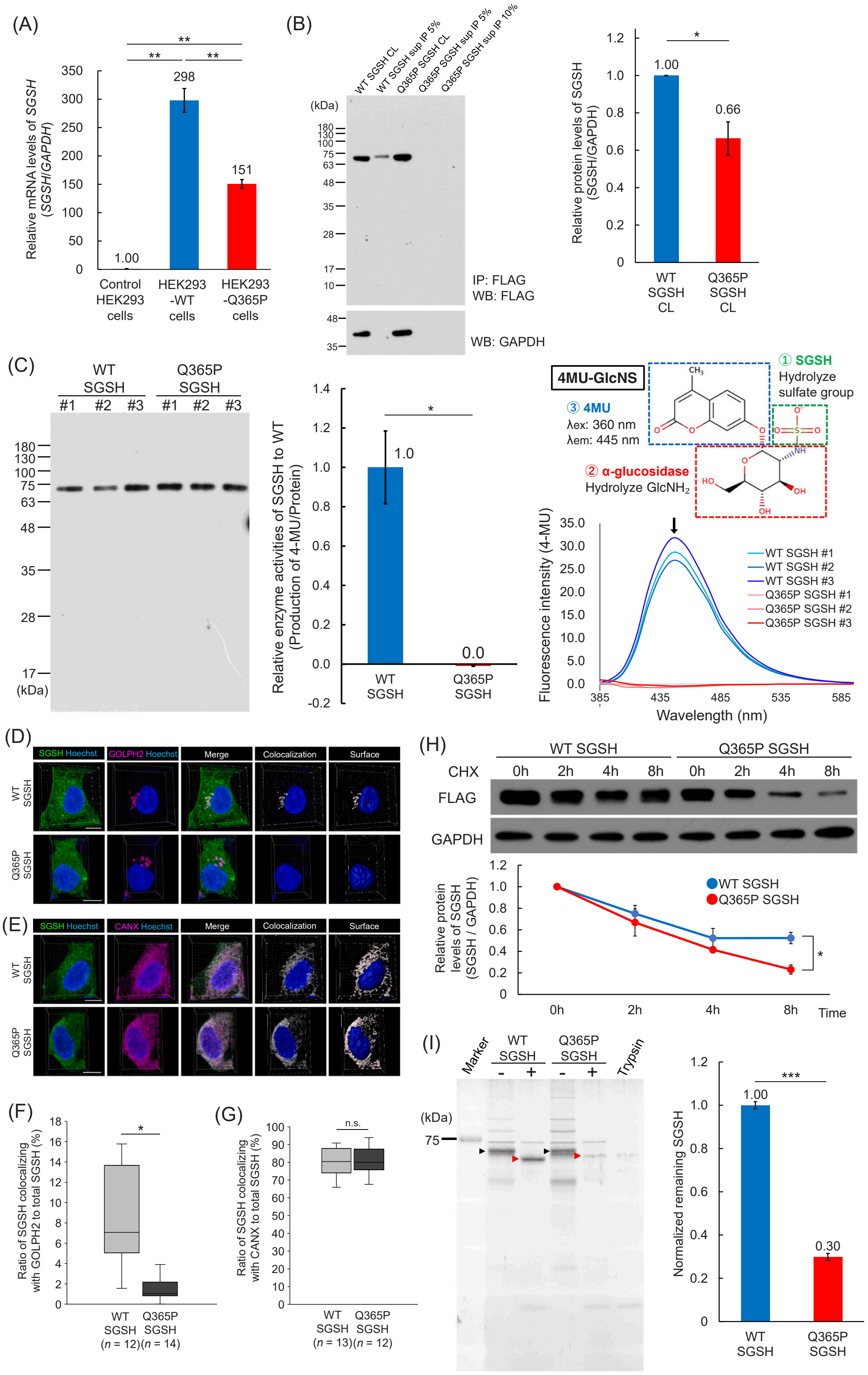
very high (pLDDT>90). Whereas the majority of pathogenic variants had high pLDDT values (pLDDT>70), a significant proportion of benign variants had lower pLDDT values (pLDDT<50) and only a small proportion had high pLDDT values (pLDDT>70) (Fig. 1A).

Intrinsically disordered regions are frequently associated with low pLDDT scores [20–22], which suggests that the mutations in pathogenic variants are predominantly located in well-structured (folded) regions of proteins. By contrast, those in benign variants tend to occur in unstructured regions, and are present to a lesser extent in folded regions. This trend highlights the importance of considering protein folding status when evaluating the potential pathogenicity of missense variants.

We then analyzed the correlation between pathogenicity and allele frequency for the 296 pathogenic and 240 benign variants. The resulting histogram revealed that most pathogenic variants tend to have lower allele frequencies compared with benign variants (Fig. 1B), consistent with previous findings that lower allele frequencies are strongly associated with pathogenicity [23,24]; however, there is an overlap in allele frequencies between pathogenic and benign variants within the range of 0.000001–0.0001. While allele frequency can be a valuable parameter for distinguishing pathogenic variants from benign ones, this overlap highlights the need for careful interpretation.

Next, we calculated mutation energy and nSASA for the 240 benign and 296 pathogenic variants. The mean ± SD pLDDT for benign variants was 53.1 ± 27.9, while that for pathogenic variants was significantly higher at 88.2 ± 14.2 (Table 1). The mean nSASA for benign variants was 0.64 ± 0.29 as compared with 0.21 ± 0.28 for pathogenic variants. A negative correlation was observed between pLDDT and nSASA for both benign and pathogenic variants (Fig. 2). Variants with low pLDDT scores (20−40) tended to have high nSASA values (0.6–1.0), indicating that the mutated residues are generally solvent-exposed, while those with high pLDDT scores (80−100) tended to have low nSASA values (0–0.4), reflecting the location of the mutations in more structured, buried region. This result is consistent with previous studies reporting that pathogenic variants are more buried than benign variants [25,26].

The correlation between mutation energy and pLDDT was analyzed for both benign and pathogenic variants (Fig. 2, upper and lower middle panels). On average, benign variants exhibited a mutation energy of 0.1 ± 2.0 kcal/mol, while pathogenic variants displayed a significantly higher average of 3.9 ± 9.6 kcal/mol (Table 1). This trend confirmed that pathogenic variants generally show greater destabilization of protein structure, reflected by higher mutation energy, and is consistent with the findings of previous studies that used using Gibbs free energy ($\Delta\Delta G$) to distinguish between benign and pathogenic variants [27–30].



(caption on next page)

Fig. 6. Enzymatic activity of the wild-type and Q365P SGSH proteins. (A) *SGSH* mRNA levels in HEK293-WT and HEK293-Q365P cells stably expressing the respective wild-type and Q365P SGSH-FLAG fusion proteins, were analyzed by real-time PCR and normalized to *GAPDH* mRNA in the same sample. Expression levels are shown relative to those of control HEK293 cells (1.0) in which an empty vector was transfected. $^{**}p < 0.001$ by one-way ANOVA followed by Tukey-Kramer test ($n = 3$). (B) Western blot analysis of wild-type (WT) and Q365P SGSH FLAG fusion proteins in the culture supernatant (sup) and cell lysate (CL) from HEK293-WT cells and HEK293-Q365P cells using anti-FLAG. For cell lysates, 10 µg of total protein was applied; for supernatants, 5 % of WT SGSH-FLAG or 5 % or 10 % of immuno-precipitated Q365P SGSH-FLAG was applied. (left) Representative Western blots. (right) Quantification of the protein bands in each cell lysate. Protein levels are shown relative to those of HEK293-WT cells (1.0). $^{*}p < 0.05$ by Student's *t*-test ($n = 3$). (C) Activity of the Q365P SGSH protein. (left) Western blot analysis using anti-FLAG of purified wild-type (WT) and Q365P SGSH FLAG fusion proteins used in the activity assay ($n = 3$). (middle) 4-MU production by Q365P SGSH is normalized by the protein amount used and shown relative to that of WT SGSH. Ratios are given as mean ± S.E. of three independent experiments. $^{*}p < 0.05$ by Student's *t*-test. (right top) Scheme of activity assay. Production of 4-MU after SGSH enzymatic reaction (16 h) using 4MU-GlcNS as a substrate, followed by α-glucosidase reaction (24 h). (right bottom) The fluorescence spectra of each sample are shown. The amount of 4-MU was determined from the fluorescence intensity at 445 nm. (D and E) Three-dimensional confocal images of the intracellular localization of wild-type (WT) and Q365P SGSH-FLAG fusion proteins in HEK293-WT cells and HEK293-Q365P cells, respectively. SGSH (green) and nucleus (blue) are labeled with anti-FLAG and Hoechst 33258, respectively. Cis-Golgi body (magenta in D) and endoplasmic reticulum (ER; magenta in E) are labeled with anti-GOLPH2 and anti-CANX, respectively. Colocalization of SGSH/GOLPH2 (D) or SGSH/CANX (E) is indicated in white in the second panels from the right. Surface rendering models of the colocalization image are shown in the rightmost panels. Scale bar: 10 µm. (F and G) Quantitative analysis of SGSH colocalization with Golgi and ER markers. The ratio of SGSH colocalizing with GOLPH2 (F) or CANX (G) to total SGSH in HEK293-WT cells and HEK293-Q365P cells is shown. $^{**}p < 0.001$, n.s.: not significant by Student's *t*-test. (H) CHX chase assay performed in HEK293-WT and HEK293-Q365P cells. Cells were incubated for 0 hours (untreated), 2, 4, and 8 hours (h) with 100 µg/mL of CHX. (Upper) Representative Western blot images of wild-type (WT) and Q365P SGSH-FLAG fusion proteins detected using anti-FLAG antibody. (Lower) Quantification of protein bands normalized to GAPDH and expressed relative to the 0-hour sample. $^{*}p < 0.05$ by Student's *t*-test ($n = 3$). (I) Stability of the wild-type (WT) and Q365P SGSH proteins. (left) SDS–PAGE analysis of SGSH proteins treated with trypsin. Black arrowhead indicates full-length SGSH; red arrowhead indicates the trypsin-resistant fragment. (right) Proportion of trypsin-resistant fragment remaining after trypsin digestion, calculated as the intensity of the trypsin digestion-resistant band relative to that of full-length SGSH in untreated samples. The values shown have been normalized to WT SGSH. $^{***}p < 0.0001$ by Student's *t*-test ($n = 3$).

However, a subset of pathogenic variants showed low mutation energy (Fig. 2, lower middle). One possible explanation is that the mutations in these variants occur at functionally critical sites, such as residues involved in ligand or protein binding, where even minor structural changes might impair function. Assessing this phenomenon in future studies will be important for further refining pathogenicity prediction models.

As observed in our previous study [3], there was a negative correlation between nSASA and mutation energy (Fig. 2, right panels). In benign variants, high nSASA values were typically associated with low mutation energy (Fig. 2, upper right); in pathogenic variants, conversely, low nSASA values tended to be connected with relatively higher mutation energy (Fig. 2, lower right). This suggests that pathogenic mutations are more likely to occur in buried, structurally important regions, where even small changes may have a significant destabilizing effect.

3.2. Classification of missense variants based on Mahalanobis distance

To classify pathogenic and benign variants using mutation energy, nSASA and pLDDT, we applied Mahalanobis distance [31], a statistical method used in various applications to determine members of groups. First, the mean, variance and covariance of the three variables were calculated for the pathogenic and benign variants (Table 1). These parameters were then used to calculate Mahalanobis distance for both pathogenic and benign variants (Supplemental Table S1 and Table 2). These Mahalanobis distances (D_P and D_B) were used to update the VarMeter prediction method as VarMeter2, integrating mutation energy, nSASA, and pLDDT scores into a unified framework (Fig. 3). To evaluate the performance of VarMeter2, defined by Condition (ii) below, we applied it to the ClinVar dataset alongside the original VarMeter [3], AlphaMissense [10], and CADD [11] methods, defined by Conditions (i), (iii) and (iv), respectively.

Condition (i), original VarMeter [3]:
Requirement #1: nSASA ≤ 0.11
Requirement #2: mutation energy ≥ 0.88 kcal/mol

- Damaging (pathogenic) variant (D): satisfying both requirements #1 and #2
- Possibly damaging (pathogenic) (PD): satisfying either requirement #1 or #2
- Benign variant (B): satisfying neither requirement #1 or #2

Condition (ii), VarMeter2 (this study):

- Pathogenic variant (P): $D_P^2 < D_B^2$
- Benign variant (B): $D_P^2 > D_B^2$

Condition (iii), AlphaMissense [10]:

- Pathogenic variant (P): 0.564 < AM score < 1
- Ambiguous variant (A): 0.340 < AM score < 0.564
- Benign variant (B): 0 < AM score < 0.340

Condition (iv), CADD [11,32]:

- Pathogenic variant (P): CADD score ≥ 25.3
- Ambiguous variant (A): 22.7 < CADD score < 25.3
- Benign variant (B): CADD score ≤ 22.7

The prediction accuracies are summarized in Table 3. VarMeter2 showed 85 % prediction accuracy for pathogenic variants and 78 % accuracy for benign variants, yielding an overall accuracy of 82 %. This represents a significant improvement over the prediction accuracy of the original VarMeter (74 %). However, AlphaMissense [10], a deep learning approach adapted from AlphaFold [4], outperformed VarMeter2 with an overall accuracy of 91 % for the ClinVar dataset. CADD showed a similar accuracy of 91 %.

To evaluate whether prediction accuracy depends on allele frequency, we plotted accuracy against allele frequency (Fig. 4). The results indicated that the prediction accuracy of VarMeter2 is not significantly influenced by the allele frequency, probably because VarMeter2 relies solely on 3D structural parameters and does not incorporate allele frequency into calculations. In contrast, CADD demonstrated higher prediction accuracy for variants with lower allele frequencies, probably because it incorporates annotations reflecting population-level metrics as part of its scoring framework [33]. Despite these advances, challenges remain in 3D structure-based prediction, including the refinement of variant protein models, more precise mutation energy calculations ($\Delta\Delta G$) and optimizing structural parameters for greater predictive power. Further developments in these areas will be essential for improving pathogenicity predictions.

3.3. Evaluation of VarMeter2 using SGSH variant data

To validate VarMeter2 further, we applied it to published data on variants of SGSH, an enzyme responsible for catalyzing the conversion of *N*-sulfo-D-glucosamine into D-glucosamine during the degradation of heparan sulfate. This enzyme was selected due to the large number of reported pathogenic variants linked to Sanfilippo syndrome A. The crystal structure of glycosylated SGSH has been reported at 2 Å resolution (Fig. 5A) [34]; the root mean square deviation (RMSD) of Cα atoms between the crystal structure and the corresponding AlphaFold model is 0.2 Å, indicating that the AlphaFold model of SGSH is sufficiently accurate for calculating mutation energy and solvent-accessible surface area. We therefore mapped the positions of the pathogenic variants onto the 3D structural model of SGSH (Figs. 5B, 5C).

First, we calculated the mutation energy, nSASA and pLDDT for the pathogenic ($n = 24$) and benign ($n = 8$) SGSH variants (Table 4). VarMeter2 was then used to classify the SGSH variants as pathogenic or benign. For comparison, we also carried out predictions using the original VarMeter, AlphaMissense and CADD methods (Table 4). The overall accuracy of VarMeter2 (84 %) to predict variants of SGSH was higher than either the original VarMeter (78 %) or CADD (79 %), and lower than AlphaMissense (89 %) (Table 5). In addition, the accuracy of VarMeter2 to predict pathogenic variants was 96 %. This evaluation demonstrates VarMeter2's potential to accurately classify SGSH variants, offering an improved approach over previous prediction methods.

3.4. Characterization of a novel SGSH variant

To test the ability of VarMeter2 to identify novel pathogenic variants, we applied it to the prediction of SGSH variants identified in our in-house database of 900 individuals with rare or undiagnosed diseases. Among the missense variants in the database, the Q365P variant was newly identified and predicted to be pathogenic by VarMeter2 with the following parameters: D_P^2, 0.7; D_B^2, 14.8; nSASA, 0.03; mutation energy, 7.2 kcal/mol; and pLDDT, 98.89. Additionally, the AlphaMissense score for this variant was 0.90, similarly predicting it as pathogenic. Based on the clinical symptoms observed in the patient, we suspected that this variant might be pathogenic. Therefore, to determine whether this newly identified Q365P mutation directly contributes to the observed symptoms and to support diagnosis, we assessed the activity and expression of the Q365P SGSH protein through detailed experimental studies.

The wild-type and Q365P SGSH proteins were expressed, purified, and subjected to an enzymatic assay using the artificial substrate 4-MU-GlcNS (Fig. 6). First, HEK293-WT cells and HEK293-Q365P cells, stably expressing the respective wild-type and Q365P SGSH-FLAG fusion proteins, were established. The exogenous expression of *SGSH* mRNA was approximately 300 times higher in HEK293-WT cells, and about 150 times higher in HEK293-Q365P cells, as compared with the expression of endogenous *SGSH* mRNA in control HEK293 cells (Fig. 6A). The amount of the Q365P protein in HEK293-Q365P cells was 66 % of that of the wild-type protein in HEK293-WT cells (Fig. 6B). In contrast, the Q365P SGSH-FLAG fusion protein was not detected in the culture medium, whereas the wild-type SGSH-FLAG fusion protein was secreted.

To test the enzymatic activity of the Q365P variant, we purified the wild-type and Q365P SGSH proteins from their respective cell lysates using anti-FLAG magnetic beads. The SGSH enzymatic activity using 4MU-αGlcNS, showed that the Q365P protein had no activity, whereas the wild-type protein had high activity (Fig. 6C).

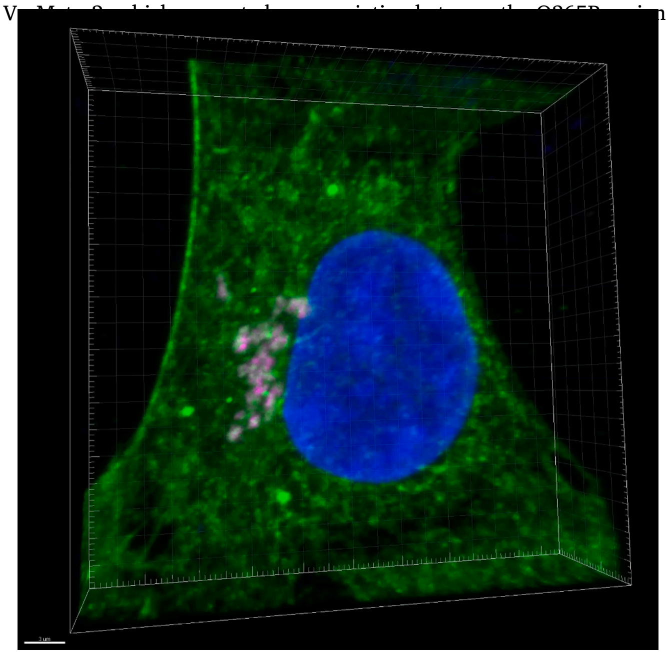
Next, we compared the intracellular localization of the Q365P and wild-type SGSH proteins by immunostaining of the respective HEK293 stable transformants. Whereas wild-type SGSH colocalized with the cis-Golgi marker GOLPH2, the Q365P SGSH protein showed minimal colocalization with this marker (Fig. 6D and Supplemental Movies 1–4), consistent with our above observation that the Q365P SGSH-FLAG fusion protein was not secreted. However, both wild-type and Q365P SGSH proteins colocalized with the endoplasmic reticulum (ER) marker calnexin (CANX) (Fig. 6E and Suppmenental Movies 5–8). Quantitative analysis showed that the ratio of SGSH colocalizing with the Golgi marker GOLPH2 to total SGSH was significantly decreased in HEK293-Q365P cells as compared with HEK293-WT cells (Fig. 6F); similarly, the ratio of GOLPH2 colocalizing with SGSH to total GOLPH2 was significantly lower in HEK293-Q365P cells than in HEK293-WT cells (Supplemental Fig. S2A). In contrast, both the ratio of SGSH colocalizing with the ER marker CANX to total SGSH and that of CANX colocalizing with SGSH to total CANX were comparable between HEK293-WT and HEK293-Q365P cells (Fig. 6G and Supplemental Fig. S2B). Thus, almost all of the Q365P SGSH protein was localized in the ER rather than in the Golgi.

Collectively, these results suggest that the Q365P variant SGSH protein fails to fold into a stable 3D structure, cannot be transferred to the Golgi, and is retained in the ER, which is likely to be the cause of its loss of activity. Moreover, a cycloheximide chase assay showed that the stability of Q365P SGSH protein was decreased as compared with wild-type SGSH (Fig. 6H and Supplemental Fig. S3). Incompletely folded proteins, such as destabilized proteins, are much more sensitive to protease digestion than native proteins [35]. Therefore, we compared the protease resistance of the two purified SGSH proteins. After treatment with trypsin, the intensity of the trypsin-resistant band was significantly reduced for the Q365P protein as compared with wild-type SGSH (Fig. 6I), further confirming that the loss of activity of Q365P SGSH is due to a decrease in protein stability, as predicted by VarMeter2.

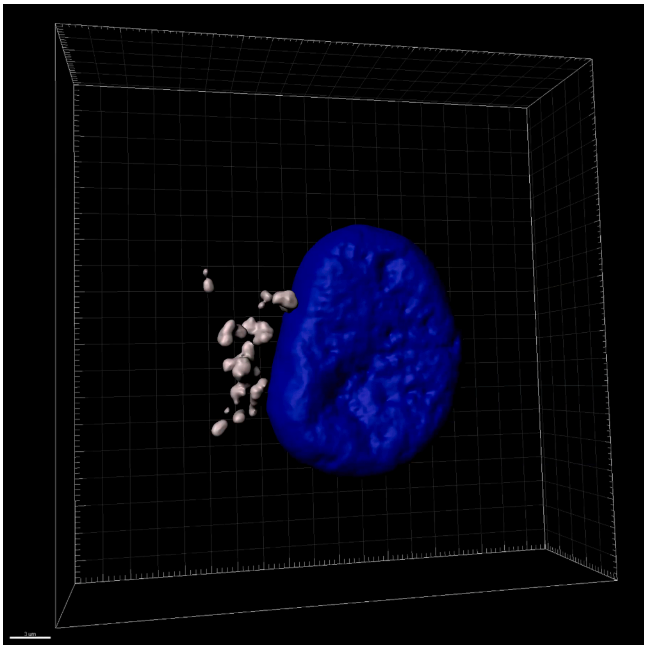
Mutations such as Q365P SGSH, identified here as causing protein destabilization, not only may reduce enzymatic activity and serve as a primary cause of disease but also, as reported in conditions like Congenital Insensitivity to Pain with Anhidrosis (CIPA) [36], may lead to the accumulation of destabilized proteins in the ER, inducing ER stress and further exacerbating disease pathology.

3.5. Clinical manifestation of the patient with SGSH variant Q365P

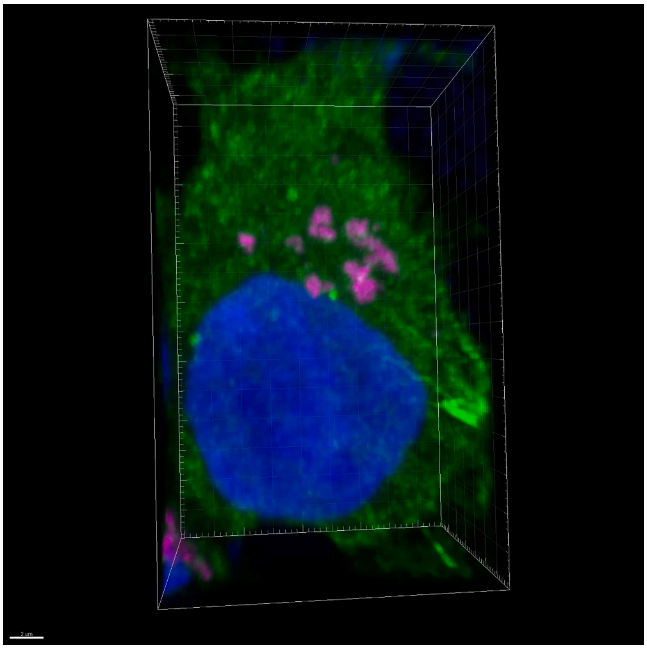
The patient with the Q365P variant of SGSH was diagnosed with Sanfilippo syndrome or mucopolysaccharidosis type III, a rare autosomal recessive lysosomal storage disease caused by impaired degradation of heparan sulfate. This diagnosis aligns with the prediction by VarMeter2, which suggested an association between the Q365P variant



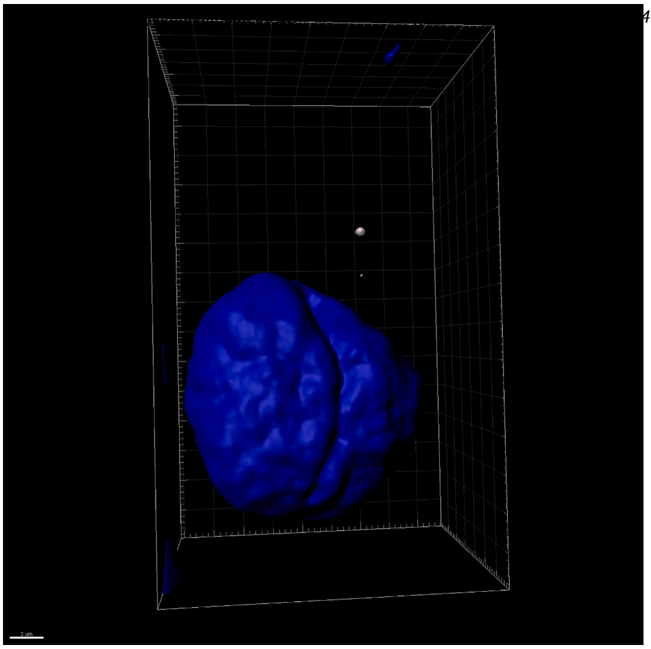
Supplemental Movie S1. Three-dimensional confocal image of a wild-type SGSH-expressing cell with visualization of the cis-Golgi body. SGSH protein, cis-Golgi body, and the nucleus are indicated in green, magenta, and blue, respectively. Movie corresponds to the image in Fig. 6D.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.



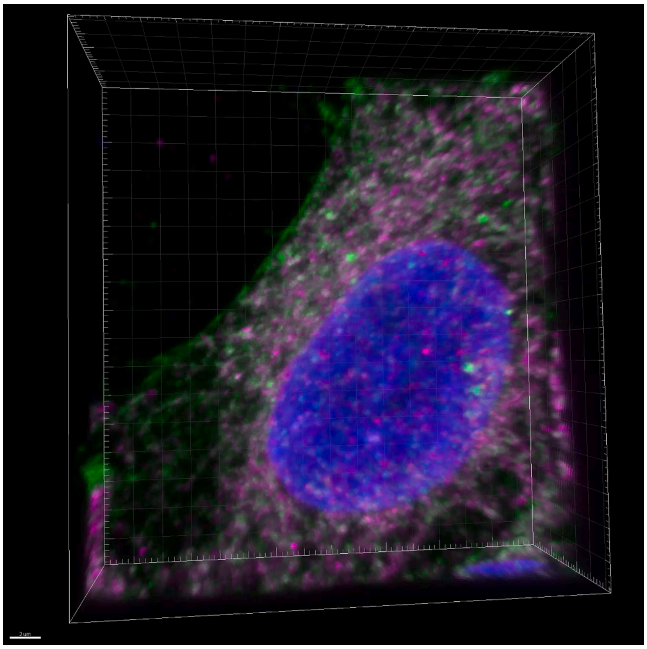
Supplemental Movie S2. Surface rendering model of a wild-type SGSH-expressing cell in which SGSH protein and GOLPH2 are colocalized. Colocalization is indicated in white; the nucleus is indicated in blue. Movie corresponds to the image in Fig. 6D.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.



Supplemental Movie S3. Three-dimensional confocal image of a Q365P SGSH-expressing cell with visualization of the cis-Golgi body. SGSH protein, cis-Golgi body, and the nucleus are indicated in green, magenta, and blue, respectively. Movie corresponds to the image in Fig. 6D.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.



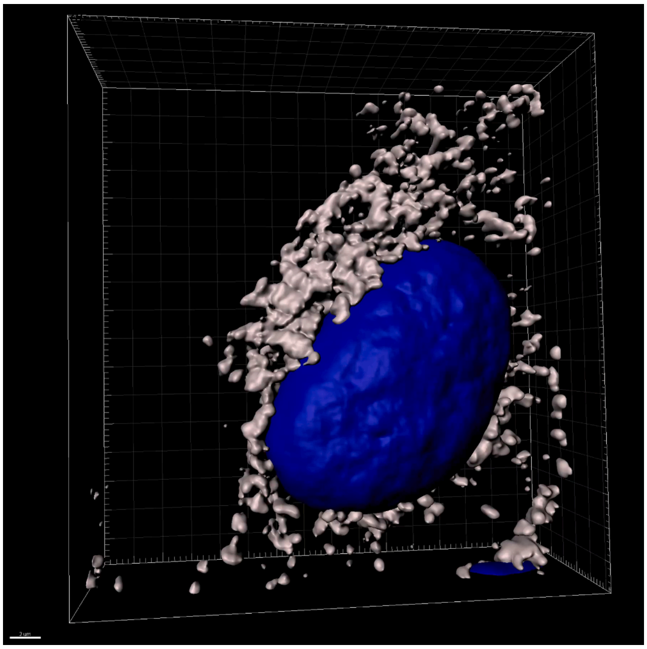
Supplemental Movie S4. Surface rendering model of a Q365P SGSH-expressing cell in which SGSH protein and GOLPH2 are colocalized scarcely. Colocalization is indicated in white; the nucleus is indicated in blue. Movie corresponds to the image in Fig. 6D.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.



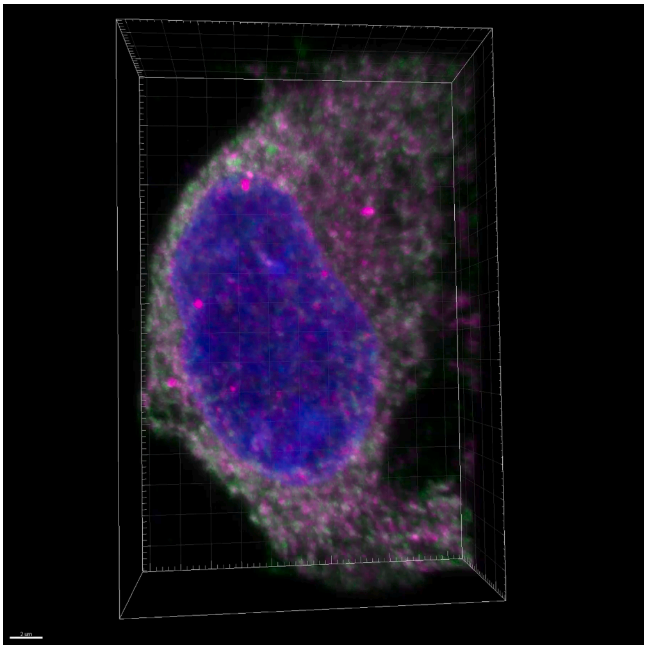
Supplemental Movie S5. Three-dimensional confocal image of a wild-type SGSH-expressing cell with visualization of the ER. SGSH protein, ER, and the nucleus are indicated in green, magenta, and blue, respectively. Movie corresponds to the image in Fig. 6E.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.

and reduced protein stability, contributing to the disease manifestation. Sanfilippo syndrome is characterized by severe degeneration of the central nervous system.

The patient was a 16-year-old female. She was born at 28 weeks of gestation with a birth weight of 1338 g. At the age of 2 years and 1 month, she underwent surgery for a right inguinal hernia, and 5 months later she underwent surgery for a ventral wall scarring hernia. She was noted to have hearing loss, hepatomegaly, hypertrichosis and developmental delay. She also had adenoids, which were removed. At the age of 5 years and 3 months, an increase in urinary mucopolysaccharides and a decrease in heparan *N*-sulfatase activity in the blood were detected. At the age of 11, the homozygous pathogenic variant (p.Q365P) in the

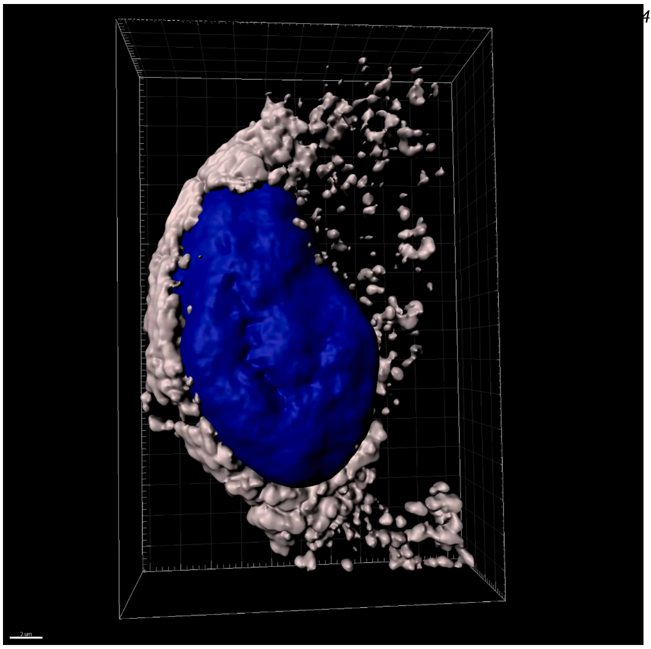


Supplemental Movie S6. Surface rendering model of a wild-type SGSH-expressing cell in which SGSH protein and CANX are colocalized. Colocalization is indicated in white; the nucleus is indicated in blue. Movie corresponds to the image in Fig. 6E.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.



Supplemental Movie S7. Three-dimensional confocal image of a Q365P SGSH-expressing cell with visualization of the ER. SGSH protein, ER, and the nucleus are indicated in green, magenta, and blue, respectively. Movie corresponds to the image in Fig. 6E.. A video clip is available online. Supplementary material related to this article can be found online at doi:10.1016/j.csbj.2025.02.008.

SGSH gene was identified in the patient by trio-whole-exome sequencing analysis. This diagnosis was further supported by VarMeter2, which predicted that the Q365P variant is associated with reduced protein stability, contributing to disease manifestation. She gradually showed regression, and her digital quotient score was less than 10 at the age of 16 years.



Supplemental Movie S8. Surface rendering model of a Q365P SGSH-expressing cell in which SGSH protein and CANX are colocalized. Colocalization is indicated in white; the nucleus is indicated in blue color. Movie corresponds to the image in Fig. 6E.. A video clip is available online. Supplementary material related to this article can be found online at Movie 8doi:10.1016/j.csbj.2025.02.008.

In summary, we have developed VarMeter2, an improved tool for predicting the pathogenicity of missense variants. Utilizing nSASA, mutation energy and pLDDT values, VarMeter2 applies Mahalanobis distance calculations to distinguish pathogenic from benign variants with 82 % accuracy based on the ClinVar dataset. We validated the method with both reported and novel SGSH variants, demonstrating its practical application and confirming its effectiveness. Unlike many other prediction tools, VarMeter2 relies solely on physical parameters derived from 3D structural models, setting it apart from conservation-based tools but also complementing them by focusing on physicochemical changes that may be directly associated with the mechanism of pathogenicity [37]. As a result, VarMeter2 may serve as a complementarily approach, especially for variants that are challenging for conservation-based methods. However, this reliance on structural data also introduces limitations. First, the performance of VarMeter2 is contingent on the accuracy of AlphaFold models, which may be less reliable for disordered regions or proteins with low-confidence structural predictions. Second, the absence of sequence conservation analysis might limit its predictive power for variants in highly conserved regions where evolutionary information is critical. Ongoing efforts are underway to refine the tool further to address these limitations and to enhance its power to identify pathogenic missense variants. VarMeter analyses for missense variants can be conducted collaboratively upon request, and in future we will develop an online tool to further facilitate accessibility.

Funding

This work was supported by AMED (Japan Agency for Medical Research and Development) under Grant Number JP22ek0109446.

CRediT authorship contribution statement

Togayachi Akira: Writing – review & editing, Validation, Supervision. **Aoki-Kinoshita Kiyoko:** Writing – review & editing, Validation, Supervision. **Manabe Noriyoshi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation. **Angata**

Kiyohiko: Writing – review & editing, Validation, Supervision. **Inokuchi Jin-Ichi:** Writing – review & editing, Validation, Supervision. **Kaname Tadashi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation. **Ohno Shiho:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation. **Furukawa Jun-ichi:** Writing – review & editing, Validation, Supervision. **Inamori Kei-ichiro:** Writing – review & editing, Validation, Supervision. **Itoh Kazuyoshi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation. **Nishihara Shoko:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization. **Yabuki Akane:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation. **Yamaguchi Yoshiki:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Ogura Chika:** Writing – review & editing, Writing – original draft, Visualization, Validation, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Eiji Akama (Tohoku Medical and Pharmaceutical University) for his initial analysis of SGSH variants. The authors also thank Dr. Takashi Sato (National Institute of Advanced Industrial Science and Technology (AIST)) for providing the pFLAG-CMV-3-DEST-IRES-puro plasmid. Shoko Nishihara, Kazuyoshi Itoh, Kiyohiko Angata, Akira Togayachi, Kiyoko Aoki-Kinoshita, and Jun-ichi Furukawa are supported by the “Human Glycome Atlas Project” as one of the “Large-Scale Academic Frontier Promotion Projects” of the Ministry of Education, Culture, Sports, Science and Technology (MEXT). GaLSIC and iGCORE are also supported by J-GlycoNet, one of the official Joint Usage/Research Centers of Japan by MEXT.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.02.008.

References

[1] Keskin Karakoyun H, Yuksel SK, Amanoglu I, Naserikhojasteh L, Yesilyurt A, et al. Evaluation of AlphaFold structure-based protein stability prediction on missense variations in cancer. Front Genet 2023;14:1052383.

[2] David A, Sternberg MJE. Protein structure-based evaluation of missense variants: resources, challenges and future directions. Curr Opin Struct Biol 2023;80:102600.

[3] Aoki E, Manabe N, Ohno S, Aoki T, Furukawa J, et al. Predicting the pathogenicity of missense variants based on protein instability to support diagnosis of patients with novel variants of ARSL. Mol Genet Metab Rep 2023;37:101016.

[4] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9.

[5] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 2022;50:D439–44.

[6] Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, et al. ClinVar: improvements to accessing data. Nucleic Acids Res 2020;48:D835–44.

[7] Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants. Am J Hum Genet 2018;103:474–83.

[8] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. Am J Hum Genet 2016;99:877–85.

[9] Andrade F, Aldamiz-Echevarria L, Llarena M, Couce ML. Sanfilippo syndrome: overall review. Pedia Int 2015;57:331–8.

[10] Cheng J, Novati G, Pan J, Bycroft C, Zemgulyte A, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 2023;381: eadg7492.

[11] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019; 47:D886–94.

[12] UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51:D523–31.

[13] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature 2024;625:92–100.

[14] Mitsuhashi N, Toyo-oka L, Katayama T, Kawashima M, Kawashima S, et al. TogoVar: a comprehensive Japanese genetic variation database. Hum Genome Var 2022;9:44.

[15] Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Res 2024;52:D1143–54.

[16] Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:641–56.

[17] Sato T, Sato M, Kiyohara K, Sogabe M, Shikanai T, et al. Molecular cloning and characterization of a novel human β1,3-glucosyltransferase, which is localized at the endoplasmic reticulum and glucosylates *O*-linked fucosylglycan on thrombospondin type 1 repeat domain. Glycobiology 2006;16:1194–206.

[18] Boado RJ, Lu JZ, Hui EK, Pardridge WM. Reduction in brain heparan sulfate with systemic administration of an IgG Trojan Horse-sulfamidase fusion protein in the mucopolysaccharidosis type IIIA mouse. Mol Pharm 2018;15:602–8.

[19] Karpova EA, Voznyi YaV, Keulemans JL, Hoogeveen AT, Winchester B, et al. A fluorimetric enzyme assay for the diagnosis of Sanfilippo disease type A (MPS IIIA). J Inherit Metab Dis 1996;19:278–85.

[20] Wilson CJ, Choy WY, Karttunen M. AlphaFold2: a role for disordered protein/region prediction? Int J Mol Sci 2022;23:4591.

[21] Guo HB, Perminov A, Bekele S, Kedziora G, Farajollahi S, et al. AlphaFold2 models indicate that protein sequence determines both structure and dynamics. Sci Rep 2022;12:10696.

[22] Akdel M, Pires DEV, Pardo EP, Janes J, Zalevsky AO, et al. A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol 2022;29: 1056–67.

[23] Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol 2017;18:77.

[24] Duzkale H, Shen J, McLaughlin H, Alfares A, Kelly MA, et al. A systematic approach to assessing the clinical significance of genetic variants. Clin Genet 2013;84: 453–63.

[25] de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N, et al. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. PLoS Comput Biol 2013;9:e1003382.

[26] Iqbal S, Pérez-Palma E, Jespersen JB, May P, Hoksza D, et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. Proc Natl Acad Sci USA 2020;117:28201–11.

[27] Scheller R, Stein A, Nielsen SV, Marin FI, Gerdes AM, et al. Toward mechanistic models for genotype-phenotype correlations in phenylketonuria using protein stability calculations. Hum Mutat 2019;40:444–57.

[28] Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. eLife 2019;8:e49138.

[29] Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. PLoS Genet 2017;13:e1006739.

[30] Buonfiglio PI, Bruque CD, Lotersztein V, Luce L, Giliberto F, et al. Predicting pathogenicity for novel hearing loss mutations based on genetic and protein structure approaches. Sci Rep 2022;12:301.

[31] Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.

[32] Pejaver V, Byrne AB, Feng BJ, Pagel KA, Mooney SD, et al. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. Am J Hum Genet 2022;109:2163–77.

[33] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5.

[34] Sidhu NS, Schreiber K, Propper K, Becker S, Uson I, et al. Structure of sulfamidase provides insight into the molecular pathology of mucopolysaccharidosis IIIA. Acta Crystallogr D Biol Crystallogr 2014;70:1321–35.

[35] Vestweber D, Schatz G. Point mutations destabilizing a precursor protein enhance its post-translational import into mitochondria. EMBO J 1988;7:1147–51.

[36] Franco ML, Melero C, Sarasola E, Acebo P, Luque A, et al. Mutations in TrkA causing congenital insensitivity to pain with Anhidrosis (CIPA) induce misfolding, aggregation, and mutation-dependent neurodegeneration by dysfunction of the autophagic flux. J Biol Chem 2016;291:21363–74.

[37] 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;14:77.

[38] Pollard LM, Jones JR, Wood TC. Molecular characterization of 355 mucopolysaccharidosis patients reveals 104 novel mutations. J Inherit Metab Dis 2013;36:179–87.

[39] Lee-Chen GJ, Lin SP, Ko MH, Chuang CK, Chen CP, et al. Identification and characterization of mutations underlying Sanfilippo syndrome type A (mucopolysaccharidosis type IIIA). Clin Genet 2002;61:192–7.

[40] Blanch L, Weber B, Guo XH, Scott HS, Hopwood JJ. Molecular defects in Sanfilippo syndrome type A. Hum Mol Genet 1997;6:787–91.

[41] Weber B, Guo XH, Wraith JE, Cooper A, Kleijer WJ, et al. Novel mutations in Sanfilippo A syndrome: implications for enzyme function. Hum Mol Genet 1997;6:1573–9.

[42] Bunge S, Ince H, Steglich C, Kleijer WJ, Beck M, et al. Identification of 16 sulfamidase gene mutations including the common R74C in patients with mucopolysaccharidosis type IIIA (Sanfilippo A). Hum Mutat 1997;10:479–85.

[43] Wen Z, Cheng TL, Yin DZ, Sun SB, Wang Z, et al. Identification of the genetic cause for childhood disintegrative disorder by whole-exome sequencing. Neurosci Bull 2017;33:251–4.

[44] Muschol N, Storch S, Ballhausen D, Beesley C, Westermann JC, et al. Transport, enzymatic activity, and stability of mutant sulfamidase (SGSH) identified in patients with mucopolysaccharidosis type III A. Hum Mutat 2004;23:559–66.

[45] Beesley CE, Young EP, Vellodi A, Winchester BG. Mutational analysis of Sanfilippo syndrome type A (MPS IIIA): identification of 13 novel mutations. J Med Genet 2000;37:704–7.

[46] Knottnerus SJG, Nijmeijer SCM, L IJ, Te Brinke H, van Vlies N, et al. Prediction of phenotypic severity in mucopolysaccharidosis type IIIA. Ann Neurol 2017;82:686–96.

[47] Piotrowska E, Jakóbkiewicz-Banecka J, Tylki-Szymańska A, Czartoryska B, Węgrzyn A, et al. Correlation between severity of mucopolysaccharidoses and combination of the residual enzyme activity and efficiency of glycosaminoglycan synthesis. Acta Paediatr 2009;98:743–9.

[48] Bekri S, Armana G, De Ricaud D, Osenda M, Maire I, et al. Early diagnosis of mucopolysaccharidosis III A with a nonsense mutation and two *de novo* missense mutations in *SGSH* gene. J Inherit Metab Dis 2005;28:601–2.

[49] Miyazaki T, Masuda N, Waragai M, Motoyoshi Y, Kurokawa K, et al. An adult Japanese Sanfilippo A patient with novel compound heterozygous S347F and D444G mutations in the sulphamidase gene. J Neurol Neurosurg Psychiatry 2002;73:777–8.

[50] Di Natale P, Balzano N, Esposito S, Villani GR. Identification of molecular defects in Italian Sanfilippo A patients including 13 novel mutations. Hum Mutat 1998;11:313–20.

[51] Valstar MJ, Neijs S, Bruggenwirth HT, Olmer R, Ruijter GJ, et al. Mucopolysaccharidosis type IIIA: clinical spectrum and genotype-phenotype correlations. Ann Neurol 2010;68:876–87.

[52] Meyer A, Kossow K, Gal A, Steglich C, Mühlhausen C, et al. The mutation p. Ser298Pro in the sulphamidase gene (*SGSH*) is associated with a slowly progressive clinical phenotype in mucopolysaccharidosis type IIIA (Sanfilippo A syndrome). Hum Mutat 2008;29:770.

[53] Di Natale P, Pontarelli G, Villani GR, Di Domenico C. Gene symbol: SGSH. Disease: Sanfilippo type A syndrome, mucopolysaccharidosis IIIA. Hum Genet 2006;119:679.