






Design of a stable human acid- β -glucosidase: towards improved Gaucher disease therapy and mutation classification

Sarka Pokorna^{1,2} , Olga Khersonsky¹, Rosalie Lipsh-Sokolik¹, Adi Goldenzweig¹, Rebekka Nielsen^{1,*}, Yacov Ashani¹, Yoav Peleg³, Tamar Unger³, Shira Albeck³, Orly Dym³, Asa Tirosh³, Rana Tarayra⁴, Michaël Hocquemiller⁵, Ralph Laufer⁵, Shifra Ben-Dor³ , Israel Silman⁶, Joel L. Sussman⁴ , Sarel J. Fleishman¹  and Anthony H. Futerman¹ 

1 Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot, Israel

2 J. Heyrovsky Institute of Physical Chemistry of the Czech Academy of Sciences, Prague, Czech Republic

3 Department of Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, Israel

4 Department of Chemical and Structural Biology, Weizmann Institute of Science, Rehovot, Israel

5 Lysogene, Neuilly-sur-Seine, France

6 Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel

Keywords

Gaucher disease; gene therapy; *in silico* mutation classification; PROSS; Rosetta; SNPs

Correspondence

A. H. Futerman, Department of Biomolecular Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel
 Tel: +972-8-934-2704
 E-mail: tony.futerman@weizmann.ac.il

Present address

*Aarhus University Hospital, Denmark

(Received 24 September 2022, revised 23 January 2023, accepted 20 February 2023)

doi:10.1111/febs.16758

Acid- β -glucosidase (GCase, EC3.2.1.45), the lysosomal enzyme which hydrolyzes the simple glycosphingolipid, glucosylceramide (GlcCer), is encoded by the *GBA1* gene. Biallelic mutations in *GBA1* cause the human inherited metabolic disorder, Gaucher disease (GD), in which GlcCer accumulates, while heterozygous *GBA1* mutations are the highest genetic risk factor for Parkinson's disease (PD). Recombinant GCase (e.g., Cerezyme[®]) is produced for use in enzyme replacement therapy for GD and is largely successful in relieving disease symptoms, except for the neurological symptoms observed in a subset of patients. As a first step toward developing an alternative to the recombinant human enzymes used to treat GD, we applied the PROSS stability-design algorithm to generate GCase variants with enhanced stability. One of the designs, containing 55 mutations compared to wild-type human GCase, exhibits improved secretion and thermal stability. Furthermore, the design has higher enzymatic activity than the clinically used human enzyme when incorporated into an AAV vector, resulting in a larger decrease in the accumulation of lipid substrates in cultured cells. Based on stability-design calculations, we also developed a machine learning-based approach to distinguish benign from deleterious (i.e., disease-causing) *GBA1* mutations. This approach gave remarkably accurate predictions of the enzymatic activity of single-nucleotide polymorphisms in the *GBA1* gene that are not currently associated with GD or PD. This latter approach could be applied to other diseases to determine risk factors in patients carrying rare mutations.

Abbreviations

AAV, adeno-associated virus; ERT, enzyme replacement therapy; GCase, acid- β -glucosidase; GD, Gaucher disease; GlcCer, glucosylceramide; GlcSph, glucosylsphingosine; LSDs, lysosomal storage disorders; PD, Parkinson's disease; SNP, single-nucleotide polymorphism; SRT, substrate reduction therapy.

Introduction

Gaucher disease (GD), one of the two most common lysosomal storage diseases (LSDs) [1,2], is caused by biallelic mutations in the *GBA1* gene (see Table S1 for a comprehensive list of all known *GBA1* missense mutations). *GBA1* encodes the lysosomal enzyme acid- β -glucosidase (GCase, EC3.2.1.45) [3] which hydrolyzes the simple glycosphingolipids, glucosylceramide (GlcCer), and glucosylsphingosine (GlcSph). Malfunction of GCase leads to intracellular accumulation of both lipids, primarily in the lysosomes of macrophages and monocytes [4]. GD is classified into three clinical subtypes. Type 1 is manifested by hepatosplenomegaly, anemia, thrombocytopenia, and bone disease, while GD types 2 and 3 cause severe neurological disease (nGD) [5].

Two approved treatments for GD are currently available, namely enzyme replacement therapy (ERT) and substrate reduction therapy (SRT). Patients treated by ERT receive periodic intravenous infusions of a recombinantly expressed GCase, of which Cerezyme[®] is the most widely used, while SRT uses inhibitors of GlcCer synthesis, thereby reducing its accumulation. However, neither ERT nor SRT can be currently used to treat nGD [6,7]. Similar to other neurological diseases, gene therapy offers an attractive option for the treatment of nGD. Gene delivery mediated by adeno-associated viruses (AAVs) [8] has the advantage of low immunogenicity, high efficiency, and the possibility of targeting specific tissues or cell types, including neurons [9,10]. AAV gene therapy is safe and efficient in mouse models of LSDs, including GD [11], and has been used in preclinical trials on human patients with LSDs (see Ref. [10] and references therein).

GCase comprises 497 amino acids and contains two disulfide bridges and five glycosylation sites, four of which are usually occupied [12]. Despite the success of ERT using recombinant GCase for the treatment of type 1 GD, no attempts have been made to optimize treatment strategies using, for instance, more stable forms of GCase or of other enzymes used in ERT in other LSDs. If such stabilized enzymes were available, they might remain active for longer times, reducing infusion frequency and enhancing therapeutic outcomes and economic benefit.

Due to the marginal stability of many proteins [13], protein engineering is frequently used to improve protein stability, although not, so far, for the enzymes used in ERT or in gene therapy for rare metabolic diseases. One approach to stabilize proteins is the use of computer-based algorithms, such as PROSS, which

combine atomistic Rosetta design calculations and phylogenetic sequence analysis to design stable variants [14,15]. PROSS has been successfully applied to many proteins, including those that have several disulfide bonds and glycosylation sites [14,16–18]. PROSS designs often exhibit higher recombinant expression levels and increased thermal stability while maintaining activity.

Protein destabilization or loss of expression caused by missense mutations can lead to a range of human diseases [19], and the ability to be able to predict the pathogenicity of missense mutations is highly desirable. A number of *in silico* tools for predicting functional and structural consequences of missense mutations are available, with most using sequence and conservation-based methods, protein sequence and structure, or supervised learning methods [20]. Nevertheless, such predictions often disagree, raising questions about their reliability. By way of example, results obtained with seven available *in silico* algorithms using a dataset of 97 nonsynonymous single-nucleotide polymorphisms (nsSNPs) in *GBA1* [21] suggested that 22 should result in GD. However, the limitations of this study can be appreciated since only six of the algorithms recognized L444P, and only three identified N370S as disease-causing mutations, even though they are the two most prominent mutations associated with GD. A more useful approach might be to train an algorithm based on mutations with known pathologies, that is, benign and disease-causing, such as was done successfully for the MLH1 variant in Lynch syndrome [22]. For such an approach to be effective, a sufficient number of known mutations, both benign and disease-causing, should be available.

In the present study, we use the PROSS algorithm to generate a more stable form of GCase. Notably, one of the GCase designs is secreted at a higher level, and upon transduction into neuroblastoma cells using an AAV vector, results in more effective clearance of GlcCer compared with WT GCase. Based on these results, we hypothesized that PROSS could enrich data from clinical studies to train an algorithm to predict the clinical severity of various mutations. We verified these predictions experimentally and by analysis of published clinical data. We conclude that the PROSS-designed GCase may help improve the efficacy of ERT or of gene therapy (at least in the brain, which is an immune-privileged site). Furthermore, predictions of the clinical outcome of additional *GBA1* mutations could be used for diagnostic purposes, with particular relevance to novel *GBA1* mutations in Parkinson's disease (PD), in which *GBA1* mutations are the highest genetic risk factor [23,24].

Results

A stabilized GCcase design for recombinant expression

We used PROSS to design GCcase variants based on its crystal structure (PDB: 3gxi [25]), while not permitting design calculations within the active site pocket. Designs dGCcase1, dGCcase2, and dGCcase3 containing 35, 45, and 55 amino acid substitutions, respectively, were expressed in mammalian HEK293T cells, along with WT human GCcase (plasmids are shown in Fig. 1). The proteins were purified from growth media by one-step affinity chromatography using the Twin-Strep tag. Designs were screened for enzymatic activity and secretion (Fig. 2A). In contrast to WT GCcase, all three designs were secreted, and their enzymatic activity increased with the number of mutations. Additionally, dGCcase3 exhibited activity when expressed and purified in *Escherichia coli*. By contrast, WT GCcase did not display any enzymatic activity even though it could be expressed in *E. coli* (Fig. 2B). We concluded that increasing GCcase stability led to correct protein folding independent of glycosylation (which does not occur in *E. coli*).

Design dGCcase3 carries 55 amino acid mutations compared with WT GCcase. The mutations are distributed across the entire protein, except for the active site which was restricted from the design calculations (Fig. 2C,E). To approximate the impact of the mutations on protein structure, ALPHAFOLD2 [26,27] was used to predict the structure of dGCcase3. Alignment with the crystal structures of human GCcase (PDB: 3gxi and 1ogs) [25,28] revealed very close agreement (< 0.5 Å root mean square deviation; Fig. 3). This modeling suggests that no major structural rearrangements occur in dGCcase3, in agreement with the fact that it retains catalytic activity.

Purified dGCcase3 was assessed for *in vitro* enzyme activity and thermal stability and compared to purified recombinant GCcase (r-GCcase, purchased from Biotest) and to Cerezyme[®]. Cerezyme[®] and dGCcase3 displayed very similar kinetic properties. By contrast, r-GCcase exhibited an ~ 5 -fold lower $k_{\text{cat}}/K_{\text{M}}$, mainly due to a lower reaction rate (k_{cat} ; Table 1). The melting temperature (T_{m}) of dGCcase3 was 67.9 ± 0.7 °C, which is ~ 17 °C and ~ 12 °C higher than Cerezyme[®] and r-GCcase, respectively (Table 1, Fig. 2D). dGCcase3 exhibits almost the same K_{M} values as Cerezyme[®], confirming that the active site is intact despite the 55 designed mutations.

One of the designed mutations, N370D, which is present in all of the PROSS designs, impacts the same

position as one of the most common GD-causing mutations, N370S [29]. Whereas the clinical mutation to Ser results in low enzymatic activity, the Asp mutation maintains both expression and activity levels comparable to those of WT GCcase (Fig. 4). This observation demonstrates that even positions associated with disease-causing mutations can be optimized using a judicious choice of mutation.

Expression of dGCcase3 using AAV exhibits high GCcase activity

Two therapeutic regimes are currently attracting significant interest for neurological forms of LSDs, namely the use of small compounds for SRT that cross the blood–brain barrier [30], and gene therapy using vectors injected directly into the brain [10]. Since the immune response is attenuated in the central nervous system [31], designed proteins could in principle be used in the brain with minimal risk of an immunological response. Therefore, human WT GCcase and dGCcase3 were cloned into an AAVrh10 (adeno-associated virus, serotype rh10) vector and used to transduce $\text{GBA}^{-/-}$ neuroblastoma cells in culture.

Nondifferentiated SH-SY5Y $\text{GBA}^{-/-}$ cells (Fig. 5) displayed increased GCcase activity upon transduction with both vectors in a dose-dependent manner. The activity of cells transduced with AAV-dGCcase3 was two–threefold higher than that of those transduced with AAV-WT GCcase (Fig. 5A). The AAV-dGCcase3-transduced cells [5×10^5 vg per cell (viral genome per cell)] exhibited the same levels of GCcase activity as SH-SY5Y $\text{GBA}^{+/+}$ cells (Fig. 5A). Likewise, when SH-SY5Y $\text{GBA}^{-/-}$ cells were differentiated (to allow them to survive longer in culture) and transduced with 5×10^5 vg per cell of the AAV vectors, the AAV-dGCcase3 transduced cells exhibited significantly higher activity 12 and 15 days post-transduction than cells transduced with AAV-WT GCcase (Fig. 5B). SH-SY5Y $\text{GBA}^{+/+}$ cells contained ~ 400 pmol·mg⁻¹ protein of GlcCer, with GlcCer levels elevated ~ 20 -fold (~ 7000 pmol·mg⁻¹ protein) in SH-SY5Y $\text{GBA}^{-/-}$ cells. Fifteen days post-transduction, AAV-WT GCcase transduced $\text{GBA}^{-/-}$ cells showed reduction of GlcCer levels to ~ 950 pmol·mg⁻¹. An even more effective reduction was obtained upon transduction with AAV-dGCcase3 (~ 450 pmol·mg⁻¹), decreasing GlcCer levels close to those of WT cells (Fig. 5C; Table 2 gives levels of individual GlcCer species with different *N*-acyl chain lengths). Similar results were obtained for the extent of reduction of GlcSph, with GlcSph levels ~ 125 and ~ 20 times lower following AAV-dGCcase3 and AAV-WT GCcase transduction compared with $\text{GBA}^{-/-}$ cells

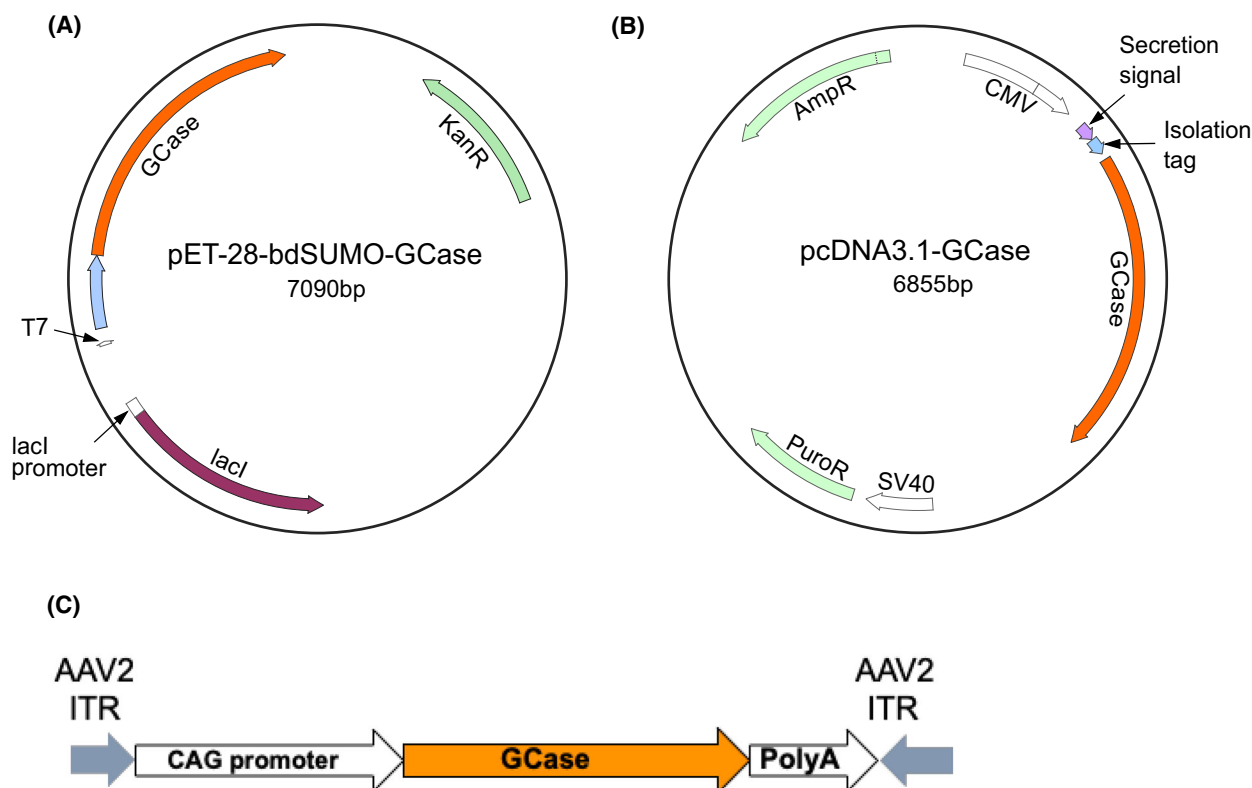


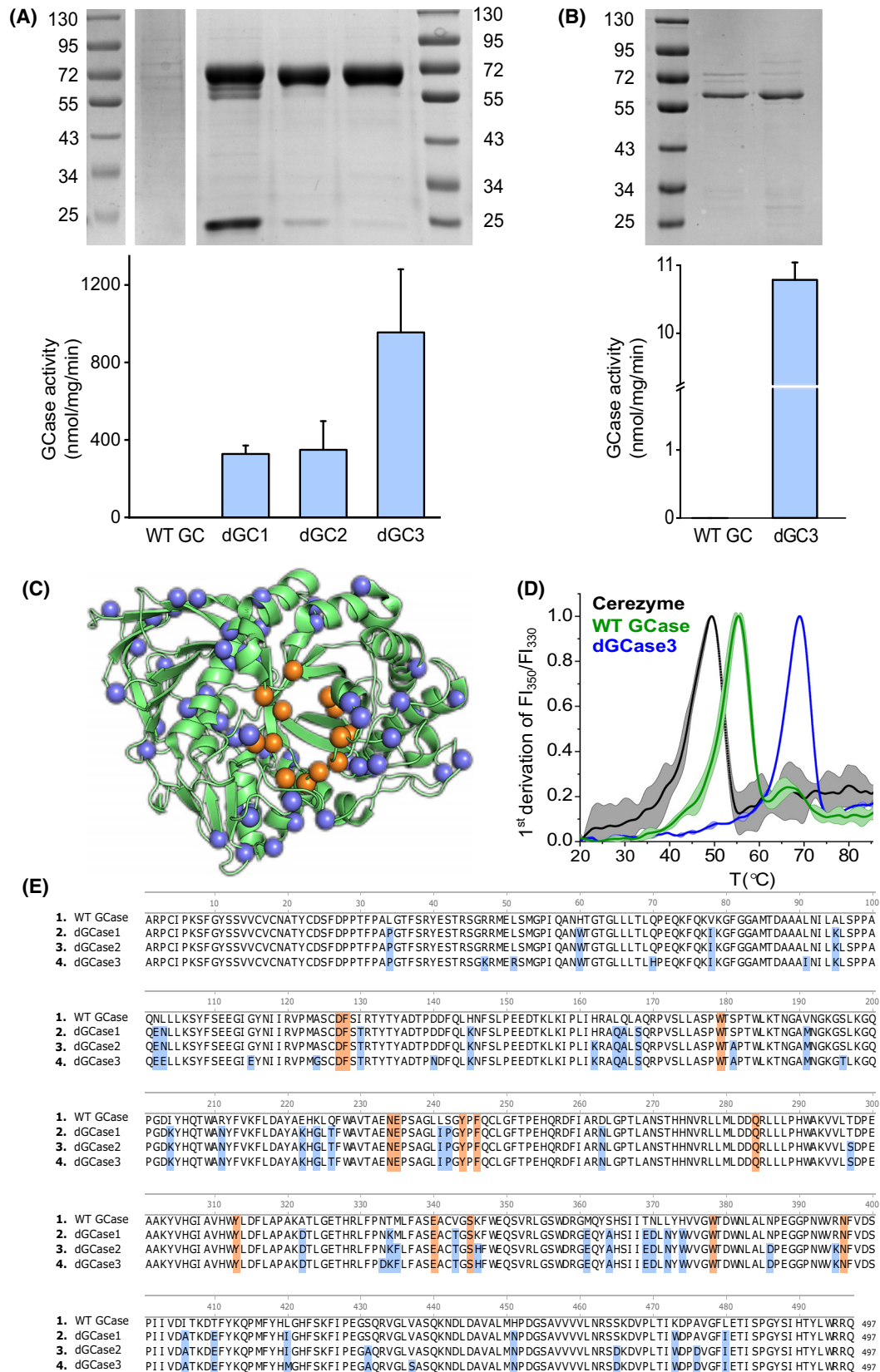
Fig. 1. Vectors used for GCase expression. Schematic representation of vectors used for expression of WT GCase and its PROSS-designed variants. The GCase sequence is in orange and tags at the N terminus, which were used for protein isolation, in blue. (A) pET-28-bdSUMO was used for *E. coli* expression in which the protein was fused to N-terminal poly His14-SUMO (SUMO, blue). (B) A pcDNA3.1 vector was used for expression of GCase in HEK293T cells in which the R-PTP-S secretion signal (purple) was used to target the protein for secretion to the extracellular medium. An N-terminal isolation tag (TwinStrep, blue) was inserted into the vectors immediately following the secretion signal. The sizes of the vectors expressed as base pairs (bp) are shown. Promoters (T7, lacI, CMV, SV40, white) and antibiotic resistance sequences (kanamycin, KanR; ampicillin AmpR, puromycin PuroR, green) are also shown. (C) The WT human GCase and dGCase3 DNA sequences were cloned into AAV2-based recombinant genomes under the control of the CAG promoter.

(Fig. 5C, Table 2). Together, our results suggest that dGCase3 may be a suitable candidate for nGD gene therapy since it is more active in cell culture and clears more GlcCer than the WT enzyme, when using the same dose of AAV.

A machine learning classifier predicts the severity of *GBA1* mutations.

We postulated that the accuracy of the PROSS stability-design algorithm, and the large number of

Fig. 2. Characterization of PROSS-designed GCase variants. WT and PROSS-designed GCase variants were purified from (A) HEK 293T cells and (B) *E. coli*. Levels of expression and secretion are visualized on Coomassie-stained SDS/PAGE gels (upper panels). Mr markers (kDa) are shown. Enzyme activity was assayed using C6-NBD-GlcCer (lower panels). Data are means \pm SD from at least three independent experiments. (C) Structure of WT GCase [28] (green) with the positions of PROSS mutations shown as blue spheres; amino acid residues of the active site (C126, D127, F128, W179, N234, E235, Y244, P245, F246, D283, Q284, H311, Y313, E340, C342, G344, S345, W381, N382, F397, and V398) are depicted in orange. (D) Melting curves (T_m) for Cerezyme (black), dGCase3 (blue) and WT GCase (r-GCase, green) were measured by differential scanning fluorimetry. Derivation of the fluorescence intensity ratio was measured (F_{I330}/F_{I350}), with the peak corresponding to the T_m . Standard deviations are indicated as areas with lighter tones. (E) Amino acid sequences of PROSS-designed variants of GCase (dGCase) together with that of WT GCase. Mutations introduced by PROSS are highlighted in blue. Amino acid residues corresponding to the active site (and restricted from the PROSS mutagenesis) are shown in orange. Protein structure was created in PYMOL (Schrödinger, LLC, New York, NY, USA).



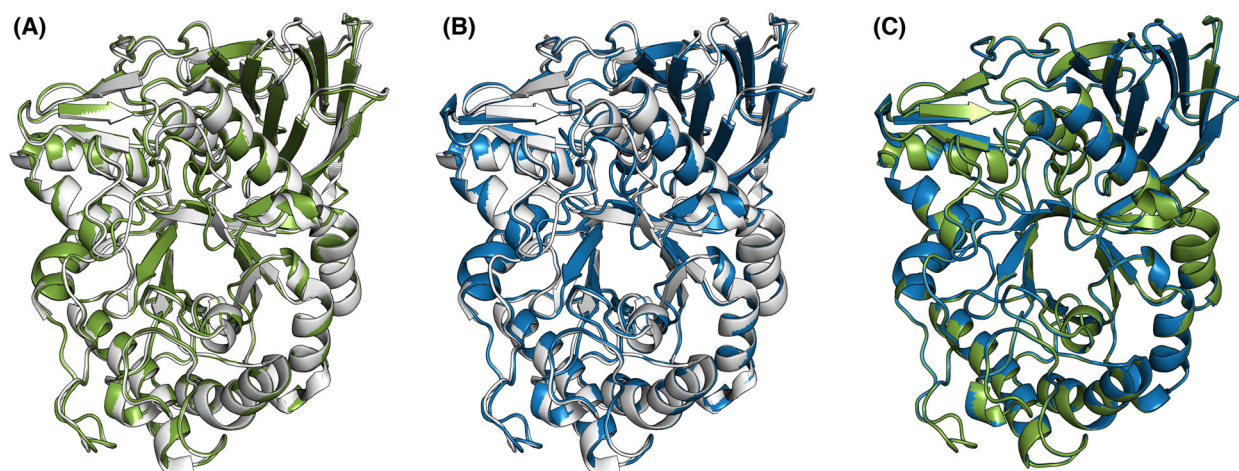


Fig. 3. Alignment of WT and dGCase3 structures. (A, B) Crystal structure of WT human GCase (gray, PDB: 3gxi) aligned with ALPHAFOLD2 structures of human GCase (A, green) and dGCase3 (B, blue). The alignment gave root mean square deviations (RMSDs) of 0.397 and 0.417 Å respectively. (C) Alignment of ALPHAFOLD2 structures of human GCase (green) and dGCase3 (blue), which yielded an RMSD of 0.152 Å. Protein structures were created in PYMOL (Schrödinger, LLC, New York, NY, USA).

	K_M (mM)	k_{cat} (min ⁻¹)	k_{cat}/K_M (min ⁻¹ ·M ⁻¹)	T_m (°C)
dGCase3	0.90 ± 0.23	1340 ± 546	1.49 ± 0.41 × 10 ⁶	67.9 ± 0.7
r-GCase	1.20 ± 0.39	337 ± 33	0.29 ± 0.06 × 10 ⁶	55.1 ± 0.5
Cerezyme®	0.70 ± 0.28	1071 ± 26	1.54 ± 0.33 × 10 ⁶	50.6 ± 2.2

Table 1. *In vitro* characterization of dGCase3. Data were collected from at least three independent experiments and are means ± SD.

stabilizing mutations that it predicts, could augment the limited clinical data on benign mutations and lead to improved discrimination of disease-causing mutations. As a striking example for the paucity of clinical data on benign mutations, only three missense mutations in *GBA1* have been classified to date as benign or likely benign (<https://www.ncbi.nlm.nih.gov/variation/view/>). The 226 GD-causing missense mutations (Table S1) and the 55 PROSS-designed mutations in dGCase3 (Table S2) were combined to train an algorithm to predict the clinical effect of unknown *GBA1* SNPs (Table S3). The analysis is based on the premise that PROSS mutations are individually neutral or stabilizing and do not impact enzyme activity. Three parameters were calculated for each mutation: (a) the change in conservation score between the WT amino acid and the mutated amino acid (Δ PSSM; based on the position-specific scoring matrix computed by PROSS); (b) the change in protein energy due to the mutation ($\Delta\Delta G$; also computed by PROSS); and (c) the exposure of the amino acid position to solvent (calculated by the Stride webserver). The best separation between mutations introduced by PROSS and the disease-causing mutations was obtained using Δ PSSM, followed by $\Delta\Delta G$ (Fig. 6A). Solvent exposure did not show a significant separation between the GD-causing

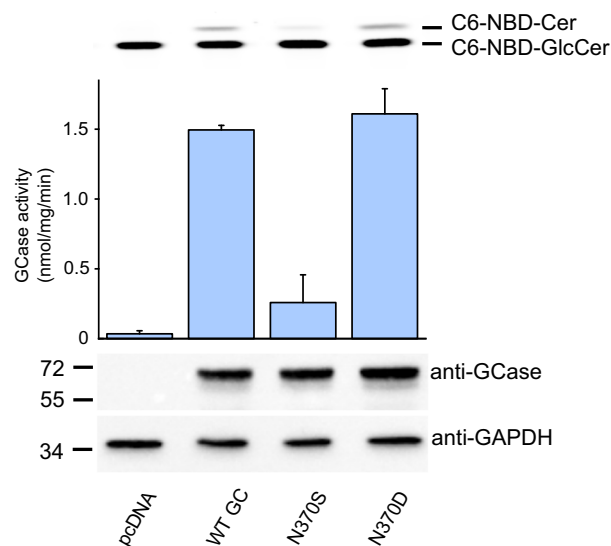


Fig. 4. Expression and activity of N370S and N370D. Activity of WT GCase (WT GC), N370S and N370D expressed in HEK293T *GBA*^{-/-} cells compared to activity in cells transfected with a control pcDNA vector. Activity was assayed using C6-NDB-GlcCer in cell homogenates. A representative TLC plate is shown, with the product (C6-NDB-Cer) and the substrate (C6-NDB-GlcCer) indicated. GCase expression was determined by western blotting using an anti-GCase antibody. Mr markers are shown. GAPDH served as a loading control. Data are means ± SD, *n* = 3.

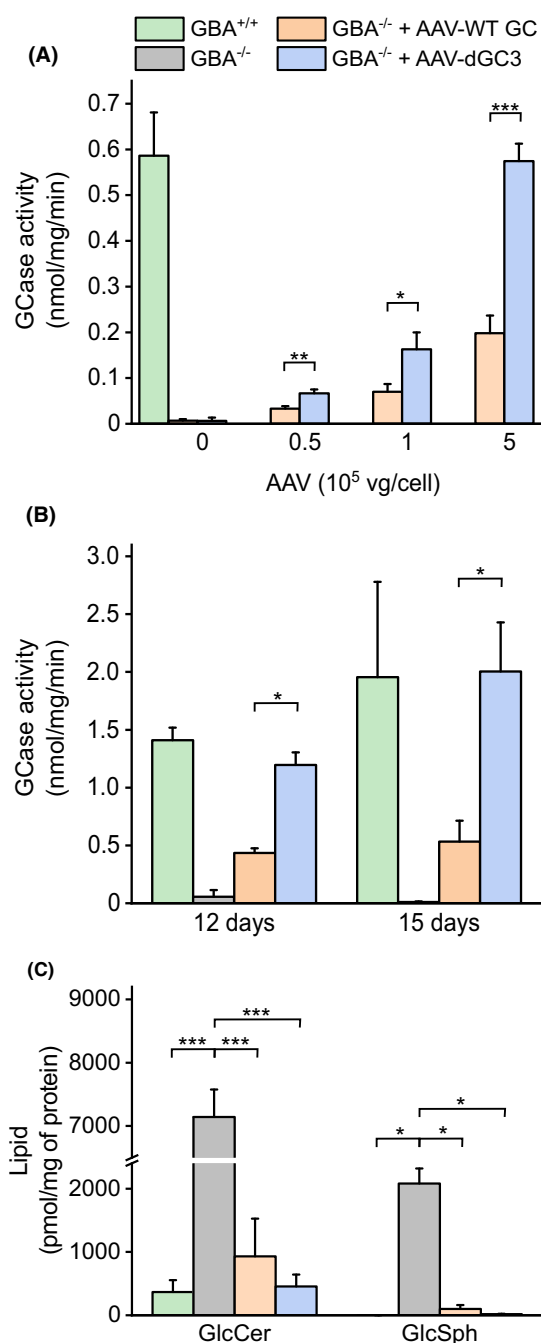


Fig. 5. GCase activity and GlcCer/GlcSph levels in SH-SY5Y GBA^{-/-} cells transduced with AAV. Cells were transduced using AAV coding for WT GCase (AAV-WT GC, orange) and dGCase3 (AAV-dGC3, blue), and compared with nontransduced SH-SY5Y GBA^{-/-} cells (gray) and wild-type SH-SY5Y GBA^{+/+} cells (green). *In vitro* enzyme activity was assayed on cell homogenates of (A) nondifferentiated cells, 5 days post-AAV transduction using several AAV doses and (B) differentiated SH-SY5Y cells, 12 and 15 days post-AAV transduction, using a dose of 5 × 10⁵ vg per cell. (C) LC-ESI-MS/MS analysis of GlcCer and GlcSph levels in differentiated SH-SY5Y cells harvested 15 days post-AAV transduction (5 × 10⁵ vg per cell). Data are means ± SD from at least three independent experiments. Statistical significance was determined using the Student's *t*-test. **P* < 0.05, ***P* < 0.01, ****P* < 0.005. Further data relating to lipid levels are given in Table 2.

A set of SNPs in *GBA1* has been documented (Table S3), although none of them have been detected in GD patients to date. Using the trained PRAMP (PROSS-based Algorithm for Mutation Prediction) classifier, we analyzed this set and separated the SNPs into putatively deleterious and benign mutations (Fig. 6C). In addition, each SNP was assigned a score (PRAMP score), determined by its distance from the separation line, with benign and harmful SNPs assigned a positive and negative score, respectively. Twenty-eight clones of GCase, bearing individual SNPs spanning the PRAMP score range, were expressed in HEK293T GBA^{-/-} cells, and their *in vitro* activity was determined (Fig. 6D, Table S3). A clear correlation between the PRAMP score and GCase activity was obtained, as seen by the Spearman coefficient of 0.8. Thus, even though GD can be caused by factors other than defective enzymatic activity (such as defective lysosomal trafficking [32]), the PRAMP score developed herein gives a remarkably good correlation with enzymatic activity.

Many other prediction tools have been developed to attempt to distinguish between disease-causing and benign mutations [20]. We compared the results of our algorithm to REVEL (<https://sites.google.com/site/revelgenomics/>), a missense mutation classifier that is based on an ensemble of 13 individual *in silico* tools [33]. Precomputed REVEL scores, in a range of 0 to 1 with 1 being the most severe, for the same *GBA1* SNP missense mutations also correlated with enzymatic activity (Fig. 6E), but with a somewhat lower Spearman coefficient (−0.7) than obtained for PRAMP. Together, these results demonstrate that the PRAMP algorithm can accurately classify missense mutations. The observation that our stability-based analysis is at least as powerful as the much more sophisticated scheme employed by REVEL, highlights the importance of stability and expression in understanding the effect of mutations in *GBA1*.

and PROSS mutations. Next, we used Δ PSSM and $\Delta\Delta G$ to train a linear support-vector machine to predict whether a particular *GBA1* mutation is likely benign or deleterious (Fig. 6B). Out of 281 mutations (226 GD-causing and 55 PROSS), only five mutations were misclassified (A476D, F216Y, H255Q, H451R, and S345F), with three of them very close to the separation line (more details about individual mutations are provided in Table S1).

Table 2. GlcCer and GlcSph levels after AAVrh10 treatment. Levels of GlcSph and GlcCer species with defined N-acyl chain lengths were quantified in differentiated SH-SY5Y cells. Cells were harvested 15 days post-AAVrh10 infection (5×10^5 vg per cell). Data are mean \pm SD, $n = 3$. dGlc3, dGCase3; GC, GCase; ns, not significant.

	GBA ^{+/+}		GBA ^{-/-}		GBA ^{-/-} + WT GCcase		GBA ^{-/-} + dGCcase3		GBA ^{+/+} versus GBA ^{-/-}		GBA ^{+/+} versus GBA ^{-/-} + WT GC		GBA ^{+/+} versus GBA ^{-/-} + dGC3		GBA ^{-/-} + WT GC versus GBA ^{-/-} + dGC3		GBA ^{-/-} versus GBA ^{-/-} + dGC3	
	Lipid levels (pmol·mg ⁻¹)								Statistical significance (<i>P</i> values)									
Total GlcSph	0.90 ± 0.46	2082 ± 237	99.8 ± 61.3	16.6 ± 6.8	< 0.05	ns	ns	ns	< 0.05	< 0.05	ns	ns	ns	< 0.05	ns	< 0.05		
Total GlcCer	366.6 ± 187.9	7142 ± 434	931.3 ± 595.5	454 ± 190	< 0.001	ns	ns	ns	< 0.001	< 0.001	ns	ns	ns	< 0.001	ns	< 0.001		
d18.1/C14.0	1.44 ± 0.39	164.8 ± 16.3	40.4 ± 10.4	25.9 ± 5.0	< 0.01	ns	ns	ns	< 0.01	< 0.01	ns	ns	ns	< 0.01	ns	< 0.01		
d18.0/C14.0	0.03 ± 0.01	3.27 ± 0.33	0.78 ± 0.19	0.48 ± 0.09	< 0.05	ns	ns	ns	< 0.05	< 0.05	ns	ns	ns	< 0.05	ns	< 0.05		
d18.1/C16.0	53.4 ± 15.4	2818 ± 177	313 ± 116	149.3 ± 38.8	< 0.005	ns	ns	ns	< 0.005	< 0.001	ns	ns	ns	< 0.001	ns	< 0.005		
d18.1/C18.1	0.09 ± 0.04	27.1 ± 2.4	0.98 ± 0.27	0.27 ± 0.04	< 0.01	ns	ns	ns	< 0.01	< 0.01	ns	ns	ns	< 0.01	ns	< 0.01		
d18.0/C18.1	0.14 ± 0.04	10.2 ± 0.5	1.13 ± 0.25	0.59 ± 0.11	< 0.005	ns	ns	ns	< 0.005	< 0.001	ns	ns	ns	< 0.001	ns	< 0.005		
d18.1/C18.0	40.8 ± 12.4	2610 ± 170	294 ± 67.4	163 ± 25.9	< 0.005	ns	ns	ns	< 0.005	< 0.005	ns	ns	ns	< 0.005	ns	< 0.005		
d18.1/C20.0	12.2 ± 3.7	267 ± 11.1	37.5 ± 13.8	19.4 ± 4.9	< 0.001	ns	ns	ns	< 0.001	< 0.001	ns	ns	ns	< 0.001	ns	< 0.001		
d18.1/C22.0	69.9 ± 18.8	280 ± 39.8	56.1 ± 27.1	28.3 ± 10.4	< 0.05	ns	ns	ns	< 0.05	< 0.05	ns	ns	ns	< 0.05	ns	< 0.05		
d18.1/C24.1	103 ± 35.1	740 ± 191	139 ± 82.6	43.9 ± 18.0	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		
d18.1/C24.0	80.7 ± 22.2	198 ± 47.7	42.4 ± 23.9	19.0 ± 8.3	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		
d18.1/C26.1	1.67 ± 0.59	8.01 ± 2.48	1.66 ± 1.08	0.60 ± 0.27	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		
d18.0/C26.1	0.06 ± 0.02	0.41 ± 0.09	0.07 ± 0.05	0.03 ± 0.02	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		
d18.1/C26.0	2.86 ± 0.62	18.6 ± 4.20	3.27 ± 1.53	2.55 ± 1.08	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns		

Prediction of clinical phenotypes using the PRAMP score

Nearly 300 GD-causing mutations have been documented in *GBA1*, including ~230 missense mutations (Table S1 and [34]). Limited genotype–phenotype correlation is available, with a few exceptions. Thus, homozygosity for N370S always results in type 1 GD [35] and homozygosity for L444P invariably leads to nGD [36], although there is significant clinical variation even among patients homozygous for these well-characterized mutations. Predicting disease course is particularly problematic in compound heterozygotes.

We attempted to predict the clinical severity of known GD mutations using the PRAMP score. The scores of the GD-causing mutations (Table S1) have a median value of -2.49 with some as low as -7.5 . As shown previously, mutations with lower scores have lower enzyme activity and likely correspond to mutations that cause a more severe form of the disease. Indeed, a clear trend of a decreased PRAMP score correlating with increased disease severity is observed (Fig. 7A, Table 3). In particular, homozygous mutations [34] causing type 1 GD have a significantly higher PRAMP score than GD type 2 ($P < 0.01$; Fig. 7A). Δ PSSM was the most important parameter for separating disease-causing from benign mutations (Fig. 6A). Even so, mild N370S and severe L444P mutations have the same Δ PSSM (7), and their distinct severity is reflected in $\Delta\Delta G$ values ($\Delta\Delta G_{(N370S)} = -11$, $\Delta\Delta G_{(L444P)} = 24$). Our atomistic calculations performed with PROSS are consistent with studies showing that GCase with an N370S mutation gives a stable protein with reduced enzyme activity [37], whereas the L444P mutation leads to protein structure destabilization which results in ER-assisted degradation [38,39].

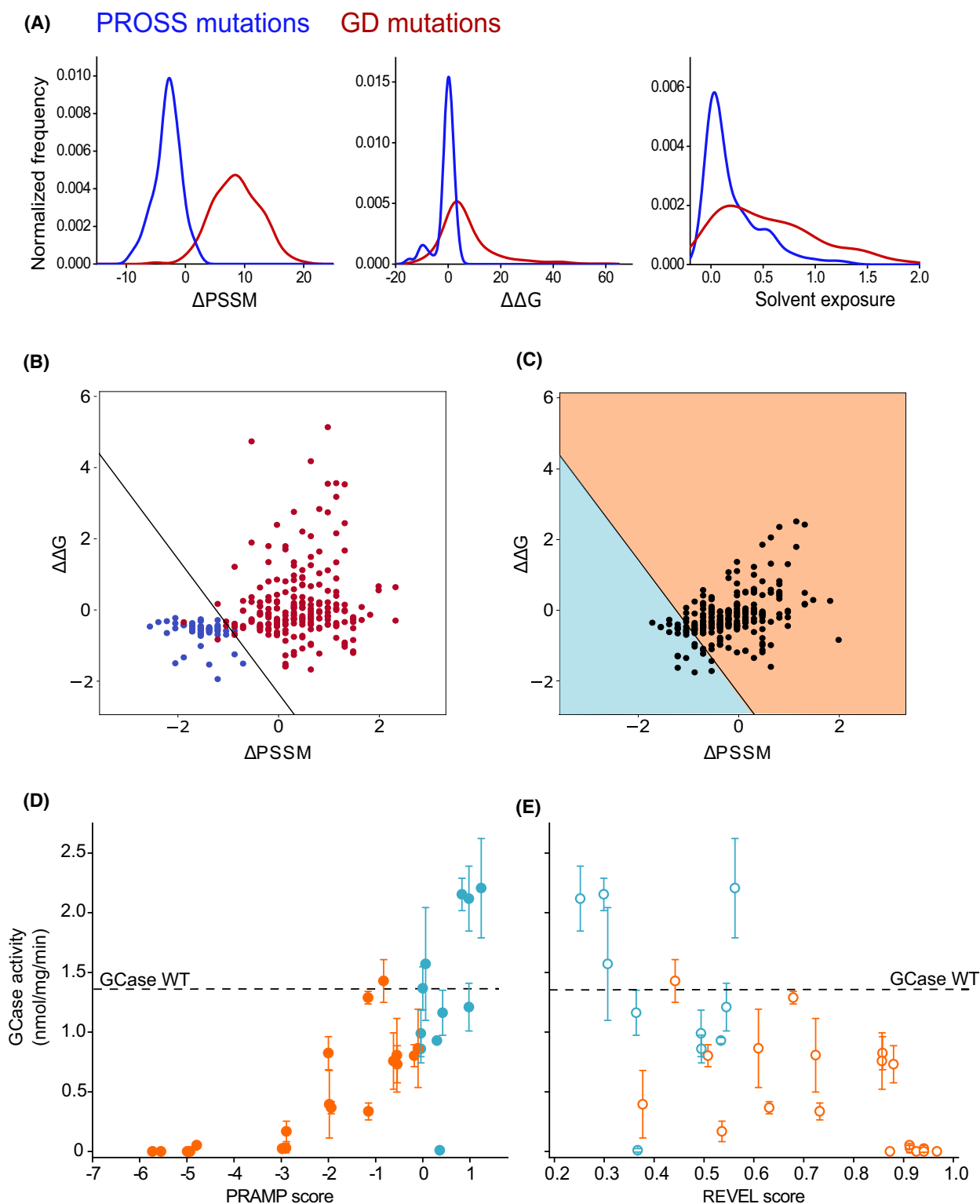
We next tested whether a similar approach could be used to predict the clinical severity exhibited by patients with different mutations in each allele, that is, compound heterozygotes. Normally, N370S in one allele, even if the second mutation is a more severe mutation (e.g., L444P), results in a disease closer to that observed with homozygous N370S than with homozygous L444P, suggesting that N370S can protect against the more severe (neurological) disease associated with L444P [36]. This being the case, a geometric rather than an arithmetic average was used to calculate the score of compound heterozygotes, since it is weighted in favor of the milder allele (lower score). PRAMP scores for compound heterozygous and the few homozygous mutations for which clinical data are available were taken from [40,41]. As for homozygous mutations, the PRAMP score for compound heterozygous

mutations also decreases with disease severity, yielding significant differences between mutations related to GD type 1 and GD type 2 ($P < 0.005$) and GD types 1 and 3 ($P < 0.05$; Fig. 7C). The same mutation sets were also assessed using REVEL (Fig. 7B,D). Although a trend of higher REVEL scores, that is, classified as more harmful, was observed for mutations related to more severe disease, the only significantly distinct REVEL scores were obtained for GD type 2 and GD type 1 for compound heterozygous mutations ($P < 0.05$). Moreover, a correlation between the age of disease onset and the mutation score was observed (Table 4) yielding Spearman coefficients of 0.94 and -0.77 for PRAMP and REVEL scores, respectively. Taken together, our results indicate that the clinical outcome of *GBA1* mutations can be predicted to a large extent by the impact of the mutation on protein stability and expression. Comparison of the outcome of the PRAMP algorithm with the REVEL classifier showed similar trends, but better performance of our prediction algorithm documented by higher correlation coefficients and significantly distinct PRAMP scores between the individual GD types.

Discussion

Our study makes two important contributions based on stability-design calculations. First, a GCase design comprising 55 mutations exhibits several potential advantages relative to the WT human enzyme for possible use in gene therapy since the design exhibits higher *in vitro* GCase activity and better performance upon AAV transduction in terms of enzymatic activity and GlcCer/GlcSph clearance. Second, by assuming that all PROSS-designed mutations are benign, we augmented clinical data to generate a classifier of the effect of mutations. The PRAMP classifier correctly predicted the functional characteristics of SNPs that have not been assigned disease status and demonstrated promise in predicting disease severity. Taken together, these results suggest that this simple predictor may provide a novel diagnostic tool, although clearly other factors, such as genetic and environmental factors, are also likely to play a role in disease severity.

In terms of gene therapy, dGCase variants could be expressed via viral vectors, which is an attractive approach for overcoming the neurological symptoms in nGD patients [10]. While the higher stability is likely to be of great advantage, the finding that dGCase3 has similar kinetic parameters to Cerezyme[®], along with its ability to clear GlcCer better than WT GCase, suggests that using the designed GCase may

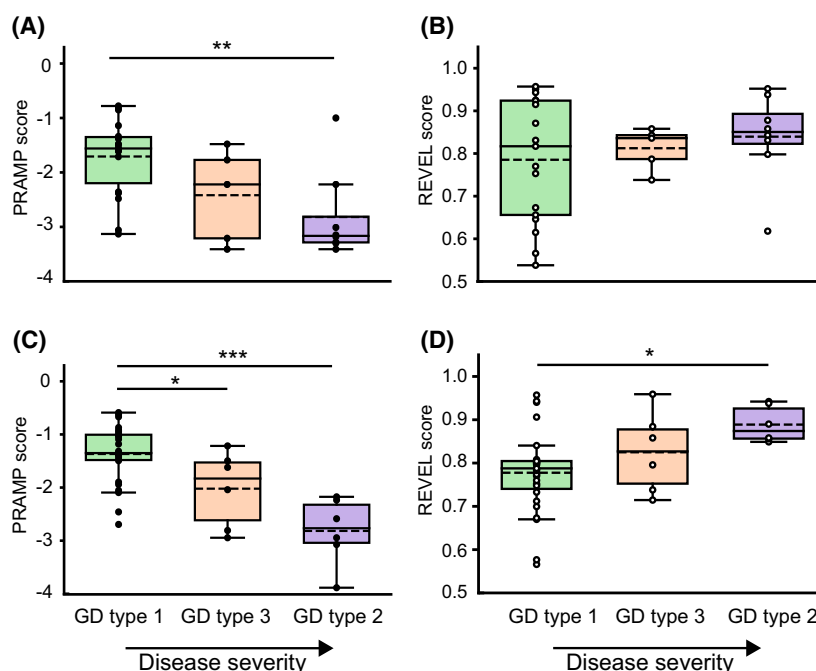


indeed be of great value in gene therapy approaches. Vector dose-dependent immune responses and toxicity have been observed in several gene therapy trials. The

optimization of transgene product levels and activity may permit reduction of the vector dose required to achieve therapeutic efficacy [42].

Fig. 6. SNP classification predicts functional outcome. Changes introduced into the GCase sequence by PROSS, together with GD-causing mutations, were used to construct a PRAMP algorithm to identify harmful and nonharmful mutations. (A) Separation of PROSS (blue) and GD (red) mutations according to one of the calculated parameters. Histograms of Δ PSSM, $\Delta\Delta G$, and solvent exposure fraction were calculated for each mutation. (B) Classification of the mutation training sets (PROSS, blue, and GD-causing, red) by PRAMP according to their Δ PSSM and $\Delta\Delta G$ scores (Δ PSSM and $\Delta\Delta G$ scores were normalized); the separation line is depicted in black. The y-axis is in Rosetta energy units. (C) Prediction of the algorithm using the SNP data (black dots; see Table S3). The separation line is depicted in black; the fields with potentially harmful and benign mutations are colored in orange and blue, respectively. The y-axis is in Rosetta energy units. (D, E) Enzymatic activity of WT GCase, together with that of the individual GCase point missense mutants expressed in HEK293T GBA^{-/-} cells. GCase activity was assayed on cell homogenates. Mutations with positive (benign) and negative (harmful) PRAMP scores are shown in blue and orange, respectively. The dashed lines indicate the level of enzyme activity of WT GCase. GCase activity was correlated with (D) PRAMP scores or to (E) REVEL scores. Values of the PRAMP score decrease with mutation severity whereas values of the REVEL score increase with disease severity. Data are means \pm SD, $n = 3$. Further details are given in Table S3.

Fig. 7. Correlation of the PRAMP score with GD types. Each mutation was assigned (A, C) a PRAMP score or (B, D) a REVEL score. (A, B) Homozygous GD-causing mutations together with their clinical classification were taken from Ref. [34]. (C, D) Compound heterozygous and several homozygous GD-causing mutations together with their clinical classification were taken from Ref. [41]. The PRAMP and REVEL scores of mutations are depicted according to which GD type they are associated with type 1 (green), type 2 (purple), and type 3 (orange). Means and medians are depicted by solid and dashed lines, respectively. Lists of individual mutations together with their scores are given in Tables S4 and S5. Statistical significance was calculated by ANOVA and *post hoc* Tukey's test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.005$.



Our results also suggest that stability design can be successfully applied to other proteins that cause LSDs. Although LSDs are individually rare, taken together they are found in $\sim 1 : 5000$ individuals [43], and most LSDs are caused by missense mutations, as in the case of GD. While our stability-based PRAMP classifier is attractive, other factors need to be considered when predicting disease severity, such as genetic background [44] and environmental factors [45]. For instance, patients homozygous for L444P present with a quite different clinical course depending on their genetic background [46], even though all have nGD. This suggests that the PRAMP score could be used to predict the type of GD, that is, type 1, 2, or 3, but may need to be combined with other factors in order to distinguish subtle differences in the clinical course of each type of disease in individual patients. One area that has not yet received attention is patients who are compound heterozygotes.

The PRAMP score predicted for compound heterozygotes in the current study gave a reasonable fit with the relatively limited clinical data available, supporting the possibility that the PRAMP score could be used to predict the clinical course of GD in compound heterozygotes. Such predictions could guide treatment regimes.

The ability to predict disease severity based on a classifier that distinguishes between stabilizing and disease-causing mutations not only paves the way to redefining genotype–phenotype correlations but also has exciting implications for understanding protein structure and function. This is exemplified in the mutations found at N370, with N370S the most common mutation leading to type 1 GD, whereas N370D is a stabilizing mutation identified by PROSS. Remarkably, 30 of the mutations in dGCase3 impact positions in which disease-causing mutations or predicted disease-causing SNPs have been identified. Thus,

	ΔPSSM	ΔΔG	Solvent exposure	PRAMP score	REVEL score	Phenotype
T369M	1	0.91	0.35	−0.20	0.731	Mild GD, PD
E326K	1	1.32	0.53	−0.23	0.595	PD
R496H	2	5.41	0.35	−0.78	0.566	Mild
N370S	7	−10.95	0.04	−0.82	0.673	Mild
V394L	5	4.53	0.47	−1.47	0.851	Severe
D409H	4	7.92	0.27	−1.48	0.738	Severe
R463C ^a	9	3.20	0.74	−2.38	0.817	Mild
L444P	7	23.58	0.12	−3.41	0.858	Severe

^aThis mutation is usually classified as mild, but it has been found in several patients with type 3 GD [36,59].

Table 4. Correlation of PRAMP score with the age of onset of GD. Genotypes and the average age of onset (years) (from Ref. [40]) are shown along with the PRAMP and REVEL scores. Spearman coefficients are 0.94 and −0.77 for PRAMP and REVEL scores, respectively.

Genotype	Age of onset	PRAMP score	REVEL score
N370S/N370S	30	−0.82	0.673
N370S/other ^a	17	−1.43	0.770
N370S/L444P	19	−1.47	0.760
D409H/D409H	4–5	−1.48	0.738
L444P/other ^a	2–7	−2.91	0.869
L444P/L444P	1–3	−3.41	0.858

^aIn some clinical reports, the second mutation is not identified. In these cases, we used the median of the score of all GD-causing mutations (a value of −2.49 and 0.88 for PRAMP and REVEL scores, respectively).

PROSS stability-design calculations are sensitive enough to suggest mutations that stabilize the protein even at positions where other mutations lead to severe disease. Our results open the way to designing candidate proteins for improved enzyme replacement or gene therapy and suggest sensitive tools for diagnosing the pathological effects of SNPs.

Materials and methods

PROSS design

Stabilized GCase variants were designed by PROSS2 [47] using the PROSS web server (<https://pross.weizmann.ac.il/step/pross-terms/>). Designs were generated based on the GCase crystal structure (PDB: 3gxi [25]). Mutations were manually curated to avoid the active site (C126, D127, F128, W179, N234, E235, Y244, P245, F246, D283, Q284, H311, Y313, E340, C342, G344, S345, W381, N382, F397, and V398) and known disease-causing mutations (Table S1); no GD-causing mutations appear in the set of PROSS mutations; none of the four glycosylation sites (N19, N59, N146, and N270) were altered by PROSS. The

Table 3. Correlation of PRAMP score with the most common GD mutations. The most common GD-causing mutations are listed together with the three parameters calculated for each mutation. Mutations that are risk factors for PD are also indicated. Mutations are ordered according to the PRAMP score from the least harmful (lowest score) to the most harmful (highest score). REVEL scores are also shown. The reported clinical phenotype is given for each mutation.

set of PROSS mutations is given in Table S2. ΔPSSM, ΔΔG, and solvent fraction exposure parameters are documented in Supplementary Dataset (10.17632/pkcjn539b5.1).

Reagents

Escherichia coli and human codon-optimized genes encoding GCase PROSS variants were obtained from IDT (Coralville, IA, USA). The affinity resins for protein isolation were purchased from the following: Ni²⁺ chelate chromatography from Adar Biotech (Rehovot, Israel) and Strep-Tactin[®]XT 4Flow high-capacity resin from IBA GmbH (Göttingen, Germany). Recombinantly expressed WT GCase (r-GCase; Cat. Nr. 7410-GHB-020) was purchased from R&D Systems (Minneapolis, MN, USA). The fluorescent substrate for the GCase activity assay *N*-[6-[(7-nitro-2-1,3-benzoxadiazol-4-yl) amino]hexanoyl]-D-glucosyl-β1-l'-sphingosine (C6-NBD-GlcCer) was obtained from Avanti Polar Lipids (Alabaster, AL, USA). The following antibodies were used for western blotting: anti-GCase (Cat. Nr. G4171, Sigma Aldrich, Darmstadt, Germany), anti-His (Cat. Nr. A7058, Sigma Aldrich), anti-StrepMAB (Cat. Nr. 2-1509-001, IBA GmbH) and anti-GAPDH (Cat. Nr. MAB374, Sigma Aldrich).

Cell culture

Protein expression was performed in the *E. coli* SHuffle T7 Express strain (New England BioLabs, Ipswich, MA, USA), grown in LB media containing Kanamycin (30 μg·mL^{−1}). HEK 293T and SH-SY5Y cells were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 110 μg·mL^{−1} sodium pyruvate, 100 IU·mL^{−1} penicillin, 100 μg·mL^{−1} streptomycin and nonessential amino acids in a humidified incubator at 37 °C with 5% CO₂.

Protein expression in *E. coli*

WT GCase and dGCase3 were cloned into a pET28-bdSUMO [48] vector (Fig. 1A) containing an N-terminal His-tag for purification and expressed in *E. coli*. Cells were

grown at 30 °C until OD₆₀₀ reached 0.6–0.8, followed by induction of protein expression (by 200 μ M IPTG) at 15 °C for ~18 h. Proteins were isolated from *E. coli* lysates using Ni²⁺ chelating chromatography followed by release from the column using SUMO protease [49]. Protein purity was assessed on 10% Tris-glycine SDS/PAGE gels stained with Coomassie blue. GCase variants were identified by western blotting using anti-His antibodies.

Protein expression in HEK293T cells

Genes coding for WT GCase and dGCase1, dGCase2, and dGCase3 were cloned into a pcDNA 3.1 vector, together with an N-terminal Twin-Strep isolation tag (Fig. 1B). Proteins were targeted extracellularly using the N-terminal R-PTP-S secretion signal (MGILPSPGMPALLSLVSLLSVLLMGCVA) [50]. For protein expression, HEK293T cells were grown in 10 cm culture dishes and transiently transfected using the polyethyleneimine reagent with 10 μ g of plasmid per dish (DNA : PEI ratio was 1 : 1.5 w/w). Growth media were collected 36–48-h post-transfection.

Purification of WT GCase and dGCase from the extracellular medium

GCase was isolated from the medium using a Strep-Tactin®XT 4Flow high-capacity resin. Medium was transferred to 250 mL tubes and centrifuged at 10 000 *g* (4 °C, 20 min) to remove detached and dead cells. The medium was then transferred into 50 mL Falcon tubes and 200 μ L of affinity resin suspension in Tris buffer (150 mM NaCl/50 mM Tris, pH 7.4) was added. Tubes were placed on a rotator at 4 °C overnight. The tubes were then centrifuged (4000 *g*, 4 °C, 20 min) and the medium aspirated. The resin was washed with an excess of Tris buffer by three centrifugation steps (4000 *g*, 4 °C, 20 min). GCase was released from the Strep-TactinXT resin using five consecutive elution steps with 50 mM biotin. Biotin was dissolved in sodium citrate buffer (40 mM trisodium citrate, 15 mM disodium hydrogen citrate, 187 mM D-mannitol, and 0.1% (v/v) mL Tween 80, pH 6.1). The eluted protein was stored in sodium citrate buffer. Protein purity was assessed on 10% Tris-glycine SDS/PAGE gels stained with Coomassie blue. GCase was identified by western blotting using anti-GCase and anti-StrepMAB antibodies. Protein preparations were assayed for enzymatic activity and subjected to differential scanning fluorimetry. The kinetic and spectroscopic data were compared with the corresponding data for Cerezyme® and for recombinant WT GCase (r-GCase).

Differential scanning fluorimetry

Differential scanning fluorimetry (DSF) was performed using a NanoDSF Prometheus NT.48 instrument (NanoTemper

Technologies GmbH, Munich, Germany). Samples were heated at 1 °C·min⁻¹ steps over a 20–95 °C temperature range. The fluorescence emission of tyrosine and tryptophan was recorded at 330 and 350 nm, respectively. Data were analyzed using a PR.THERMCONTROL v2.1.1 instrument (NanoTemper Technologies GmbH). The melting temperature (T_m) was defined as the inflection point of the fluorescence intensity (FI) ratio curve, where $R(FI) = FI_{350nm}/FI_{330nm}$.

Enzyme activity

Enzymatic activity was determined using a fluorescently labeled substrate of GCase, C6-NBD-GlcCer, as described [51,52]. The assay was performed using 0.1 μ g of pure protein or 7 μ g of cell homogenate in a final volume of 20 μ L McIlvarine buffer, pH 4.2. The reaction was run at 37 °C for 5 min and terminated by the addition of 1.5 mL of chloroform-methanol (1 : 2, v/v) prior to lipid extraction.

For the kinetic study, the activity of purified GCase was determined using *p*-nitrophenyl- β -D-glucopyranoside [37]. An aliquot of the enzyme was incubated with 0.2–4 mM *p*-NP-Glc, pH 5.9, at 25 °C, for 60 min. The reaction was terminated by 50-fold dilution into 1 M glycine buffer, pH 10.0, and absorbance of the *p*-nitrophenol was measured at 405 nm using an Agilent Cary 3500 spectrophotometer (Agilent Technologies, Santa Clara, CA, USA). Data were analyzed using the Michaelis–Menten eq.

SH-SY5Y and HEK293T *GBA*^{-/-} cells

HEK293T *GBA*^{-/-} and SH-SY5Y *GBA*^{-/-} cells were produced by Crispr/Cas9 genome editing [53]. A guide sequence (CATAGCGGCTGAAGGTACCA) was chosen to optimize for both off- and on-targeting using the MIT CRISPR design tool [54], and the sgRNA designer Rule set 1 [55], respectively. The guide sequence was cloned into a pSpCas9(BB)-2A-GFP vector and transfected into cells. Isolation of clonal cell populations was performed 24 h after transfection by FACS sorting. Single cells were sorted using GFP fluorescence, into 96-well plates containing 100 μ L medium in each well. After 1–3 weeks, viable colonies were transferred to 24-well plates and collected for verification of *GBA1* knock-out by western blotting. Endogenous GCase activity and GlcCer levels were determined in cell homogenates (Fig. S1).

AAV vector preparation

Vectors were generated at the translational vector core (CPV) of the University Hospital of Nantes by packaging AAV2-based recombinant genomes containing DNA sequences encoding WT GCase or dGCase3 under the control of a ubiquitous CAG promoter (Fig. 1C) into AAVrh10 capsids using helper virus-free transfection of

HEK293 cells. The vectors were purified using an optimized CsCl gradient-based purification protocol [56]. Viral protein purity and identity were verified by SDS/PAGE silver staining, and vector titers quantified by qPCR with primers targeting the flanking sequence of ITR2.

Transduction of SH-SY5Y cells using AAV

Nondifferentiated SH-SY5Y cells were seeded in 6-well plates (300 000 cells per well in 2 mL culture medium) (Day 1). On Day 2, 0.5 mL of medium was replaced by the same volume of transduction medium containing the vector at 0.5×10^5 , 1×10^5 , and 5×10^5 vg per cell. On Day 3, 0.5 mL of fresh cell culture medium was added. Cells were collected on Day 5 using trypsin/EDTA and pelleted by centrifugation (5 min, 1000 g, 4 °C). Pellets were used immediately or stored at -80 °C.

SH-SY5Y GBA^{+/+} and GBA^{-/-} cells were differentiated as follows [57]. Three hundred thousands cells were seeded in 6-well plates (35 mm diameter) precoated with poly-L-lysine. The next day (Day 1), medium was replaced by low-FBS media (DMEM, 1% FBS, 110 $\mu\text{g}\cdot\text{mL}^{-1}$ sodium pyruvate, 100 IU $\cdot\text{mL}^{-1}$ penicillin, 100 $\mu\text{g}\cdot\text{mL}^{-1}$ streptomycin and nonessential amino acids) with 10 μM retinoic acid (RA) added just prior to use. On Day 2, the medium was replaced by fresh low-FBS medium containing 10 μM RA and AAV (5×10^5 vg per cell). On Day 5, the low-FBS medium was exchanged with Neurobasal medium (Neurobasal medium containing B-27 supplement, 20 mM KCl, 100 IU $\cdot\text{mL}^{-1}$ penicillin, 100 $\mu\text{g}\cdot\text{mL}^{-1}$ streptomycin, nonessential amino acids, 2 mM glutamine, 50 ng $\cdot\text{mL}^{-1}$ BDNF, 2 mM db-cAMP) with 10 μM RA. The medium was replaced every 3 days by fresh Neurobasal medium containing RA. Cells were harvested after 12 or 15 days.

Cell pellets were lysed by five freeze–thaw cycles in 150 μL of McIlvaine buffer (41 mM Na_2HPO_4 , 59 mM citric acid, pH 4.2) with a protease inhibitor cocktail (1 : 100) and DNase (1 : 200). Protein concentrations of cell homogenates were determined using the BCA reagent. Enzyme activity was assayed using C6-NBD-GlcCer.

Cell transfection using GCase missense mutants

Single-point missense mutations were introduced into the WT GCase sequence in pcDNA 3.1 plasmids. HEK 293T GBA^{-/-} cells were cultured in 6-well plates and transfected using the polyethyleneimine reagent using 2 μg of plasmid per well. Cells were collected 36–48 h post-transfection. Enzymatic activity was measured as described in the previous section.

Lipidomic analysis

Cell homogenates were prepared as described in previous sections except that cell pellets were lysed in double-distilled

water with a protease inhibitor cocktail (1 : 100). Quantitative analysis of sphingolipids in cell homogenates was performed by liquid chromatography–tandem mass spectrometry [58].

PRAMP algorithm

All mutations used are listed in Supplementary Dataset and Tables S1 (GD-causing), S2 (PROSS) and S3 (SNPs). A comprehensive list of GD-causing missense mutations was created via literature review. To generate the list of SNPs (that have not been detected in GD patients), variants of *GBA1* were downloaded from the NCBI Variation Viewer (<https://www.ncbi.nlm.nih.gov/variation/view/>) human genome version GRCh38.12. The list was filtered prior to download with the molecular consequence ‘missense variant’. The list was manually annotated to ensure that all protein-coding changes were included in the dataset and duplicates or synonymous mutations had been removed. Mutations without documented clinical significance and with uncertain significance were chosen.

ΔPSSM was calculated by subtracting the PSSM score of the mutated amino acid from that of human GCase. The PSSM table was extracted from the PROSS stability-design calculations, as were the $\Delta\Delta\text{G}$ calculations. The solvent exposure fraction was calculated using the STRIDE webserver (<http://webclu.bio.wzw.tum.de/stride/>). All the parameters can be found in the Supplementary dataset (10.17632/pkcn539b5.1). The PRAMP algorithm was built by a custom-written Python script using LinearSVC function (scikit-learn).

REVEL scores were downloaded from the website: <https://sites.google.com/site/revelgenomics/downloads> and parsed manually to create a file with only the GBA locus for comparison to PRAMP.

The PRAMP score for compound heterozygous mutations was calculated as the negative geometric average of the two individual PRAMP scores, and the REVEL score for compound heterozygous was calculated as geometric average of the two individual REVEL scores. Statistical significance was evaluated either by the Student's *t*-test or by analysis of variance (ANOVA) followed by *post hoc* pairwise comparisons using the Tukey's honest significant difference test. Correlations were evaluated using the Spearman coefficient.

Acknowledgements

We thank Drs Maxim Itkin and Sergey Malitsky (Lipidomics) and Dr Yael Fridman-Sirkis (DSF measurements) from the Life Sciences Core Facilities at the Weizmann Institute of Science, Dr Ron Rotkopf and Shani Blumenreich-Kashani for help with statistical analysis, Yochai Maytal for help with collating data

for Table S1 and Chen Yaacobi for creating the GBA^{-/-} cell lines. Research in the Futerman laboratory was supported by a Sponsored Research Agreement between the Weizmann Institute of Science (via Yeda, its technology transfer office) and Lysogene. Research in the Fleishman laboratory was supported by a Consolidator Award from the European Research Council (815379), the Israel Science Foundation (1844), the Dr Barry Sherman Institute for Medicinal Chemistry and by a charitable donation in memory of Sam Switzer. SP was partially supported by the Czech Academy of Sciences (Czech/Israel scientific program). AHF is the Joseph Meyerhoff Professor of Biochemistry at the Weizmann Institute of Science.

Author contributions

JLS, IS, SJF, and AHF designed the research. SP, RK, YA, YP, TU, SA, OD, AT, and RT performed the biochemical experiments. OK, RL-S, and AG performed the computational work. MH and RL contributed new reagents. SP, YA, OK, RL-S, and SBD analyzed data. SP wrote the manuscript. AHF wrote the manuscript and obtained funding. All authors discussed data and edited the manuscript.

Conflict of interest

Yeda Research & Development, on behalf of the Weizmann Institute of Science, has applied for patent applications corresponding to PCT/IL2021/050357 on the acid- β -glucosidase designs, naming AHF, IS, JLS, SJF, AG, SP, YA, and OK as inventors. SJF and AG are named inventors on stability-design patents corresponding to PCT/IL2016/050812. MH and RL are employees and shareholders of Lysogene.

Peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/febs.16758>.

Data availability statement

The experimental data that support the findings of this study are included in this article and supplementary material (Fig. S1 and Tables S1–S5). The structural data used within this study are openly available in the wwPDB (PDB 10.2210/pdb3GXI/pdb). The data used for the construction of the PRAMP algorithm are openly available in Mendeley data at 10.17632/pkcn539b5.1. The list of single-nucleotide polymorphism data is openly available at NCBI Variation

Viewer at <https://www.ncbi.nlm.nih.gov/variation/view/>.

References

- 1 Futerman AH & van Meer G (2004) The cell biology of lysosomal storage disorders. *Nat Rev Mol Cell Biol* **5**, 554–565.
- 2 Meikle PJ, Hopwood JJ, Clague AE & Carey WF (1999) Prevalence of lysosomal storage disorders. *JAMA* **281**, 249–254.
- 3 Futerman AH & Zimran A (2006) Gaucher Disease. 1st edn. Taylor and Francis Group, CRC Press, Boca Raton, FL.
- 4 Grabowski GA, Gaft S, Horowitz M & Kolodny EH (1990) Acid beta-glucosidase: enzymology and molecular biology of Gaucher disease. *Crit Rev Biochem Mol Biol* **25**, 385–414.
- 5 Grabowski GA (2008) Phenotype, diagnosis, and treatment of Gaucher's disease. *Lancet* **372**, 1263–1271.
- 6 Schiffman R, Fitzgibbon EJ, Harris C, DeVile C, Davies EH, Abel L, van Schaik IN, Benko W, Timmons M, Ries M *et al.* (2008) Randomized, controlled trial of miglustat in Gaucher's disease type 3. *Ann Neurol* **64**, 514–522.
- 7 Shayman JA (2013) The design and clinical development of inhibitors of glycosphingolipid synthesis: will invention be the mother of necessity? *Trans Am Clin Climatol Assoc* **124**, 46–60.
- 8 Weinberg MS, Samulski RJ & McCown TJ (2013) Adeno-associated virus (AAV) gene therapy for neurological disease. *Neuropharmacology* **69**, 82–88.
- 9 Cearley CN & Wolfe JH (2006) Transduction characteristics of adeno-associated virus vectors expressing cap serotypes 7, 8, 9, and Rh10 in the mouse brain. *Mol Ther* **13**, 528–537.
- 10 Hocquemiller M, Giersch L, Audrain M, Parker S & Cartier N (2016) Adeno-associated virus-based gene therapy for CNS diseases. *Hum Gene Ther* **27**, 478–496.
- 11 Massaro G, Hughes MP, Whaler SM, Wallom K-L, Priestman DA, Platt FM, Waddington SN & Rahim AA (2020) Systemic AAV9 gene therapy using the synapsin I promoter rescues a mouse model of neuronopathic Gaucher disease but with limited cross-correction potential to astrocytes. *Hum Mol Genet* **29**, 1933–1949.
- 12 Berg-Fussman A, Grace ME, Ioannou Y & Grabowski GA (1993) Human acid β -glucosidase. N-glycosylation site occupancy and the effect of glycosylation on enzymatic activity. *J Biol Chem* **268**, 14861–14866.
- 13 Magliery TJ (2015) Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol* **33**, 161–168.
- 14 Goldenzweig A, Goldsmith M, Hill SE, Gertman O, Laurino P, Ashani Y, Dym O, Unger T, Albeck S,

- Prilusky J *et al.* (2016) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol Cell* **63**, 337–346.
- 15 Goldenzweig A & Fleishman SJ (2018) Principles of protein stability and their application in computational design. *Annu Rev Biochem* **87**, 105–129.
 - 16 Campeotto I, Goldenzweig A, Davey J, Barford L, Marshall JM, Silk SE, Wright KE, Draper SJ, Higgins MK & Fleishman SJ (2017) One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. *Proc Natl Acad Sci USA* **114**, 998–1002.
 - 17 Kriegl M, Wiederanders HJ, Alkhashrom S, Eichler J & Muller YA (2021) A PROSS-designed extensively mutated oestrogen receptor α variant displays enhanced thermal stability while retaining native allosteric regulation and structure. *Sci Rep* **11**, 1–13.
 - 18 Peleg Y, Vincentelli R, Collins BM, Chen KE, Livingstone EK, Weeratunga S, Leneva N, Guo Q, Remans K, Perez K *et al.* (2021) Community-wide experimental evaluation of the PROSS stability-design method. *J Mol Biol* **433**, 1–14.
 - 19 Redler RL, Das J, Diaz JR & Dokholyan NV (2016) Protein destabilization as a common factor in diverse inherited disorders. *J Mol Evol* **82**, 11–16.
 - 20 Yazar M & Özbek P (2021) *In silico* tools and approaches for the prediction of functional and structural effects of single-nucleotide polymorphisms on proteins: an expert review. *OMICS* **25**, 23–37.
 - 21 Manickam M, Ramanan P, Singh P & Talwar P (2014) *In silico* identification of genetic variants in glucocerebrosidase (GBA) gene involved in Gaucher's disease using multiple software tools. *Front Genet* **5**, 1–10.
 - 22 Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, Hoffmann ER, Bernstein I, Anne-Marie G, Takahashi M *et al.* (2019) Computational and cellular studies reveal structural destabilization and degradation of mlh1 variants in lynch syndrome. *Elife* **8**, e49138.
 - 23 Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, Bar-Shira A, Berg D, Bras J, Brice A *et al.* (2009) Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N Engl J Med* **361**, 1651–1661.
 - 24 Blumenreich S, Jenkins BJ, Barav OB, Milenkovic I & Futerman AH (2020) The lysosome and nonmotor symptoms: linking Parkinson's disease and lysosomal storage disorders. *Mov Disord* **35**, 2150–2155.
 - 25 Lieberman RL, D'Aquino JA, Ringe D & Petsko GA (2009) Effects of pH and iminosugar pharmacological chaperones on lysosomal glycosidase structure and stability. *Biochemistry* **48**, 4816–4827.
 - 26 Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S & Steinegger M (2022) ColabFold – Making protein folding accessible to all. *Nat Methods* **19**, 679–682.
 - 27 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A, Potapenko A *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
 - 28 Dvir H, Harel M, McCarthy AA, Toker L, Silman I, Futerman AH & Sussman JL (2003) X-ray structure of human acid- β -glucosidase, the defective enzyme in Gaucher disease. *EMBO Rep* **4**, 704–709.
 - 29 Beutler E, Gelbart T, Kuhl W, Zimran A & West C (1992) Mutations in Jewish patients with Gaucher disease. *Blood* **79**, 1662–1666.
 - 30 Cabrera-Salazar MA, Deriso M, Bercury SD, Li L & Lydon JT (2012) Systemic delivery of a glucosylceramide synthase inhibitor reduces CNS substrates and increases lifespan in a mouse model of type 2 Gaucher disease. *PLoS ONE* **7**, e43310.
 - 31 Ransohoff RM & Engelhardt B (2012) The anatomical and cellular basis of immune surveillance in the central nervous system. *Nat Rev Immunol* **12**, 623–635.
 - 32 Ron I & Horowitz M (2005) ER retention and degradation as the molecular basis underlying Gaucher disease heterogeneity. *Hum Mol Genet* **14**, 2387–2398.
 - 33 Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al.* (2016) REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99**, 877–885.
 - 34 Hruska KS, LaMarca ME, Scott CR & Sidransky E (2008) Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum Mutat* **29**, 567–583.
 - 35 Horowitz M, Tzuri G, Eyal N, Berebi A, Kolodny EH, Brady RO, Barton NW, Abrahamov A & Zimran A (1993) Prevalence of nine mutations among Jewish and non-Jewish Gaucher disease patients. *Am J Hum Genet* **53**, 921–930.
 - 36 Koprivica V, Stone DL, Park JK, Callahan M, Frisch A, Cohen IJ, Tayebi N & Sidransky E (2000) Analysis and classification of 304 mutant alleles in patients with type 1 anti type 3 gaucher disease. *Am J Hum Genet* **66**, 1777–1786.
 - 37 Wei RR, Hughes H, Boucher S, Bird JJ, Guziewicz N, Van Patten SM, Qiu H, Pan CQ & Edmunds T (2011) X-ray and biochemical analysis of N370S mutant human acid β -glucosidase. *J Biol Chem* **286**, 299–308.
 - 38 Ohashi T, Hong CM, Weiler S, Tomich JM, Aerts JMFG, Tager JM & Barranger JA (1991) Characterization of human glucocerebrosidase from different mutant alleles. *J Biol Chem* **266**, 3661–3667.
 - 39 Liou B (2006) Analyses of variant acid beta-glucosidases: effects of Gaucher disease mutations. *J Biol Chem* **281**, 4242–4253.
 - 40 Grabowski GA, Zimran A & Ida H (2015) Gaucher disease types 1 and 3: Phenotypic characterization of

- large populations from the ICGG Gaucher Registry. *Am J Hematol* **90**, S12–S18.
- 41 Alfonso P, Aznarez S, Giralt M, Pocovi M & Giraldo P (2007) Mutation analysis and genotype/phenotype relationships of Gaucher disease patients in Spain. *J Hum Genet* **52**, 391–396.
 - 42 Colella P, Ronzitti G & Mingozzi F (2018) Emerging issues in AAV-mediated *in vivo* gene therapy. *Mol Ther Methods Clin Dev* **8**, 87–104.
 - 43 Meikle PJ, Broox DA, Ravenscroft EM, Yan M, Williams RE, Jaunzems AE, Chataway TK, Karageorgos LE, Davey RC, Boulter CD *et al.* (1997) Diagnosis of lysosomal storage disorders: evaluation of lysosome-associated membrane protein LAMP-1 as a diagnostic marker. *Clin Chem* **43**, 1325–1335.
 - 44 Klein AD, Ferreira NS, Ben-Dor S, Duan J, Hardy J, Cox TM, Merrill AH & Futerman AH (2016) Identification of modifier genes in a mouse model of Gaucher disease. *Cell Rep* **16**, 2546–2553.
 - 45 Jakóbkiewicz-Banecka J, Gabig-Cimińska M, Banecka-Majkutewicz Z, Banecki B, Węgrzyn A & Węgrzyn G (2014) Factors and processes modulating phenotypes in neuronopathic lysosomal storage diseases. *Metab Brain Dis* **29**, 1–8.
 - 46 Goker-Alpan O, Hruska KS, Orvisky E, Kishnani PS, Stubblefield BK, Schiffmann R & Sidransky E (2005) Divergent phenotypes in Gaucher disease implicate the role of modifiers. *J Med Genet* **42**, 1–7.
 - 47 Weinstein JJ, Goldenzweig A, Hoch S & Fleishman SJ (2020) PROSS 2: a new server for the design of stable and highly expressed protein variants. *Bioinformatics* **37**, 123–125.
 - 48 Zahradník J, Kolářová L, Peleg Y, Kolenko P, Svidenská S, Charnavets T, Unger T, Sussman JL & Schneider B (2019) Flexible regions govern promiscuous binding of IL-24 to receptors IL-20R1 and IL-22R1. *FEBS J* **286**, 3858–3873.
 - 49 Frey S & Görlich D (2014) A new set of highly efficient, tag-cleaving proteases for purifying recombinant proteins. *J Chromatogr A* **1337**, 95–105.
 - 50 Aricescu AR, Lu W & Jones EY (2006) Biological crystallography A time-and cost-efficient system for high-level protein production in mammalian cells. *Acta Crystallogr D* **62**, 1243–1250.
 - 51 Meivar-Levy I, Horowitz M & Futerman AH (1994) Analysis of glucocerebrosidase activity using N-(1-[14 C] hexanoyl)-D-erythro-glucosylsphingosine demonstrates a correlation between levels of residual enzyme activity and the type of Gaucher disease. *Biochem J* **303**, 377–382.
 - 52 Shaaltiel Y, Bartfeld D, Hashmueli S, Baum G, Brill-Almon E, Galili G, Dym O, Boldin-Adamsky SA, Silman I, Sussman JL *et al.* (2007) Production of glucocerebrosidase with terminal mannose glycans for enzyme replacement therapy of Gaucher's disease using a plant cell system. *Plant Biotechnol J* **5**, 579–590.
 - 53 Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA & Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308.
 - 54 Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827–832.
 - 55 Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ & Root DE (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267.
 - 56 Ciron C, Cressant A, Roux F, Raoul S, Cherel Y, Hantraye P, Déglon N, Schwartz B, Barkats M, Heard JM *et al.* (2009) Human α -iduronidase gene transfer mediated by adeno-associated virus types 1, 2, and 5 in the brain of nonhuman primates: vector diffusion and biodistribution. *Hum Gene Ther* **20**, 350–360.
 - 57 Shipley MM, Mangold CA & Szpara ML (2016) Differentiation of the SH-SY5Y human neuroblastoma cell line. *J Vis Exp* 53193.
 - 58 Kim JL, Ben-Dor S, Rosenfeld-Gur E & Futerman AH (2021) A novel C-terminal DxRSDxE motif in ceramide synthases involved in dimer formation. *J Biol Chem* **298**, 101517.
 - 59 Sidransky E, Bottler A, Stubblefield B & Ginns EI (1994) DNA mutational analysis of type 1 and type 3 gaucher patients: how well do mutations predict phenotype? *Hum Mutat* **3**, 25–28.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Characterization of HEK293T and SH-SY5Y GBA^{-/-} cells.

Table S1. Comprehensive list of missense mutations which cause Gaucher disease.

Table S2. PROSS mutations and their PRAMP scores.

Table S3. List of SNPs in GCaSe which have not been shown to cause Gaucher disease.

Table S4. Homozygous GD mutations used to generate Fig. 7A,B.

Table S5. Compound heterozygous and homozygous GD mutations used to generate Fig. 7C,D.