Mechanism of Glucocerebrosidase Activation and Dysfunction in Gaucher Disease Unraveled by Molecular Dynamics and Deep Learning

Raquel Romero[1^], Arvind Ramanathan[2^], Tony Yuen[3], Debsindhu Bhowmik[2], Mehr Mathew[3], Lubna Bashir Munshi[3], Seher Javaid[3], Madison Bloch[3], Daria Lizneva[3,4], Alina Rahimova[3], Ayesha Khan[3], Charit Taneja[3], Se-Min Kim[3], Li Sun[3], Maria New[5*], Shozeb Haider[1*] and Mone Zaidi[3*]

[1]Department of Pharmaceutical and Biological Chemistry, University College London School of Pharmacy, London WC1N 1AX, UK

[2]Computational Science and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

[3]Department of Medicine and Mount Sinai Bone Program, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[4]Department of Reproductive Health Protection, Scientific Center of Family Health and Human Reproduction, Irkutsk, Russian Federation

[5]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

[^]Joint first authors

*Corresponding authors: shozeb.haider@ucl.ac.uk, maria.new@mssm.edu and mone.zaidi@mssm.edu

Key Words: multiscale simulations; gene mutations; lysosomal storage disease; rare disease

SIGNIFICANCE STATEMENT

Gaucher disease is a rare genetic disorder that has crippling health consequences. Mutations in the *GBA1* gene are known to disrupt the enzyme glucocerebrosidase-1, but it is not known, at atom-level detail, as to how enzyme function is lost. This study uses multiscale simulations and deep learning to define precisely the mechanism underlying the disruption of glucocerebrosidase-1, and in particular, its interaction with the facilitator protein, saposin C.

ABSTRACT

The lysosomal enzyme glucocerebrosidase-1 (GCase) catalyzes the cleavage of a major glycolipid glucosylceramide into glucose and ceramide. The absence of fully functional GCase leads to the accumulation of its lipid substrates in lysosomes causing Gaucher disease, an autosomal recessive disorder that displays profound genotype–phenotype non-concordance. More than 250 disease-causing mutations in *GBA1*, the gene encoding glucocerebrosidase-1, have been discovered, although only one of these, N370S, causes 70% of disease. Here, we have used a knowledge-based docking protocol that considers experimental data of protein-protein binding to generate a complex between GCase and its known facilitator protein saposin C (SAPC). Multiscale molecular dynamics simulations were used to study lipid self-assembly, membrane insertion, and the dynamics of the interactions between different components of the complex. Deep learning was applied to propose a model that explains the mechanism of GCase activation, which requires SAPC. Notably, we find that conformational changes in the loops at the entrance of the substrate-binding site are stabilized by direct interactions with SAPC, and that the loss of such interactions induced by N370S and another common mutation L444P result in destabilization of the complex and reduced GCase activation. Our findings provide an atomistic-level explanation for GCase activation and the precise mechanism through which N370S and L444P cause Gaucher disease.

/body

INTRODUCTION

The enzyme glucocerebrosidase-1 (GCase) catalyzes the cleavage of the major glycolipid glucosylceramide (GL-1) into glucose and ceramide, and the minor lipid glucosylsphingosine (Lyso-GL-1) into sphingosine and water (1-4). The lipid tails of both glycolipids are embedded within the intra-lysosomal membrane, such that both substrates are inaccessible and require the assistance of an 84-residue facilitator protein saposin C (SAPC), a member of the Sphingolipid Activator Protein family (5-8). There is experimental evidence that both GCase and SAPC associate in the intra-lysosomal membrane, but the mechanism through which SAPC destabilizes intra-lysosomal vesicles to make lipids accessible to GCase is not well understood (2, 9).

Loss-of-function mutations of the *GBA1* gene encoding GCase result in a crippling human disorder, Gaucher disease (10). Despite being a monogenic disorder, Gaucher disease presents with extreme phenotypic variability, ranging from an asymptomatic form to disease characterized by severe organ damage (11). Hepatosplenomegaly, anemia, thrombocytopenia, osteoporosis and bone marrow infiltration are hallmarks, with neurodegeneration noted with certain mutations. Although ~250 *GBA1* mutations have been reported thus far (12), just one, N370S, is responsible for more than 70% of the cases of Gaucher disease type 1 in the Ashkenazi Jewish population (13, 14). Another mutation, L444P, accounts for ~40% of Gaucher disease types 2 and 3 worldwide (15-19). Our earlier studies have recapitulated the entire human phenotype in mice through the selective deletion *GBA1* in cells of the hematopoietic and mesenchymal cell lineages using the Cre-lox technology (20-22). However, there is no clear explanation yet as to how N370S and L444P mutations cause human disease and whether there is a role for SAPC in this process (13, 14, 23, 24).

SAPC not only mediates the contact of the GCase with its natural substrates, but is also known to induce a conformational change to stimulate enzyme activity directly (6, 9, 25). As a consequence, mutations in GCase that affect its association with SAPC would result not only in diminished GCase activity, but also in vulnerability of GCase to early degradation. Likewise, mutations in the *PSAP* gene that causes malfunction or absence of SAPC in the lysosomal compartment cause a juvenile form of Gaucher disease (7, 8).

The interaction between GCase and SAPC has been modeled earlier and although this serves as a good starting point, the study has considerable limitations (26). First, the available model is unable to account for experimental data (27, 28), a limitation in itself. Second, there is no structural information on how the GCase-SAPC complex interacts with the membrane. In separate studies, selected mutants have been modeled through molecular dynamics (MD) simulations (29, 30), but these studies lack information on the GCase-SAPC interface, specifically membrane anchoring and the influence of membrane lipid and substrate on the complex.

Here, we report a model of GCase in complex with SAPC, which has been constructed employing structural bioinformatics, including knowledge-based protein-protein docking. Multi-scale MD simulations were conducted to understand the structural mechanism underlying the association of GCase with SAPC in its membrane environment. The results from our deep learning approach explain the activation mechanism of GCase by SAPC and provide a structural explanation at the atomistic level on how the two most commonly occurring mutations N370S and L444P cause Gaucher disease.

RESULTS

The GCase-SAPC Complex

GCase is a globular protein composed of three domains (SI Appendix, Figure S1) (31). Domain I (residues 1–27 and 383–414) is a small three stranded anti-parallel β-sheet; domain II (residues 30–75 and 431–497) is an independent 8 stranded β-barrel; and domain III (residues 76–381 and 416–430) is an (α/β) and Triose-Phosphate Isomerase (TIM) Barrel, containing the active site. Domains I and III interact tightly and are linked by one of the loops at the entrance of the binding site. Domains II and III are separated by a long loop that acts as a hinge, with structural folds that are similar to other hydrolases, such as α-galactosidase (31, 32). The active site, containing two catalytic residues, namely E340 (catalytic nucleophile) and E235 (acid-base residue), lies in a cavity formed at the center of the TIM barrel motif surrounded by residues R120, D127, F128, W179, N234, Y244, F246, Y313, C342, S345, W381, N396, F397 and V398. Loops 1 through 5, containing residues 311–319, 345–349, 394–399, 237–248 and 283–288, respectively, presenting at the entrance of the active site, can rearrange in different conformations to regulate substrate accessibility (4, 31, 32). Two different conformations of Loop-1 have been reported, namely extended and helical. Notably, in the active state, Loop-1 is in a helical conformation with the side chain of residue D315 pointing towards residue N370, while in the inactive state, it adopts an extended conformation with residue D315 pointing towards Loop-2. Furthermore, the bulky side chain of W348 (Loop-2) is oriented towards outside the binding site in the active state, whereas in the inactive enzyme, it points towards the entrance of the active site and thus blocks it. Likewise, in the inactive state, the side chain of R395 (Loop-3) orients towards the catalytic residue E340, but points outside the binding site when the enzyme is active (4, 31, 32).

To understand the structural mechanism of GCase activation and the role of SAPC

binding, a GCase–SAPC complex was generated employing a knowledge-based protocol. Using the CPORT algorithm (33), the SAPC binding site was predicted to be located on helix-7, flanked by helix-6, Loop-1, Loop-2 at the entrance of the active site and Domain II (Figure 1A). This predicted location is consistent with experimental evidence, which identified the N370S binding site to be positioned on helix-7 (34). The Protein-Protein docking (PPD) program Hex (35) was used to generate the model, which was corroborated with an alternative program Haddock (36). The complexes were constructed using the crystal structures of the active and inactive conformations of the enzyme GCase (4, 32), and that of SAPC in its closed and open conformations (Figure 1B) (5, 37). Molecular docking studies demonstrated that the extended Loop-1 clashes with the binding site. Moreover, the lipid substrate glucosylceramide (GluCer) could not be properly positioned within the binding site when the Loop-1 is in the extended conformation. Three residues from Loop-3 (R395, N396, F397) play important roles in substrate accessibility to the active site (31).

After applying screening criteria to the top 20 solutions of each docking run, only one docking solution was identified that was common to both active and inactive complexes using both docking programs. This docked pose was among the top results after selection using different correlation methods. We noted that the SAPC-binding site on GCase lay between Domain III Loops-1 (H311–P319) and -2 (S345–S351) at the entrance of the active site, helix-6 (K321–L330), helix-7 (W357–L372) and Domain II (T43–S45, Q440–D445, L461–S465 and Y487). Table S1 shows the residues involved in the protein-protein binding and electrostatic interactions.

Coarse–Grained Molecular Dynamics

To understand the structural basis of the catalytic function of GCase, we studied its dynamics and the conformational changes arising from interactions with other components,

such as the lipid bilayer, SAPC, and GluCer. We thus used coarse-grained MD (CG-MD) to study lipid self–assembly, specifically to determine how the complex anchors to the bilayer. GCase was simulated to position the complex on the lipid bilayer interface in the absence or presence of GluCer and SAPC. A total of five 1.2 μs-long CG simulations were conducted employing the Martini coarse-grained force field (Table 1). Membrane assembly occurred between 40 and 120 ns in all simulations (SI Appendix, Figure S2). The proteins/complexes became inserted into the lipid membrane immediately after their formation and remained anchored throughout the course of the simulations. Importantly, the entrance of the GCase active site oriented towards the bilayer, consistent with allowing GluCer access into the binding site anchored from within the membrane. The orientation of SAPC anchoring to the membrane was also consistent with that observed in experimental studies (37). Membrane anchoring became stronger during the initial equilibration phase, in all the instances. The distance between the centers of mass between GCase and the lipid bilayer decreased as the equilibration progressed and thereafter remained stable. Furthermore, in atomistic simulations of the complex (3a and 3b in Table 2), the equilibrium distance to the membrane increased as SAPC became positioned between GCase and the membrane (SI Appendix, Figure S3).

Atomistic Molecular Dynamics of Wild Type GCase

Atomistic MD (AT-MD) allowed us to study, at atom-level detail, the dynamics of the interactions between different components in the system. Once the membrane-protein complexes were assembled in CG-MD (above), they were converted to atomistic models and simulated using classical AT-MD. A total of 10 simulations were performed (Table 2). To evaluate conformational stability of the complex over time, root-mean-square deviation (RMSD) of Cα atoms from the initial structure (SI Appendix, Figure S4) and RMS Fluctuation (RMSF) *per* residue (SI Appendix, Figure S5) were calculated for simulations 2a (active, no SAPC); 2b

(inactive, no SAPC); 3a (active complex), and 3b (inactive complex). All four configurations were found to be stable over the simulated period, with each reaching equilibration at ~250 ns.

Conformational changes at the protein-protein interface and at the surface were monitored by following changes in surface electrostatics in both active and inactive GCase conformations in the complex (see SI Appendix, Figures S6–S8). In simulation 3a, the electrostatic surface of GCase did not alter significantly throughout the simulation. During simulation 3b, the change in the electrostatic surface was found to be prominent. Notably, at its start, the SAPC binding region on GCase was positively charged. The area of SAPC was flat when compared in simulation 3a, in which GCase showed more cavities and was noted as slightly more negative. Towards the end of simulation 3b, the binding area of SAPC appeared more negative and irregular, with some cavities appearing that were equivalent to those observed in simulation 3a. There was also considerable difference in the electrostatic surface in the catalytic pocket. In simulation 3a, the catalytic pocket was deeper and wider than in the first part of 3b. Towards the end of 3b, however, the electrostatics of the catalytic site had changed, appearing wider, and similar to that of simulation 3a. We postulate that changes in electrostatic surface pattern in GCase are a result of conformational changes and could possibly be under the influence of SAPC binding.

Active Site Loop Dynamics of Wild Type GCase

Analysis of the loop dynamics at the entrance of the active site shed light on the activation mechanism of the enzyme. In simulation 2a (no SAPC), Loop-1 lost its helical structure as the simulation progressed. However, in simulation 3a (active complex), when SAPC was present, the interaction between D315 of GCase and K34 of SAPC stabilized the helical conformation of Loop-1 and this configuration was maintained over the simulation (Figure 2A). In inactive state complex (simulation 3b), residue K34 of SAPC interacted with the

backbone atoms of L314 and Y373, which are maintained throughout the simulation and force Loop-1 to adopt a near-helical conformation. It is important to note that the latter two residues surround D315, a key residue that forms interactions with SAPC (31). Of note also is that the impetus for Loop-1 to adopt a helical state is absent without SAPC in simulation 2b (Figure 2B).

We also observed differences in the conformation of Loop-2 and Loop-3 in the presence or absence of SAPC (Figure 3A). In both simulations 3a and 3b, the side chain of W348, located in Loop-2 of GCase, was oriented towards the outside of the binding site. In the active complex (simulation 3a), the side chain of W348 was tucked in a hydrophobic pocket formed by SAPC (Figure 3B). However, in simulation 2b (inactive GCase, no SAPC), the bulky indole side chain of W348 was found to partially obstruct the entrance to the binding site, while in simulation 2a, W348 became embedded in the membrane. In the inactive state (simulation 2b), residue R395 (Loop-3) and catalytic residue E340 formed a stable hydrogen bond, which occluded the entrance to the active site, thus preventing substrate access (Figure 3C). This interaction was not observed in simulation 3b (inactive complex), in which R395 oriented towards the outside of the binding site, with a final orientation as observed in the active state (Figure 3C).

We found that a number of Protein-Protein interactions (PPI) stabilized the GCase–SAPC complex. In simulation 3a (active complex, Figure 4 and SI Appendix, Figure S9), residue K34 of SAPC formed a stable hydrogen bond with residue D315, which is essential in maintaining the helicity of Loop-1. There was a stable interaction in Loop-2 between residues S44 (SAPC) and W348 (GCase). In helix-7, there were PPIs between residues D30 (SAPC) and H365 (GCase). Finally, interactions in Domain II of GCase included those between: D52 (SAPC) and R44 and Y487 (GCase); S60 (SAPC) and S464 (GCase); S60 (SAPC) and S464 (GCase); and K26 (SAPC) and N442, D445, D443 (backbone) and L444 (backbone) of GCase.

In simulation 3b (inactive complex), residues T24 and K34 of SAPC formed stable interactions with Loop-1, as well as with the side chain and backbone of residues K321 and L314, respectively. In Loop-2, two stable PPIs formed between residues S44 (SAPC) and E349 (GCase) and between S37 (SAPC) and K346 (GCase). In helix-7, there were PPIs between residues D33 (SAPC) and H365 (GCase), D30 (SAPC) and Y373 (GCase) and K34 (SAPC) and Y373 (GCase). Finally, interactions within Domain II of GCase included Q48 (SAPC) with S45 (GCase); D52 (SAPC) with R44 and S465 (GCase); S56 (SAPC) with S465 and S464 (GCase); and K26 (SAPC) and D443, L444 with D445 (GCase) (SI Appendix, Figure S10 and S11).

Atomistic Molecular Dynamics of Mutant GCases

AT-MD simulations were also performed for two of the most clinically prevalent Gaucher mutations in GCase, namely: N370S and L444P. Mutant GCases were simulated in complex with SAPC in an intraluminal membrane environment, using both active and inactive conformations (Table 2). Cα-RMSD of GCase was calculated in all four simulations and compared to the wild type. Notably, there was an overall conformational stability of GCase in the three active state simulations, where Loop-1 adopted a helical conformation (simulations 3a, 5a and 6a, SI Appendix, Figure S12). The equilibration time in the wild type simulation (3a) was shorter (~100 ns) than in $GCase^{N370S}$ and $GCase^{L444P}$ mutants (~250 ns) (5a and 6a). In simulation 3a, the average Cα-RMSD from the end of the equilibration was lower in simulation 3a (2.4 ± 0.1 Å) than in simulation 5a (3.1 ± 0.2 Å) or 6a (3.4 ± 0.1 Å), indicating a more stable wild type GCase in the complex. Analysis of RMSDs of simulations containing the extended, inactive state conformation of Loop-1, namely simulations 3b, 5b and 6b, showed a similar trend after equilibration (SI Appendix, Figure S13). In simulation 3b, unlike its active counterpart, the average RMSD value from the end of the equilibration was slightly higher (3.8 ± 0.1 Å) than in simulation 5b (3.6 ± 0.2 Å) and similar to the average in simulation 6b (3.8 ± 0.1 Å). In

simulation 6b, GCase exhibited the greatest conformational drift compared with all other simulations.

However, the mutant GCase^{N370S}–SAPC and GCase^{L444P}–SAPC complexes were unstable; thus affecting the conformation of GCase over the course of the simulations. These simulations also showed that point mutations affected loop dynamics. In the first 300 ns of simulation 5a (GCase^{N370S}–SAPC), the helical conformation of Loop-1 was lost. This helicity, however, partly recovered when interactions between K34 (SAPC) and the side chain of D315 (GCase) occurred during the second half of the simulation 5a (Figure 5).

In simulation 6a (GCase^{N370S}–SAPC), Loop-1 retained the helical conformation, although the helix was deformed and moved towards Loop-2. In both mutants, the poor coupling between GCase and SAPC rendered Loop-2 free, unlike in the wild type simulations, where Loop-2 remained tucked in a hydrophobic pocket formed in SAPC. The evolution of Loop-3 was also different in the two mutants. While in the wild type, residue R395 was oriented towards the outside of the active site, in simulation 5b (GCase^{N370S}–SAPC, inactive state), it was oriented towards the inside creating interactions with residue S350 of Loop-2. This interaction lies adjacent to a bulky phenylalanine side chain that impeded the return of Loop-3 to an open conformation. In simulation 6b (GCase^{L444P}–SAPC, inactive state), the guanidinium side chain of residue R395 pointed towards the exterior of the binding pocket, although Loop-3 was more closed than in the wild type.

A comparison of active site loop dynamics in simulations of the inactive state (Loop-1, extended form) of wild type and mutant GCase, also highlight some important differences (Figure 5D–F). In simulations 5b (GCase^{N370S}–SAPC, inactive state) and 6b (GCase^{L444P}–SAPC, inactive state), Loop-1 extended towards helix-7. Residue W348 did not remain consistently tucked in the hydrophobic pocket on SAPC, as was noted in the wild type

simulation 3b. In simulations 5b and 6b, Loop-3 adopted a closed conformation, whereby residue R395 interacted with the catalytic residue E340. This interaction completely obstructed the binding site, and was similar to that observed in simulation 2b, where inactive state GCase (Loop-1 extended conformation) was simulated without SAPC.

PPIs were also affected in mutant complex simulations, with the disruption of many interactions identified in the wild type GCase–SAPC complex. These differences were most pronounced in simulations 5a and 5b, where residue N370 was mutated to serine (SI Appendix, Figure S14). In active wild type GCase–SAPC complex (simulation 3a), the interaction between residue H365 in helix-7 and D30 of SAPC was stable throughout the simulation. This interaction was completely lost in the GCase^{N370S}–SAPC mutant complex. In active state simulation 5a, a PPI between D315 of GCase and K34 of SAPC was formed at ~400 ns, and partially recovered Loop-1 helicity. PPIs between residue K26 of SAPC and residue N442 and D443 in the proximities of L444 were disrupted, while those between K26 (SAPC) and L444 and D445 (GCase) were maintained (SI Appendix, Figure S15). Other disrupted PPIs that included those between residue W348 (GCase) and S44 (SAPC) and Q440 (GCase) and E64 (SAPC). In inactive GCase^{N370S}–SAPC (simulation 5b), SAPC became loosely attached to the GCase^{N370S} after ~400 ns. At this point, SAPC was positioned near a completely deformed Loop-1 and the PPIs formed were between residue K321 near Loop-1 and residues D30 and E27 of SAPC. Towards the end of the simulation, new PPIs between the GCase^{N370S} and SAPC formed; these however did not involve helix-7 (which contains residue 370). Finally, interactions between K26 and L444 and surrounding residues were completely abrogated in this GCase^{N370S}–SAPC mutant simulation. The interaction between D30 of SAPC and Y373 in the proximities of N370 was also disrupted; this is otherwise stable in the corresponding wild type protein complex (SI Appendix, Figure S16).

Equally prominent differences were noted when L444 was mutated to proline in simulations 6a and 6b (SI Appendix, Figure S17). In the active state GCase^{L444P}–SAPC (simulation 6a), interactions between residues K26 (SAPC) and P444 and D445 (GCase) were disrupted from ~600 ns onwards, although the interaction with residue D443 was maintained beginning at ~500 ns. Interactions of residues in SAPC with Loop-1 of GCase were almost non-existent towards the end of the simulation, but some interactions between SAPC and Domain I and II of GCase remained stable from 500 ns onwards (SI Appendix, Figure S18). In the inactive state GCase^{L444P}–SAPC simulation 6b, interactions between residue K26 (SAPC) and residues P444 and other surrounding residues, including D445 and D443, were completely lost. The disruption of these interactions makes SAPC partially detached and translates towards the end of helix-7 near Domain I. Interactions with Loop-1 were almost non-existent. Stable interactions that remained included those between S44 (SAPC) and Q350 (GCase) and between D52 (SAPC) and R353 or W357 (backbone) of GCase (SI Appendix, Figure S19).

Deep Clustering of AT-MD Simulations

Using the AT-MD simulations, we next probed how the wild type and mutant simulations differ with respect to their dominant motions. We posited that the conformational motions of GCase, especially when subjected to interactions with SAPC, would be non-linear. Hence, linear models such as principal component analysis (PCA) may not sufficiently capture the conformational diversity in these simulations (38, 39). To account for the non-linearity in protein conformational fluctuations, we recently developed a deep clustering approach to identify intermediate states from folding trajectories (see Methods) (40). We examined whether our deep clustering approach based on a convolutional variational auto-encoder (CVAE; see Methods) could elucidate (a) the differences in the conformational motions between the active and inactive states of the wild type and mutant GCase, and (b) the different conformational

states that are influenced by the motions within the wild type and mutant GCase AT-MD. We used contact matrices from GCase as a starting point for our analysis.

We first examined how many intrinsic latent dimensions are necessary to describe the conformational diversity observed from the wild type AT-MD simulations. To estimate this, we plotted the overall loss (L) as a function of the number of dimensions in the latent space (Figure 6A). This is similar to the cumulative variance plots used to estimate the total number of principal modes needed to describe the observed variance in the simulations for techniques such as PCA (38). Our results for the CVAE show that, as the number of intrinsic dimensions increase, the RMS loss also decreases. As shown in Figure 6A, however, the loss in the validation dataset increased beyond 14 dimensions, indicating that the CVAE is over fitting. Hence, for the GCase system, we can use a 14-dimensional latent space to describe the conformational motions sampled in the simulations.

We evaluated the performance of CVAE on the mutant AT-MD simulations. As shown in Figure 6A (inset), the RMS loss for the mutant simulations was higher on average compared to the wild type AT-MD simulations (average RMS loss of 7.93 in wild type *vs* 24.77 in mutant AT-MD simulations). This difference in RMS loss allowed us to posit that the conformational motions in the wild type and mutant AT-MD simulations are different. Note that this is significant given that we used contact matrices from GCase (without the substrate or SAPC included) to build the latent space representations and the RMS loss captures how well the model trained on the wild type AT-MD simulations captures the conformational motions in the mutant simulations.

To understand how the conformational motions in the GCase simulations are different, we next examined the CVAE latent space. Given that a 14-dimensional latent space is difficult to visualize, we used the t-distributed stochastic neighborhood embedding (t-SNE) to examine the latent space in 3 dimensions. As shown in Figure 6C, the t-SNE based visualization allowed

us to distinguish the conformational states visited by the simulations – especially in the context of the distance between the Cα atoms of residues 340 and 395. The distance between these two residues is critical, considering the formation of the ion–pair interaction between E340-R395 obstructs the binding pocket and locks it down in a closed conformation. A histogram of the distances between these two residues (Figure 6B) represents the presence of three distinct states that are labeled I–III – we examined if the CVAE-based clustering can recapitulate these states. The CVAE clustering of the AT-MD simulations shows a clear distinction between the loop conformations as depicted in the 3D-representations (Figure 6D). Note that the CVAE representation does not use the distance between these residues as input, but discovers these as a consequence of the differences in the conformational motions. Corresponding representations of these sample conformations are shown in cartoon form in Figure 6D.

DISCUSSION

GCase–SAPC protein-protein binding sites have been defined both experimentally and *in silico*. Based upon the competition of synthetic lipids, two binding sites located at position 6–27 and 45–60, and two activation sites at positions 27–34 and 41–48 on SAPC have been defined in an earlier study (28). In a separate study, chimeric saposins were used to identify a single activation site between residues 47–62 (27). Together these studies led to two sets of possibilities: one was that the two activation sites lay adjacent to the loops at the entrance of the active site in GCase and exerted actions on the surrounding environment, and two, that the SAPC activation site lay adjacent to the loops at the entrance of the active site in GCase. A pose consistent with the first premise was identified *via* Hex docking when active and inactive conformations of GCases were used. This pose was also identified in results from Haddock docking, *albeit* being the only plausible pose. In this study, we have followed a knowledge-based docking protocol to characterize the complete GCase–SAPC protein-protein interface in

depth, which satisfies all the requirements to be an optimal pose. The predicted binding site is in agreement with experimental data (27, 28) and lies adjacent to the loops at the entrance of the active site of GCase (31). CG-MD was further used to characterize the association of the GCase–SAPC complex in the membrane, providing an opportunity to observe the lipid self-assembly process. Quality controls demonstrated that the membrane was formed correctly, and that wild type and mutant GCase–SAPC complexes anchored to the membrane as peripheral membrane proteins in all the simulations.

To understand the conformational changes within the GCase–SAPC complexes in depth, the CG coordinates of all simulations were transformed to atomistic, with further bursts of 1000 ns simulations, with a total sampling time of 9 μs. Notably, the values of RMSD, which reflected conformational drifts in the systems, were more stable when GCase was simulated along with SAPC than when simulated alone (Simulation 3a and 3b, respectively), suggesting that SAPC normally stabilizes GCase. In contrast, the mutated GCase^{N370S}–SAPC and GCase^{L444P}–SAPC complexes were unstable in both inactive and active states. Furthermore, when GCase was simulated in inactive state (Loop-1 extended), it exhibited higher RMSD values than in active state (Loop-1 helical). A comparison of interactions made by N370 and L444 in the wild type (SI Appendix, Figures S20–S23) has been made with N370S and L444P mutants (SI Appendix, Figures S24–S31).

We also analyzed RMSF values to study loop dynamics, interactions within the binding site, and PPIs. The highest RMSF peaks corresponded to surface loops, whereas the core structure was stable, with some differences in the loops at the active site entrance (SI Appendix, Figures S5 and S32–S33). Loop-1 partially lost its helical form in active state (simulation 2a), whereas it extended towards helix-7 in inactive state (simulation 2b). In the active state complex (simulation 3a), Loop-1 conserved its helicity during the course of the entire simulation due to the restraint placed by interaction with residue K34 of SAPC. In the inactive state

complex (simulation 3b), however, Loop-1 did not change its extended conformation, and Loop-2 displayed RMSF values above 2Å in two simulations, 2a and 3b. In simulation 2a (active state, no SAPC), the loop moved from a helical conformation to become embedded inside the lipid membrane, and in 3b (inactive complex), the loop was tucked in a hydrophobic pocket present in SAPC. Loop-3 displayed high mobility only during simulation 3b, moving from a closed conformation, where the side chain of residue R395 pointed towards the interior of the binding pocket, to an open conformation.

Loop dynamics at the entrance of the active site in GCase shed further light on the GCase–SAPC interactions, and their aberration with the two disease-causing mutations GCase^{N370S} and GCase^{L444P}. Loop-1 (residues 311–319) normally adopts a helical conformation in the active state when simulated as a complex (simulation 3a), specifically because of interactions between D315 (GCase) and K34 (SAPC). The helicity of Loop-1 is lost partly when GCase is simulated without SAPC in simulation 2a (no SAPC), wherein Loop-1 establishes interactions with residues of Loop-2, namely W348 and K346. The simulation of mutant GCase^{N370S}–SAPC and GCase^{L444P}–SAPC complexes (simulations 5a and 6a; SI Appendix, Figure S34) mimicked the simulation of GCase without SAPC. The mutants show a complete loss of helical conformation, with partial recovery once an interaction between D315 and K34 is established. Subsequent stability results from interactions between H365 and S366 in helix-7.

Furthermore, in simulations with the wild type GCase–SAPC complex (3a and 3b), residue W348 is tucked in a hydrophobic pocket on SAPC at the protein-protein interface (SI Appendix, Figure S35). In simulations with both mutants, except in active state GCase^{N370S}-SAPC (simulation 5a), W348 is not tucked inside this hydrophobic pocket. Among all the simulations of inactive states, the open conformation of Loop-3 is identified only in simulation 3b. Notably, in the rest of the simulations of inactive states of both mutant complexes and wild type GCase (without SAPC), namely 5b, 6b and 2b, respectively, residue R395 in Loop-3 forms

an H-bond with E340, which completely obstructs the binding pocket and locks it in a closed conformation. Together, the data provide the structural basis for poor accessibility of lipid substrates to the GCase catalytic site in both mutations, N370S and L444P, even in the presence of unaltered SAPC.

Atomistic MD although necessary, is not sufficient to derive a complete picture of the free energy surface of a protein (41). Traditional MD analyses measure conformational drifts, such as RMSD or radius of gyration, and therefore cannot be used to infer dominant motions accountable for protein function (42). Further, given the nature of complex interactions of GCase with SAPC and its substrates, we expected the conformational motions to be non-linear. Hence, we used a deep clustering approach that indicates that Loop-1 was functionally relevant, with secondary roles of other Loops. However, we noted that we used only the contact maps of GCase (generated from the trajectories) as inputs for our analysis. Our analysis was able to identify three sub-states in our simulations corresponding to (a) the inactive enzyme, (b) an intermediate and (c) an active conformation. The inactive conformation passed through an intermediate state to result in the active conformation. Of note is that the intermediate state was the same conformation as observed towards the end of simulation 3b (inactive complex), whereby W348 (Loop-2) was inserted in a hydrophobic pocket on SAPC. This sub-state enabled the stable binding of K34 (SAPC) with D315 (Loop-1), and in turn, influenced Loop-3 to orient away from the binding site. The conformation was observed only in the GCase–SAPC complex.

The structural differences identified at the GCase–SAPC interface and the stability of interactions in different simulations together reflect the dynamics of the protein-protein recognition. Proteins do not fit in a static manner as building blocks, but do so *via* a flexible and evolving process. The disruption of some of these interactions can alter this evolution, thereby making recognition at the protein-protein interface unfavorable, best exemplified by GCase^{L444P}.

Notably, the mutation of L444 to proline, a cyclic and more rigid residue, prevents its interaction with residue K26 of SAPC, essentially abrogating GCase–SAPC interactions. In contrast, in the N370S mutation, Loop-1 extended towards helix-7 resulting in the loss of its interaction with SAPC, which then destabilizes the complex.

Methods

Protein-Protein Docking

Prior information about interactions at the protein-protein interface can limit docking sampling and increase the chance of obtaining accurate results. Here we have used CPORT (33) interface predictor to propose potential sites of protein-protein interactions on GCase and SAPC surface. A model of GCase in complex with SAPC was generated using a knowledge-based protocol. The complex was constructed by docking the X-ray crystal structures of active (PDB id 2NSX) (4) or inactive (PDB id 1OGS) (32) conformation of GCase with the X-ray crystal structures of SAPC in its closed (PDB id 2GTG) (5) and open conformations (PDB id 2QYP) (37). The Protein-Protein docking (PPD) program Hex (35, 43, 44) was used to generate the model, which was corroborated with a second docking program, Haddock (36).

Docking calibration was performed in two sets of docking experiments. In the first set, geometrical parameters of the program were adjusted. The combination of correlation type and post–processing procedure was optimized in a second set of calibration experiments. Once the most suitable parameters were identified, protein-protein docking experiments were conducted in combinations of different conformations of both partner proteins. For each combination, two series of seven docking runs were carried out using the parameters obtained from the calibration experiments. In the first series, the center of mass of each protein was used as a centroid or origin of the geometrical operations. In the second series, residue H365 of GCase was used as a centroid or origin. The docking poses obtained were screened *per* different criteria including the relative position of the proteins and number of electrostatic interactions, taking into account results of protein-protein interface predictors and known experimental data, namely binding and activation regions of SAPC. A total of 6 docking poses were selected for energy minimization using the MD engine AMBER 12 (45). A second docking program,

Haddock, was also used to validate docking datasets. A total of 5 runs were conducted making different selections for passive and active residues of both proteins.

Coarse Grained Molecular Dynamics (CG-MD)

A total of five CG-MD simulations were carried out in order to study the insertion of the proteins and their complexes in a lipid bilayer representative of a lysosomal membrane. These included (1) inactive GCase, (2) inactive GCase with GluCer in its binding site, (3) active GCase with GluCer in complex with SAPC and SAPC alone in (4A) open and (4B) closed conformations. The systems were parameterized using the Martini Forcefield in the Gromacs molecular dynamics engine to group atoms in clusters of four ("beans") for evaluation of their physicochemical properties (46).

For parameterization, the atomistic models were converted using the *maritinize.py* script, with side chain beads generated employing the elastic network option (elastic bond force = 500 kJ mol^{-1} nm^{-2}, with lower and upper elastic cutoffs at 0.5 nm and 0.9 nm, respectively). In simulations 2 and 3, where GluCer was simulated, the parameters for the substrate were obtained from Martini website. The substrate was manually positioned in the active site, using atomistic coordinates extracted from the crystal structure. Lipids tails of the substrate were extended as the parameters identified accounted for a molecule with shorter acyl tails.

The majority of phospholipids in lysosomal membranes are phosphatidylcholine, typified by dipalmitoyl phosphatidylcholine (DPPC), the most widely used phosphatidylcholine for simulations. A box of DPPC was generated using CG parameters from a single DPPC molecule. The optimum number of lipids for each system was identified using trial and error.

Systems containing proteins or their corresponding complexes and the correct number of DPPC molecules were energy minimized and solvated in alternate runs until the desired

portion of water/DPPC was obtained. Energy minimization was performed in two consecutive steps employing the steepest descent and conjugate gradient (1000 cycles) algorithms. MD-CG simulations were run for 1.2 μs with a time-step for integration of 0.003 ns. Standard cut-off schemes for non-bonded interactions were used to conduct the simulations. Lennard-Jones interactions were shifted to zero between 0.9 and 1.2 nm, and electrostatic interactions were shifted to zero between 0 and 1.2 nm. The non-bonded neighbor list cut-off was set to 0.14 nm to improve energy conservation, and the list was updated every ten steps. Temperature was coupled separately for protein/complexes, lipids and solvent using Berendsen algorithm at 325 K, using a time constant for coupling of 1.5 ps. Pressure of the system was coupled semi-isotropically using Berendsen algorithm at 1 bar, a compressibility of $3x10^{-5}$, and a time constant for coupling of 3.0 ps.

To convert CG to atomistic coordinates, *a* snapshot, representative of the protein/complex, inserted in the self-assembled membrane was selected. The Sugarpie script was chosen to carry out the CG to atomistic conversion after water and GluCer were removed from the CG coordinates. The atomistic structures used as templates for the conversion were same as those used for atomistic to CG conversion. After the conversion, GluCer was repositioned in the binding site by performing an alignment with the docked structure.

Atomistic Molecular Dynamics (AT-MD)

Ten simulations (total sampling time 9μs) were run with different systems that were converted from CG-MD (Table 2). Active and inactive states were identified by surveying different crystalline forms of GCase in the Protein Data Bank. The inactive state was defined when Loop-1 adopted an extended conformation and R395-E340 ion pair occluded the entrance of the binding site, while the active state was defined when Loop-1 was in a helical conformation

and the ion pair interaction was lost. The systems were built using Gromos53A6 force field (47) and Gromacs was used as the MD engine.

For AT parameterization, force field compliant topologies were generated using proteins from the converted models. The substrate GluCer was added in some simulations, by aligning the converted models to the initial docked structure. CG models were converted using the same AT-coordinates that were used to generate them. The converted models were, in some cases, used to obtain different GCase conformations or mutants by aligning to the desired structures. Simulation 2-CG was used to obtain the atomistic coordinates for simulation 2a and 2b and simulation 3-CG was used to obtain the coordinates for simulation 3a, 3b, 5a, 5b, 6a and 6b. The mutant GCases, $GCase^{N370S}$ and $GCase^{L444P}$, were generated using the mutagenesis program in the ICM-Pro molecular modeling package (www.molsoft.com), using the same PDB structures as that of the wild type. The atomistic mutant simulations were set up by converting CG structures of the complex extracted from 3-CG simulation, using the mutated proteins instead of the wild type.

The models were solvated using Single Point Charge water (SPC) and energy minimized using 5000 steps of steepest descent method. Counter ions were added to neutralize the systems. A second round of energy minimization was conducted employing an additional 5000 steps of the steepest descent method. Two rounds of equilibration were carried out: (a) 0.1 ns of NVT equilibration with time steps of 0.002 ns, using V-rescale algorithm for temperature coupling (separately for protein/complexes, lipids and solvent) at 323 K, using a time constant for coupling of 0.5 ps; and (b) 1 ns of NPT equilibration with a time step of 0.002 ns, using Nose-Hoover temperature coupling and Parrinello-Rahman for pressure coupling. The pressure of the system was coupled semi-isotropically using Berendsen algorithm at 1 bar, a compressibility of 4.5×10^{-5} and a time constant for coupling of 5.0 ps. The production run was carried out for 1000 ns without any restraints with a time-step of 0.002. A cut-off of 1.2 nm was

chosen for neighbor list generation and coulomb and Lennard-Jones interactions. Particle-Mesh-Ewald summation was chosen for electrostatic interactions. For those systems with two proteins or ones that included the substrate, strong position restraints were applied for energy minimization runs and soft restraints for equilibration phase simulations. There were no restraints on the system during the production run.

Deep Learning

Each trajectory was processed using the MDAnalysis library (48) to extract the contact matrices using the Cα atoms; a distance cut-off of 8 Å or less was used to define two residues to be in contact. Contact matrices offer a natural advantage: they are independent of rotation/translation issues, which can be problematic while analyzing MD trajectories. We then used a convolutional variational auto-encoder (CVAE) to capture the large-scale conformational motions within a low-dimensional latent space in an unsupervised fashion (40). Auto-encoders typically have an hour-glass-like architecture, where the data (from MD trajectories) are compressed into a low dimensional latent space, and then reconstructed using successive 'output' layers. Variational auto-encoders (VAEs) force the latent space to be normally distributed (49). This constraint provides a way to overcome the issues with sparsity in latent dimensions and forces the latent representation to utilize all of the information within the MD trajectories.

We used convolutional layers as inputs to the VAE (and thus, the term convolutional VAE/CVAE) since sliding filter maps can better describe secondary and tertiary structure interactions, as quantified from the contact map dynamics from the MD trajectories. We trained the CVAE on the wild type AT-MD simulations; we then used the mutant AT-MD simulations as testing data, while simultaneously inferring how these simulations may be different from the WT simulations. Although in unsupervised learning applications, we do not require a cross-

validation step, we used 60/40 split in the training/validation data (wild type AT-MD simulations) to assess the quality of the CVAE build. This also allowed us to estimate the number of intrinsic latent dimensions required to sufficiently describe the conformational motions observed in the wild type AT-MD simulations. The objective of the CVAE is to reduce the loss (L), which is composed of two terms: (1) the reconstruction loss, E_r, which measures the ability of the CVAE in reconstructing the input contact matrix. It is calculated as the cross entropy loss between f(z), which indicates the reconstructed probability of contact between two Cα atoms and the original X conformations from the simulation, which indicate the existence of contact between two Cα atoms and (2) the latent loss, E_l, which measures the loss as a consequence of constraining the latent space to be normal distribution. The latent loss is defined as a regularizing constraint that forces the latent embeddings z to conform to a Gaussian distribution; this is calculated as the Kullback-Leibler (KL) divergence between the latent embeddings z and a normal distribution with mean 0 and standard deviation 1 (40).

$$L = E_r + E_l$$

$$E_r = -\frac{1}{n}\sum_{i=1}^{n} X_i \log(f(z_i))$$

$$E_l = KL(z||Normal(0,1))$$

Details of the reconstruction loss and the latent loss are described in our previous work (40). We used the RMSProp algorithm to train the CVAE and trained the models for 50 epochs (SI Appendix, Figures S36 and S37). Similar to PCA (50), the projections of the simulations onto the CVAE latent space representation provide information on the dominant motions (represented as VAE_i, where i represents the particular index to the latent space) sampled in the simulations. However, these modes are not ordered – for convenience, we just order the modes based on the variance accounted for in the simulations.

ACKNOWLEDGEMENTS

The authors thank Dr. Pramod Mistry (Yale) for his invaluable advice. A.R. and D.B. were supported in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy (DE-AC05-00OR22725). M.Z. gratefully acknowledges the National Institutes of Health (NIH) for grants R01 AG23176, R01 AR65932 and R01 AR67066 (to M.Z.), and DK113627 (to M.Z. and L.S.).

REFERENCES

1. Kolter T & Sandhoff K (2010) Lysosomal degradation of membrane lipids. *FEBS Lett* 584(9):1700-1712.
2. Schulze H, Kolter T, & Sandhoff K (2009) Principles of lysosomal membrane degradation: Cellular topology and biochemistry of lysosomal lipid degradation. *Biochim Biophys Acta* 1793(4):674-683.
3. Vasella A, Davies GJ, & Bohm M (2002) Glycosidase mechanisms. *Curr Opin Chem Biol* 6(5):619-629.
4. Lieberman RL*, et al.* (2007) Structure of acid beta-glucosidase with pharmacological chaperone provides insight into Gaucher disease. *Nat Chem Biol* 3(2):101-107.
5. Ahn VE, Leyko P, Alattia JR, Chen L, & Prive GG (2006) Crystal structures of saposins A and C. *Protein Sci* 15(8):1849-1857.
6. Fabbro D & Grabowski GA (1991) Human acid beta-glucosidase. Use of inhibitory and activating monoclonal antibodies to investigate the enzyme's catalytic mechanism and saposin A and C binding sites. *J Biol Chem* 266(23):15021-15027.
7. Tamargo RJ, Velayati A, Goldin E, & Sidransky E (2012) The role of saposin C in Gaucher disease. *Mol Genet Metab* 106(3):257-263.
8. Tylki-Szymanska A*, et al.* (2011) Gaucher disease due to saposin C deficiency, previously described as non-neuronopathic form--no positive effects after 2-years of miglustat therapy. *Mol Genet Metab* 104(4):627-630.
9. Sun Y, Qi X, & Grabowski GA (2003) Saposin C is required for normal resistance of acid beta-glucosidase to proteolytic degradation. *J Biol Chem* 278(34):31918-31923.
10. Sidransky E (2012) Gaucher disease: insights from a rare Mendelian disorder. *Discov Med* 14(77):273-281.
11. Grabowski GA (2008) Phenotype, diagnosis, and treatment of Gaucher's disease. *Lancet* 372(9645):1263-1271.

12. Hruska KS, LaMarca ME, Scott CR, & Sidransky E (2008) Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Hum Mutat* 29(5):567-583.

13. Charrow J*, et al.* (2000) The Gaucher registry: demographics and disease characteristics of 1698 patients with Gaucher disease. *Arch Intern Med* 160(18):2835-2843.

14. Taddei TH*, et al.* (2009) The underrecognized progressive nature of N370S Gaucher disease and assessment of cancer risk in 403 patients. *Am J Hematol* 84(4):208-214.

15. Horowitz M*, et al.* (1993) Prevalence of nine mutations among Jewish and non-Jewish Gaucher disease patients. *Am J Hum Genet* 53(4):921-930.

16. Stone DL*, et al.* (2000) Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. *Hum Mutat* 15(2):181-188.

17. Koprivica V*, et al.* (2000) Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease. *Am J Hum Genet* 66(6):1777-1786.

18. Eto Y & Ida H (1999) Clinical and molecular characteristics of Japanese Gaucher disease. *Neurochem Res* 24(2):207-211.

19. Jeong SY, Park SJ, & Kim HJ (2011) Clinical and genetic characteristics of Korean patients with Gaucher disease. *Blood Cells Mol Dis* 46(1):11-14.

20. Mistry PK*, et al.* (2010) Glucocerebrosidase gene-deficient mouse recapitulates Gaucher disease displaying cellular and molecular dysregulation beyond the macrophage. *Proc Natl Acad Sci U S A* 107(45):19473-19478.

21. Offman MN, Krol M, Silman I, Sussman JL, & Futerman AH (2010) Molecular basis of reduced glucosylceramidase activity in the most common Gaucher disease mutant, N370S. *J Biol Chem* 285(53):42105-42114.

22. Liu J*, et al.* (2012) Gaucher disease gene GBA functions in immune regulation. *Proc Natl Acad Sci U S A* 109(25):10018-10023.

23. Bendikov-Bar I, Ron I, Filocamo M, & Horowitz M (2011) Characterization of the ERAD process of the L444P mutant glucocerebrosidase variant. *Blood Cells Mol Dis* 46(1):4-10.

24. Sun QY*, et al.* (2010) Glucocerebrosidase gene L444P mutation is a risk factor for Parkinson's disease in Chinese population. *Mov Disord* 25(8):1005-1011.

25. Berent SL & Radin NS (1981) Mechanism of activation of glucocerebrosidase by co-beta-glucosidase (glucosidase activator protein). *Biochim Biophys Acta* 664(3):572-582.

26. Atrian S*, et al.* (2008) An evolutionary and structure-based docking model for glucocerebrosidase-saposin C and glucocerebrosidase-substrate interactions - relevance for Gaucher disease. *Proteins* 70(3):882-891.

27. Qi X, Qin W, Sun Y, Kondoh K, & Grabowski GA (1996) Functional organization of saposin C. Definition of the neurotrophic and acid beta-glucosidase activation regions. *J Biol Chem* 271(12):6874-6880.

28. Weiler S, Kishimoto Y, O'Brien JS, Barranger JA, & Tomich JM (1995) Identification of the binding and activating sites of the sphingolipid activator protein, saposin C, with glucocerebrosidase. *Protein Sci* 4(4):756-764.

29. Offman MN*, et al.* (2011) Comparison of a molecular dynamics model with the X-ray structure of the N370S acid-beta-glucosidase mutant that causes Gaucher disease. *Protein Eng Des Sel* 24(10):773-775.

30. Zubrzycki IZ, Borcz A, Wiacek M, & Hagner W (2007) The studies on substrate, product and inhibitor binding to a wild-type and neuronopathic form of human acid-beta-glucosidase. *J Mol Model* 13(11):1133-1139.

31. Lieberman RL (2011) A Guided Tour of the Structural Biology of Gaucher Disease: Acid-beta-Glucosidase and Saposin C. *Enzyme Res* 2011:973231.

32. Dvir H*, et al.* (2003) X-ray structure of human acid-beta-glucosidase, the defective enzyme in Gaucher disease. *EMBO Rep* 4(7):704-709.

33. de Vries SJ & Bonvin AM (2011) CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One* 6(3):e17695.

34. Salvioli R*, et al.* (2005) The N370S (Asn370-->Ser) mutation affects the capacity of glucosylceramidase to interact with anionic phospholipid-containing membranes and saposin C. *Biochem J* 390(Pt 1):95-103.

35. Ritchie DW & Kemp GJL (2000) Protein docking using spherical polar Fourier correlations. *Proteins-Structure Function and Genetics* 39(2):178-194.

36. de Vries SJ, van Dijk M, & Bonvin AM (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc* 5(5):883-897.

37. Rossmann M*, et al.* (2008) Crystal structures of human saposins C andD: implications for lipid recognition and membrane interactions. *Structure* 16(5):809-817.

38. Stein SAM, Loccisano AE, Firestine SM, & Evanseck JD (2006) Principal Components Analysis: A Review of its Application on Molecular Dynamics Data. *Annual Reports in Computational Chemistry, Vol 2* 2:233-261.

39. Doerr S, Ariz-Extreme I, Harvey MJ, & De Fabritiis G (2017) Dimensionality reduction methods for molecular simulations. (arXiv:1710.10629v2 [stat.ML]).

40. Bhowmik D, Young MT, Gao S, & Ramanathan A (2018) Deep clustering of protein folding simulations *BMC Bioinformatics* accepted.

41. Henzler-Wildman K & Kern D (2007) Dynamic personalities of proteins. *Nature* 450(7172):964-972.

42. Noe F, Schutte C, Vanden-Eijnden E, Reich L, & Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci U S A* 106(45):19011-19016.

43. Ritchie DW (2003) Evaluation of protein docking predictions using Hex 3.1 in CAPRI rounds 1 and 2. *Proteins* 52(1):98-106.

44. Ritchie DW (2008) Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci* 9(1):1-15.

45. Case DA*, et al.* (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26(16):1668-1688.

46. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111(27):7812-7824.

47. Oostenbrink C, Soares TA, van der Vegt NF, & van Gunsteren WF (2005) Validation of the 53A6 GROMOS force field. *Eur Biophys J* 34(4):273-284.

48. Michaud-Agrawal N, Denning EJ, Woolf TB, & Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* 32(10):2319-2327.

49. Doersch C (2016) Tutorial on variational autoencoders. (arXiv:1606.05908 [stat.ML]).

50. Duan M, Fan J, Li M, Han L, & Huo S (2013) Evaluation of Dimensionality-reduction Methods from Peptide Folding-unfolding Simulations. *J Chem Theory Comput* 9(5):2490-2497.

Tables

Table 1: Summary of the Coarse Grained simulations. Five different systems were inserted into the membrane via self-assembly simulations. They include (1) GCase, (2) GCase bound to GluCer, (3) GCase bound to SAPC and GluCer, and SAPC in (4A) closed and (4B) open conformations.

Simulation	System	PDB ID	Length (μs)	DPPC	Waters
1: GG	GCase	1OGS	1.20	300	5000
2: CG	GCase + GluCer	1OGS	1.20	338	6431
3: CG	CPX	2NSX + 2GTG (pose 5)	1.20	414	8500
4A: CG	SAPC (closed)	2GTG	1.20	250	4000
4B: CG	SAPC (open)	2QYP	1.20	250	4000

Table 2: List of atomistic molecular dynamics (AT-MD) simulations.

Simulation	*System*	*a*	*b*	*N.Atoms*	*Time (ns)*
1	GCase	-	inactive	20070	500
2a	GCase + GluCer	active		22084	1000
2b	GCase + GluCer		inactive	22082	1000
3a	CPX	active		26540	1000
3b	CPX		inactive	26537	1000
4	SAPC	-	-	13242	500
5a	CPX (N370S)	active	-	26536	1000
5b	CPX (N370S)		inactive	26536	1000
6a	CPX (L444P)	active	-	26535	1000
6b	CPX (L444P)	-	inactive	26535	1000

FIGURE LEGENDS

Figure 1 The Predicted GCase–SAPC Interface. (**A**) Residues in red are those considered by CPORT to take part in protein-protein binding, and those marked in blue can potentially intervene in the binding. Protein-protein binding site was identified over helix-7 of Domain III, flanked by helix-6 and Domain II**. (B)** (i) Superimposition of the poses of GCase in complex with SAPC in open (green) and closed (cyan) conformations. GCase in complex with (ii) open SAPC (green) or with (iii) closed SAPC (cyan) and GluCer (orange spheres). GCase has been illustrated in surface representation and is colored in light brown; GluCer is colored in orange.

Figure 2 A Hydrogen Bond Between GCase^{D315} and SAPC^{K33} Maintains Helical Conformation of Loop-1. (**A**) Shown are snapshots of conformations extracted from simulation 3a (active complex) at (i) 0 ns, (ii) 500 ns and (iii) 1000ns. (Colors: GCase, blue; SAPC, green). Comparison of conformations adopted by Loop-1 in simulation 2a (GCase, orange) has been made at equivalent time and superimposed on that of 3a. (**B**) Comparison of conformations adopted by Loop-1 in simulations 2b (no SAPC, red) and 3b (inactive complex, yellow) at (i) 0 ns, (ii) 500 ns and (iii) 1000 ns. Loop-1 in simulation 2b extends towards helix-7. The interaction of residue SAPC^{K34} with the neighboring GCase^{D315} in simulation 3b influences Loop-1 to adopting a helical conformation.

Figure 3 Conformation of the Loops at the Entrance of the Binding Site. (**A**) Conformation of Loop-2 in simulations 2a (active GCase, no SAPC, orange), 3a (active complex, yellow) and 3b (inactive complex, blue) at 1000ns. (**B**) Snapshot of GCase–SAPC (green) complex at 1000 ns in simulation 3b. SAPC stabilizes the active form of the Loop-2, where residue GCase^{W348} is tucked in a hydrophobic pocket formed in SAPC. (**C**) Conformation of Loop-3, highlighting the orientation of side chains of R395–E340 in different simulations at 1000 ns. (**D**) Distance

between specific atoms in the side chains of residues R395 and E340 of GCase in simulations 2a, 2b, 3a and 3b (as shown).

Figure 4 Protein-protein interactions in Simulation 3a (active complex) at 1000 ns. (**A**) Loop-1 and helix-7, (**B**) Loop-2, (**C** and **D**) Domain II are shown. SAPC is colored in green and interacting residues in GCase are colored blue. Position of residue N370 has been represented with spheres and colored in cyan. Distances of the interactions, over the course of the simulation, have been illustrated in SI Appendix, Figure S9.

Figure 5 GCase Loop Dynamics. **Left column:** Comparison of loop conformations at the entrance of the binding site in simulations 3a (active complex, blue), 5a ($GCase^{N370S}$ active state, pink) and 6a ($GCase^{L444P}$ active state, cyan). Loop-1, Loop-2, and Loop-3 are illustrated. (**A**) Loop-1 maintains the helical conformation due to the influence of SAPC. (**B**) Due to the instability of the protein–protein binding, W348 (Loop-2) does not remain inserted in the hydrophobic pocket in the mutants. (**C**) Loop-3 closes towards the binding site in simulation 6a. **Right column:** Dynamic evolution of the loops at the entrance of the active site in simulations 3b (inactive complex, yellow), 5b ($GCase^{N370S}$ inactive state, grey) and 6b ($GCase^{L444P}$ inactive state, purple). (**D**) Loop-1 extends towards helix-7 in the mutants. (**E**) Poor binding between the two proteins prevents residue W348 from occupying the hydrophobic pocket in SAPC. (**F**) Loop-3 adopts a closed conformation in the mutants. Snapshots taken at 1000 ns.

Figure 6 Differences in GCase Conformational States Identified By Deep Learning. (**A**) The root-mean-square (RMS) loss between the training and validation datasets from the trajectories (2a, 2b, 3a, 3b) modeled using the CVAE are shown. The optimum number of dimensions is determined as 14, based on the RMS-loss metric, beyond which the RMS-loss of the validation set is larger than the training data. The inset represents the RMS-loss with

respect to the mutant simulations (5a, 5b, 6a, 6b) – notably, the mutant simulatons have a higher RMS-loss value indicating distinct differences in the conformational motions sampled by the MD simulations. (**B**) Histogram of the the distance between residues E340 and R395 occupy three distinct peaks, indicative of at least three conformational states sampled in the MD simulations. (**C**) CVAE learned representation of the conformational motions embedded in a 3D-space using t-SNE (see Methods) depicting three distinct conformational states. (**D**) Cartoon representations of the three states shown as an ensemble. Ensemble members were picked with respect to the peaks in (B), illustrating E340 and R395 residues as red spheres for easy identification. Notably, the three ensembles highlight the separation between E340 and R395, indicating the distinct conformations of Loop-1 (yellow), Loop-2 (green) and Loop-3 (purple) in open state of GCase.