Supporting Materials:

Structural determinants of water permeation through the sodium-galactose transporter vSGLT

Joshua L. Adelman[1], Ying Sheng[1], Seungho Choe[2], Jeffrey Abramson[3], Ernest Wright[3], John M. Rosenberg[1] and Michael Grabe[1,4]

[1] *Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA*
[2] *School of Basic Science, College of Convergence, Daegu Gyeongbuk Institute of Science & Technology, Daegu, Korea*
[3] *Department of Physiology, University of California, Los Angeles, Los Angeles, CA*
[4] *Cardiovascular Research Institute, Department of Pharmaceutical Chemistry, University of California, San Francisco, CA*

Contents

1. System Preparation

Atomistic models of monomeric vSGLT embedded in a lipid bilayer were constructed using chain A from the x-ray structure deposited in the RCSB Protein Data Bank (PDB ID: 3DH4) (1). The first 52 residues of vSGLT including the unassigned section of the first transmembrane helix (TM-1) and the unresolved loop connecting it to TM1, were removed. These residues do not constitute part of the core structure of the transporter and the corresponding TM is not conserved among superfamily members. The MODELLER software package (2), was then used to mutate any non-wild type residues back to their WT sequence and rebuild the side chains of residues K124, R273, K454 and K547, which were missing in the 3DH4 structure. Additionally, the loop connecting TM4 and 5 (residues 179 to 184) was rebuilt using MODELLER's loop modeling routine.

Two independent sets of models were constructed that differed only in the random seed used during loop modeling and the number of flanking residues on either side of the missing loop that were permitted to move. One model was chosen from each of the two independent ensembles of putative structures based on screening using the MODELLER DOPE score and analysis with MolProbity (3). These two final models were used to construct the full systems described below.

The protein along with the galactose substrate and crystallographic sodium, were oriented with respect to the z-axis using OPM (Orientations of Proteins in Membranes) (4) and then inserted into POPE (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylethanolamine) membrane using the CHARMM-GUI Membrane Builder (5). The embedded protein with membrane was then solvated in a rectangular box with dimensions 97 x 97 x 97 $Å^3$ containing approximately 91,000 atoms. The system was neutralized using sodium and chloride counterions, which were added in sufficient quantity to approximate a physiological concentration of 150 mM.

2. Simulations

2.1 Equilibration

Each system was prepared using NAMD (6) to obtain well-equilibrated starting conformations for the production runs on the Anton special-purpose supercomputer (7). Simulations were carried out using the CHARMM22 parameter set (8) for the protein with CMAP corrections (9). The membrane was parameterized using the CHARMM36 lipid forcefield (10) and galactose parameters were taken from the CHARMM force field for pyranose monosaccharides (11). Explicit water molecules were included using the three-site TIP3P model (12).

The systems were minimized using 10,000 steps of the conjugate gradient method followed by gradual heating from 1 K to 310 K at a rate of 1 K every 400 fs using temperature reassignment. During the heating phase the dynamics were carried in the constant volume ensemble with all water molecules, galactose, Na^+, and heavy backbone and sidechain atoms restrained using a harmonic potential with a 10.0 kcal/mol/$Å^2$ force constant. Subsequent dynamics were performed with a Langevin piston barostat with a 200 fs piston period and 100 fs piston decay constant to maintain a constant pressure of 1 atm. Temperature was maintained at 310 K using Langevin dynamics with a 0.5 ps^{-1} damping coefficient. Restraints on the water molecules and side-chain heavy atoms were then removed over the next 2.1 ns. Finally, restraints on the backbone, Na^+ and galactose were removed over another 1.8 ns. During the entire equilibration protocol, an external force was

applied to all water molecules outside of the transporter to prevent aberrant hydration of the protein-membrane interface. All further dynamics were simulated in the absence of external restraints.

Bond lengths between hydrogen and heavy atoms were constrained using the SHAKE algorithm and water was held rigid using the SETTLE algorithm. This permitted us to use a discretized time step of 2 fs. The Particle Mesh Ewald summation was employed for electrostatics with a grid spacing of less than 1 *Å* along each dimension. Both van der Waals and electrostatic interactions were smoothly switched off between 8 and 10 *Å*.

2.2 Production simulations on Anton

Productions simulations were run with Anton software version 2.4.1 (7). Coordinates and velocities for each setup prepared in NAMD were converted to the Anton specified format. The same force field parameters were used in both the preparatory and production runs. Additional parameters for the Anton runs were obtained using the default outputs of the `guess_chem`, `refine_sigma` and `subboxer` pre-processor scripts to tune the cut-off distances, electrostatic settings and spatial domain decomposition. Bonds between hydrogen and heavy atoms were constrained to their equilibrium length using the M-SHAKE algorithm (13). Long-range electrostatic interactions were treated using the Gaussian split Ewald (GSE) method with a 64x64x64 mesh (14). The equations of motion were integrated using the RESPA multiple time-step method (15) with time steps of 6.0 fs for the long-range electrostatic interactions and 2.0 fs for all other interactions. Simulations were run at constant temperature (300 K) and pressure (1 atm). Constant pressure was maintained using a semi-isotropic Berendsen barostat with a relaxation time of 2 ps and temperature was controlled with the corresponding Berendsen thermostat with $\tau = 1$ ps.

3. Analysis

3.1 Calculation of permeability and diffusion coefficients

The diffusion permeability, p_d, was computed using the linear flux equation

$$J_t = p_d\,(t_o - t_i)\,. \tag{1}$$

Here, J_t (mol/s) is the flux of tracked water molecules permeating from one side of the membrane to the other and t_i and t_o are the inner and outer tracer concentrations respectively. The flux of water is calculated directly from counting permeation events through the transporter as described in the main text. In this study the channel region extended 30 Å in the z-direction, which is longer than the definition employed in Ref. (16), where the channel boundaries were defined at $\pm$ 7.5 Å from the center of the transporter. Using the larger boundary separation has the effect of decreasing the number of observed permeation events by approximately a factor of two, but ensures that waters must make a complete transition from one bulk region to the other and minimizes recrossing events. Given a molar concentration of water, $c_w \sim 0.055$ mol/cm^3, the diffusion permeability is then calculated from Eq. 1 as

$$p_d = \frac{\text{counts}_{\text{up}} + \text{counts}_{\text{down}}}{2 c_w T_{\text{total}} N_A}, \tag{2}$$

where N_A is Avogadro's Number and T_{total} is the total time of the simulation.

In the absence of an osmotic concentration or pressure gradient across the lipid membrane in our simulations, we estimate the osmotic permeability, p_f using the theoretical framework proposed by Zhu et al. (17). Briefly, this calculation is based on calculating the total incremental change in the position of all water molecules in the channel at time t

$$dn(t) = \sum_{i=1}^{M} \frac{z_i(t+dt) - z_i(t)}{L}, \tag{3}$$

where L is the length of the channel and $z_i(t)$ is the z component of the position of the ith water molecule's oxygen atom. After numerically integrating Eq. 3, the osmotic permeability is calculated from a linear regression of

$$\left\langle n(t)^2 \right\rangle = 2\frac{p_f}{v_w} t, \tag{4}$$

where v_w is the average volume of a single water molecule (18 cm^3/mol/N$_A$ $\approx 3\times10^{-23}$ cm^3) and the angular brackets denote an ensemble average over trajectory segments. In this study, trajectory segments were obtained by dividing the time series of $n(t)$ into non-overlapping windows 300 ps in length. Uncertainties in the calculated value of p_f were computed using a Monte Carlo bootstrapping procedure. The value of $n(t)$ for each trajectory segments was first averaged in blocks of 80 consecutive windows, and then a bootstrap sample was created from the block averages using 10,000 samples drawn at random with replacement. This sample was then used to determine the 95% confidence interval for the calculated value of p_f. A similar procedure was used to calculate the confidence interval for p_d.

The water channel in vSGLT is not aligned with the z-axis of the simulation box and is instead curved as it winds from the extracellular to intracellular side of the transporter. We therefore modify the protocol for estimating p_f as follows. After centering the protein at the origin of the simulation box, the water molecules were wrapped into the central periodic image; the protein was then aligned onto a reference conformation. Subsequently, to define the permeation path, the center-of-mass of water molecules within cylindrical slabs 0.5 Å in thickness, and with a radius of 20 Å from the center-of-mass of the transporter, were calculated for $-15 < z < 15$ Å. A smooth continuous path, φ, is generated by fitting

$$\varphi_\alpha(\lambda) = \varphi_0 + (\varphi_N - \varphi_0)\,\lambda + \sum_{i=1}^{N_{\text{dim}}} \left[\sigma_{i,0}\sin(\pi\lambda) + \sigma_{i,1}\sin(2\pi\lambda)\right]\cdot\hat{e}_i, \tag{5}$$

through the center-of-mass of water molecules within each slab by varying λ over the range $[0,1]$ (18). Here, $N_{\mathrm{dim}} = 3$, $\hat{e}_i$ is the unit vector of the i^{th} dimension and $\sigma_{i,j}$ are the coefficients of 2 sinusoidal basis functions in each dimension. The parameters $\sigma_{i,j}$ and λ_α are selected to minimize the difference between the center-of-mass calculated for each slab, α and the parameterized curve,

$$\chi^2 = \sum_{\alpha=0}^{N_{slab}} |\varphi_\alpha(\lambda_\alpha) - \varphi_\alpha|^2 . \tag{6}$$

Equation 3 is then reformulated in terms of the displacement along the curve, rather than a displacement along z, as

$$dn(t) = \sum_{i=1}^{M} \frac{s_i(\lambda_i(t), \lambda_i(t+dt))}{L} \tag{7}$$

where s is the contour length between $\lambda_i(t)$ and $\lambda_i(t+dt)$ along $\varphi(\lambda)$ and L is the contour length of the curved path traversing the entire channel. The projection $\lambda_i(t)$ minimizes the distance between the position of the i$^{\text{th}}$ water molecule in the channel at time t and the curve. Results using the projection obtained by identifying the point on the curve with the same z value of the oxygen atom at each time, as in (16), yielded similar results.

As with the calculation of p_d, defining the channel as the volume within the transporter between $-15 < z < 15$ Å decreases p_f by approximately a factor of 2 compared to the value obtained in Ref. (16) where the channel boundaries were between $-7.5 < z < 7.5$ Å. We find that, empirically, both permeability values scale with L in a similar manner between $L = 15$ and 30 Å. As such, the ratio of p_f/p_d for this system is largely independent of L over this range.

3.2 Estimating errors in permeation count statistics from low temporal resolution trajectories

The ability to accurately enumerate the number of water permeation events crossing the transporter is limited by the temporal resolution with which the system coordinates were recorded. At a 1 ps sampling frequency, each water molecule in the system only moves a small fraction of the total length of the transporter. This allows us to unambiguously demarcate when a specific water molecules moves from one side of the membrane to the other via the lumen of the transporter. At decreasing temporal resolution of the trajectory data, the distance that a water molecule can diffuse between observations approaches the length of the passage through vSGLT as well as the distance between the water-lipid interface and the periodic boundary. When those distances become comparable, a growing subpopulation of the crossing events will not be detected by our counting algorithm, leading to under counting of the true directional flux of water. A distribution of the crossing times for waters that move through the transporter through the central region, which is bounded by two planes separated by 30 Å along the z-axis, is shown in Fig. S1.

In addition to errors arising from *false negatives*, low temporal resolution of the trajectory data can result in *false positive* counts, wherein water molecules that move from one bulk region to the other through the periodic boundary are counted as crossing events due to transient excursions into the central region of the simulation box. While it is impossible to prevent under counting due to false negatives, counts arising from the second class of errors can be eliminated from our statistics. This is done by detecting the large jumps (> 50 Å) in the z position of a water between adjacent time steps due to a periodic boundary crossing that lead to an otherwise spurious crossing event.

To estimate the error in the counting statistics from using original trajectory data sampled at 100 ps intervals, we examine the subset of the trajectories that were re-sampled at 1 ps. Crossing statistics from the 1 ps data are considered to be a true measure of the number of permeation events in the simulation. We then down sample those trajectories, taking the coordinates of every hundredth frame and recalculate the number of permeation events while applying our filtering method to remove the false positive events. The time-dependent directional flux using the 1 ps data and the down sampled 100 ps data are shown in Fig. S2. The difference between the total number of counts using the same data sampled at different temporal resolutions is approximately 6 % and is consistent with the small fraction of the total crossing events that take on the order of the 100 ps observation interval.

4. Discussion

4.1 Discrepancies between reported osmotic permeabilities for vSGLT in the literature

The osmotic permeability, p_f, reported here (2.7×10^{-13} cm^3/s) is consistent with the value we reported previously (4.1×10^{-13} cm^3/s) based on a different set of simulations of the same inward-facing state of vSGLT. Meanwhile, the Tajkhorshid laboratory has reported a p_f value for simulations on the same conformation of the protein, which is much smaller (4.75×10^{-15} cm^3/s) (19). This discrepancy arises from a difference in the calculation of the net crossing events, $n(t)$, appearing in Eq. 4. As described above, we integrate the differential form of the collective variable (Eq. 7) to obtain $n(t)$ as originally described in Ref. 17. In this formalism, $n(t)$ is a function of all water molecules within the channel at time t and $t+dt$. Conversely, Li et al. (19) calculate $n(t)$ from the cumulative sum of the discrete efflux and influx permeation events (Dr. Emad Tajkhorshid, personal communication). Using this later method, $n(t)$ depends only on the dynamics of water molecules that fully permeate the transporter. On long timescales both methods provide qualitatively similar time series for $n(t)$, as can be seen by comparing Figures 2B and S3A in Ref. 16; however, on short timescales these two methods give different results. Specifically, $n(t)$ calculated as in Ref. (19) changes in integer steps, while $n(t)$ calculated from the collective coordinate varies continuously. When the averaging time is shorter than or comparable to the mean time

between permeation events, the discrete event method often results in $n(t)^2$ values that are nearly constant and close to zero. As a result, a linear regression fit of the the average slope of $n(t)^2$ calculated from the discrete permeation counts systematically yields a smaller estimate of p_f than the collective coordinate method. We note, however, that Eq. 4 is only applicable when t is much longer than the velocity correlation time of $n(t)$ (17). Analyzing the data presented here with the discrete event method employed in Ref. (19) also produces a much smaller value of $p_f = 9.5\times10^{-15}$ cm^3/s, suggesting that the water dynamics underlying the two studies are similar. Nonetheless, we believe that the collective coordinate method developed by Zhu et al. (17), and used here, is preferred as it has been validated against nonequilibrium simulations in the presence of a chemical potential difference (17) and experimental measurements (20–22).

References

[1] Faham, S., A. Watanabe, G. M. Besserer, D. Cascio, A. Specht, B. A. Hirayama, E. M. Wright, and J. Abramson, 2008. The crystal structure of a sodium galactose transporter reveals mechanistic insights into Na+/sugar symport. *Science* 321:810–4.

[2] Sali, A., and T. L. Blundell, 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815.

[3] Chen, V. B., W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, 2010. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12–21.

[4] Lomize, M. A., A. L. Lomize, I. D. Pogozheva, and H. I. Mosberg, 2006. OPM: orientations of proteins in membranes database. *Bioinformatics* 22:623–5.

[5] Jo, S., T. Kim, and W. Im, 2007. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLoS One* 2:e880.

[6] Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, 2005. Scalable molecular dynamics with NAMD. *J Comput Chem* 26:1781–802.

[7] Shaw, D. E., R. O. Dror, J. K. Salmon, J. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, et al., 2009. Millisecond-scale molecular dynamics simulations on Anton. *In* Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis. ACM, 39.

[8] MacKerell, A. D., D. Bashford, M. Bellott, R. Dunbrack, J. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. a. Ha, et al., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B* 102:3586–3616.

[9] Mackerell, A. D., Jr, M. Feig, and C. L. Brooks, 3rd, 2004. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* 25:1400–15.

[10] Klauda, J. B., R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell, Jr, and R. W. Pastor, 2010. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J Phys Chem B* 114:7830–43.

[11] Guvench, O., S. N. Greene, G. Kamath, J. W. Brady, R. M. Venable, R. W. Pastor, and A. D. Mackerell, Jr, 2008. Additive empirical force field for hexopyranose monosaccharides. *J Comput Chem* 29:2543–64.

[12] Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, 1983. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* 79:926.

[13] Kräutler, V., W. F. van Gunsteren, and P. H. Hünenberger, 2001. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of Computational Chemistry* 22:501–508.

[14] Shan, Y., J. L. Klepeis, M. P. Eastwood, R. O. Dror, and D. E. Shaw, 2005. Gaussian split Ewald: A fast Ewald mesh method for molecular simulation. *J Chem Phys* 122:54101.

[15] Tuckerman, M., B. J. Berne, and G. J. Martyna, 1992. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics* 97:1990.

[16] Choe, S., J. M. Rosenberg, J. Abramson, E. M. Wright, and M. Grabe, 2010. Water permeation through the sodium-dependent galactose cotransporter vSGLT. *Biophys J* 99:L56–8.

[17] Zhu, F., E. Tajkhorshid, and K. Schulten, 2004. Collective diffusion model for water permeation through microscopic channels. *Phys Rev Lett* 93:224501.

[18] Zhu, F., and G. Hummer, 2010. Pore opening and closing of a pentameric ligand-gated ion channel. *Proc Natl Acad Sci U S A* 107:19814–9.

vSGLT	hSGLT1
S76	T90
Y87	F101
N245	G272
D336	N363
Q422	Q451
F424	F453
Q425	D454
Q428	Q457

Table S1. Residues involved in modulating water flow through the outer-gate of vSGLT and the corresponding residues in hSGLT1 based on the sequence alignment in (1)

[19] Li, J., S. A. Shaikh, G. Enkavi, P.-C. Wen, Z. Huang, and E. Tajkhorshid, 2013. Transient formation of water-conducting states in membrane transporters. *Proc Natl Acad Sci U S A* 110:7696–701.

[20] Aksimentiev, A., and K. Schulten, 2005. Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophys J* 88:3745–61.

[21] Hashido, M., M. Ikeguchi, and A. Kidera, 2005. Comparative simulations of aquaporin family: AQP1, AQPZ, AQP0 and GlpF. *FEBS Lett* 579:5549–52.

[22] Jensen, M. Ø., and O. G. Mouritsen, 2006. Single-channel water permeabilities of Escherichia coli aquaporins AqpZ and GlpF. *Biophys J* 90:2270–84.

[23] Scargle, J. D., J. P. Norris, B. Jackson, and J. Chiang, 2013. Studies in Astronomical Time Series Analysis. VI. Bayesian Block Representations. *The Astrophysical Journal* 764:167.

[24] Vanderplas, J., A. Connolly, Ž. Ivezić, and A. Gray, 2012. Introduction to astroML: Machine learning for astrophysics. *In* Conference on Intelligent Data Understanding (CIDU). 47 –54.

[25] Efron, B., and R. Tibshirani, 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 54–75.

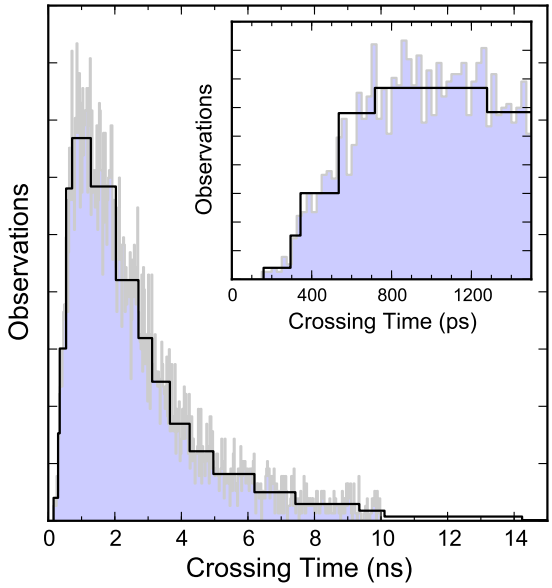

Figure S1. Distribution of crossing times for waters moving between the two planes defining the central region of the simulation box, which contains the lipid membrane and transporter. The crossing times are first partitioned using a uniform bin size of 25 ps (blue) and then using a set of variable bins determined by the Bayesian blocks approach (black) (23, 24). (Inset) The same data plotted over crossing times less than 1.5 ns.

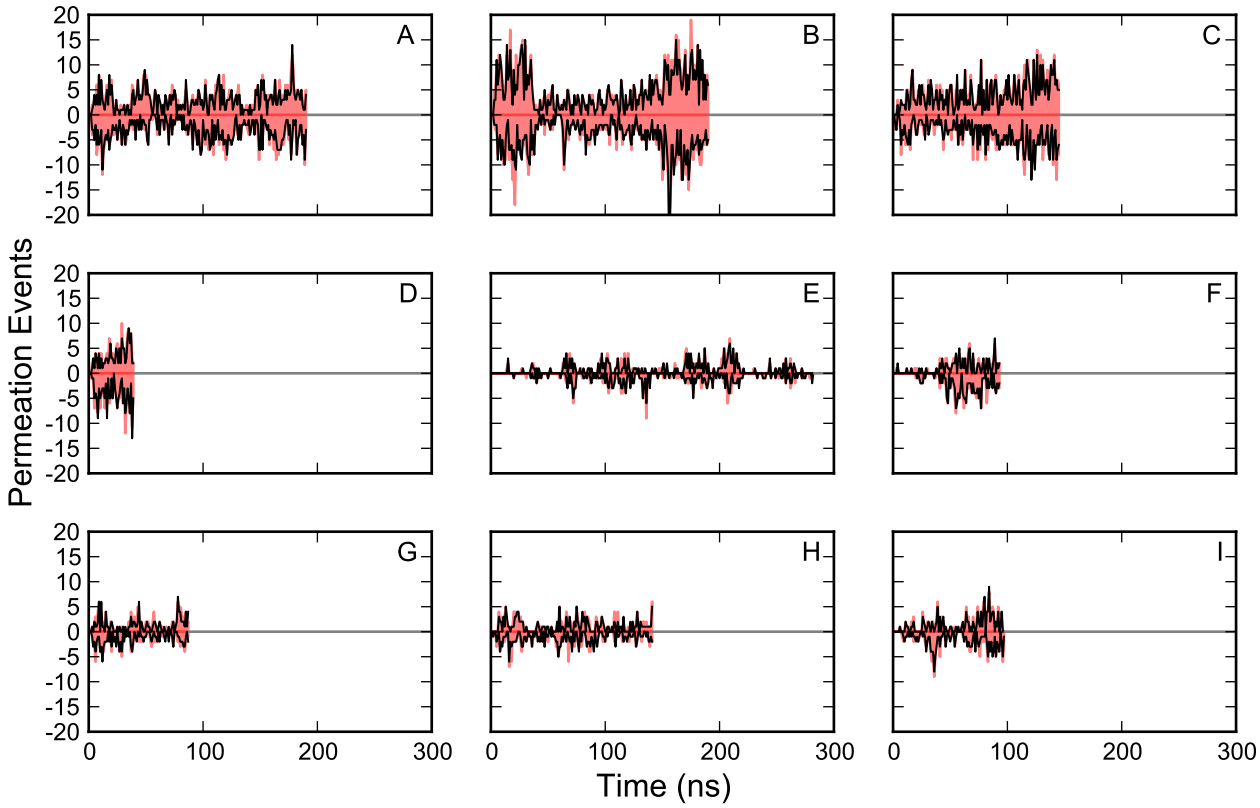

Figure S2. Comparison of time-dependent direction fluxes of water calculated from trajectories sampled at a 1 ps interval (red) and the same data down sampled to 100 ps between frames (black). Counts arising from efflux of water through the transporter is shown with positive values, and influx as negative values.

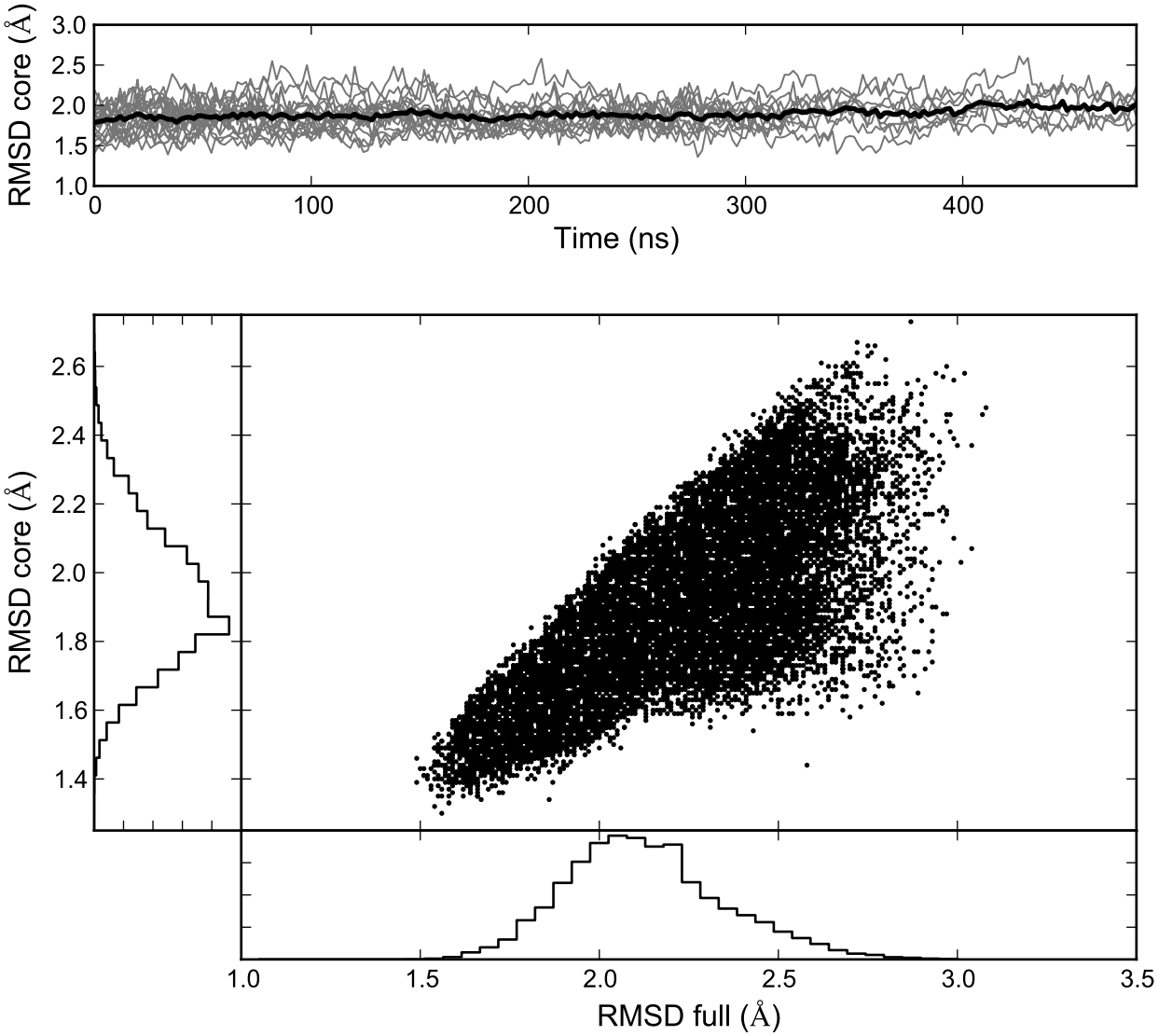

Figure S3. Cα RMSD of vSGLT. (Upper) Time series of the Cα RMSD from the crystallographic structure (PDB: 3DH4) of the ten core transmembrane helices for the 21 independent trajectories. The individual traces for each simulation and the average RMSD for the ensemble are shown in gray and black, respectively. (Lower) The Cα RMSD for core 10 TMs and the full transporter for all conformations visited in the aggregate ensemble. The probability distribution for each RMSD metric is shown adjacent to the central scatter plot, along its respective axis.

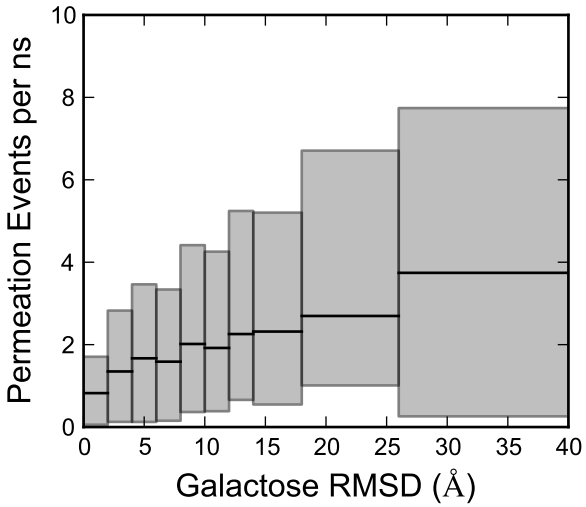

Figure S4. Galactose position dependent flow of water through vSGLT. Time series of the average inward and outward flow were calculated along with the corresponding RMSD of the galactose from its crystallographic pose after superimposing the transporter. After averaging the both quantities in 2 ns blocks, the blocks were partitioned into bins based on the mean RMSD in the block. The average flow in each bin is shown as a black line with the grey box denoting the 95 % confidence interval calculated using Monte Carlo bootstrapping (25).