Supporting Information

Adelman et al. 10.1073/pnas.1525378113

SI Materials and Methods

Markov State Model of Na$^+$ Binding/Unbinding Process. A Markov state model (MSM) of Na$^+$ binding/unbinding from the Na2 site of vSGLT was constructed using the solvent-shells framework first introduced by Harrigan et al. (14). MSMs reduce the complex dynamics of a process in the fully atomistic phase space, into a set of transition rates between discrete states, parameterized by the simulation data. Here, the movement of Na$^+$ in and out of the Na2 site is described by transforming the trajectory data into the occupancy of Na$^+$ in a set of concentric shells surrounding 13 key residues in and around the site. Specifically, we selected the backbone carbonyl oxygen atoms of A62, A63, I65, V185, and A361, the side-chain hydroxyl oxygens of S66, Y176, S183, T188, S364, S365, and S368, and the two carboxyl oxygens of D189 based on inspection of the most prevalent contacts sodium made with the transporter in the region of the Na2 site during the simulations. Around each of the selected atoms, we define $N_{\text{shells}} = 4$ shells extending radially from its center, $dr = 1.5$ Å in width. The instantaneous Na$^+$ density in each shell specifies a vectorial representation of each snapshot with dimensions $N_{\text{solute}} \cdot N_{\text{shells}}$ as follows:

$$\Theta_{(i-1)N_{\text{shells}}+j} = \frac{1}{4\pi r_{i,j}^{mid} dr} \sum_{k}^{N_{\text{sod}}} \mathbf{I}\left(r_{i,j} \le d_{i,k} \le r_{i,j} + dr\right), \qquad \text{[S1]}$$

where $i \in [1, N_{\text{solute}}]$ is the index of the solute atom (here the oxygen atoms described above), $j \in [1, N_{\text{shells}}]$ is the index of the shell, and $k \in [1, N_{\text{sod}}]$ specifies the Na$^+$ ion. Each shell j is parameterized by the radial distance, $r_{i,j}$ from solute atom i, and extends from $r_{i,j}$ to $r_{i,j} + dr$ with a midpoint, $r_{i,j}^{mid} = r_{i,j} + dr/2$. The distance between Na$^+$ k and solute atom i, $d_{i,k}$, parameterizes the indicator function $\mathbf{I}$, which equals 1 if Na$^+$ ion k is in the jth shell of solute atom i and zero otherwise. All of the elements of $\boldsymbol{\Theta}$ are therefore zero if none of the sodium ions in the simulation are within 6 Å of any of the 14 solute oxygen atoms, representing an apo Na2 site.

After transforming each of the trajectories into its solvent shell representation, we applied time-independent component analysis (tICA) to further reduce the dimensionality of the model space before constructing the MSM (15, 16). The transformation of the trajectories was performed using the WetMSM plugin (14) for MSMBuilder3 (17). All subsequent model building and analysis was also performed using MSMBuilder3 in concert with MDTraj (41).

We then built a large number of candidate MSMs by varying the set of model parameters, namely the number of microstates (clusters), tICA components, and both the lag time of the MSM and the lag time used in determining the tICA components. The grid of parameters is shown in Table S1, and iterating over all combinations of parameters resulted in 576 distinct models. These models were evaluated with regard to the convergence of their implied timescales, their generalized matrix Rayleigh quotient (GMRQ) (42) using threefold cross-validation, as well as the convergence of the first passage time distribution of Na$^+$ escape from the Na2 site determined via kinetic Monte Carlo simulations.

To select a model from among the 576 candidates, we first examined the implied timescales $\hat{t}_i$ across the range of lag times, τ, computed from the eigenvalues of each MSMs transition probability matrix, $\mathbf{T}(\tau)$, as follows:

$$\hat{t}_i = -\frac{\tau}{\ln \lambda_i(\tau)}, \qquad \text{[S2]}$$

where $\lambda_i(\tau)$ is the ith eigenvalue of $\mathbf{T}(\tau)$. The implied timescales of a model tend to plateau, becoming approximately constant for long lag times when the model's dynamics are Markovian. We therefore discarded all models whose slowest implied timescales were still increasing over the entire range of selected lag times. Furthermore, for all parameter sets, lag times less than ∼5 ns were in the rising transient phase of the implied timescale curves so models with lag times <5 ns, were also not considered (Fig. S6).

Finally, threefold cross validation using the rank-3 GMRQ objective function was used to minimize overfitting of the model (42). Models with less than 200 states tended to provide a better balance between the test and training scores.

Generating Long Escape Trajectories from a MSM Using Kinetic Monte Carlo. From each of the MSMs that showed a good balance between the training and test GMRQ scores as well as well-behaved implied timescales, we used a kinetic Monte Carlo scheme to generate long Na$^+$ escape trajectories to determine the first passage time distribution of this process. For each model we evaluated, we first mapped the crystallographic conformation of Na$^+$ in the Na2 site to one of the MSM microstates. Starting from that microstate, we then generated a long trajectory by sampling the transition probability matrix $\mathbf{T}$ of the MSM as follows. For a MSM in state i at time t, the probability of transitioning to state j at time $t + \tau$ is given by the element of $\mathbf{T}$, $T_{i,j}$. To propagate the model forward one time step τ, we generate a uniformly distributed random number, R on the interval [0,1], and select the state j if $\sum_1^j T_{i,j} < R < \sum_1^{j+1} T_{i,j}$. This process is repeated until we reach the state corresponding to the fully apo Na2 site where all of the Na$^+$ ions in the system are in the bulk and >6 Å away from any of the solute atoms. The time required to reach this state is recorded, and then a new trajectory is initiated starting at $t = 0$ from the microstate corresponding to the crystallographic conformation. We repeated this procedure 30,000 times to construct an estimate of the first passage time distribution Na$^+$ escape from each model.

Distributions of first passage times for the tested models are shown in Fig. S7. For each lag time examined, the distributions show a high degree of similarity, independent of variations in the other hyperparameters (number of microstates, number of tICA components, and tICA lag time). In agreement with the implied timescale convergence behavior, we observe little difference in the overall shape of the first passage time distributions calculated at lag times of 10 and 20 ns, whereas models with a 5-ns lag time produce escape time distributions displaying slightly faster kinetics. We therefore selected from among the models with a 10-ns lag time to maximize the amount of total sampling used.

Mathematical Model of SGLT1 Transport. The set of differential equations for the six-state kinetic model is described in the manuscript by Parent et al. (2). For the five-state model presented here, the related equations are as follows:

$$\begin{aligned}
\frac{dC_1}{dt} &= -(k_{16} + k_{12})C_1 + k_{21}C_2 + k_{61}C_6 \\
\frac{dC_2}{dt} &= k_{12}C_1 - (k_{21} + k_{23})C_2 + k_{32}C_3 \\
\frac{dC_3}{dt} &= k_{23}C_2 - (k_{32} + k_{34})C_3 + k_{43}C_4 \\
\frac{dC_4}{dt} &= k_{34}C_3 - (k_{43} + k_{46})C_4 + k_{64}C_6 \\
\frac{dC_6}{dt} &= k_{16}C_1 + k_{46}C_4 - (k_{64} + k_{61})C_6,
\end{aligned} \qquad \text{[S3]}$$

where C_i is the occupancy probability for state i, and $C_1 + C_2 + C_3 + C_4 + C_6 = 1$. Transitions between states C_i and C_j are represented by rate constants $k_{ij}: C_i \rightarrow C_j$. Eyring rate theory is used to describe the dependence of rate constants on membrane potential (V):

$$k_{ij}=k_{ij}^{o}exp\left(-\epsilon_{ij}FV/RT\right), \qquad \textbf{[S4]}$$

where k_{ij}^{o} is a voltage-independent rate, ϵ_{ij} is the equivalent charge movement (up to the transition state between $C_i \rightarrow C_j$), and F, R, and T have their usual physicochemical meanings (2). Na^+ and sugar binding on external and internal membrane surfaces is represented by rate constants:

$$\begin{aligned} k_{12} &= k_{12}^{o}[Na]_o^2 \exp\left(-\epsilon_{12}FV/RT\right) \\ k_{23} &= k_{23}^{o}[\alpha MDG]_o \\ k_{64} &= k_{64}^{o}[Na]_i^2[\alpha MDG]_i. \end{aligned}$$

The current (I_{ij}) associated with each transition $C_i \rightleftarrows C_j$ is given by $I_{ij}=e(\epsilon_{ij}+\epsilon_{ji})(k_{ij}C_i-k_{ji}C_j)$, where e is elementary charge (2). Total current (I) associated with SGLT1 is $I=N_T(I_{12}+I_{16}+I_{23}+I_{34}+I_{46})$, and N_T is the total number of transporters in the oocyte plasma membrane.

Computer simulations for the five- and six-state models were performed using Berkeley Madonna 8.0.1. The predictions of the models were examined using voltage pulse protocols for comparison with the electrophysiological experiments. The voltage pulse protocol was simulated by determining the occupancy probabilities at a given holding potential ($V_h = -50$ mV). At each test voltage (V_t ranging from +50 and −150 mV), the time course of the occupancy probabilities was obtained by numerically integrating the system of differential equations using the Runge–Kutta method. Cotransporter currents and steady-state current–voltage relations were calculated as functions of extracellular and intracellular Na^+ and glucose concentrations.

The voltage-independent rate constants (k_{ij}^{o}) and parameters for the five-state model are provided in Table S2. The base rate k_{64}^{o} is determined by enforcing microscopic reversibility:

$$k_{64}^{o}=\frac{k_{61}^{o}k_{12}^{o}k_{23}^{o}k_{34}^{o}k_{46}^{o}}{k_{43}^{o}k_{32}^{o}k_{21}^{o}k_{16}^{o}}.$$

The voltage-independent rate constants (k_{ij}^{o}) and parameters for the six-state model are provided in Table S3, where the k_{54}^{o} and k_{52}^{o} rates were determined from microscopic reversibility:

$$k_{54}^{o}=\frac{k_{23}^{o}k_{34}^{o}k_{45}^{o}k_{52}^{o}}{k_{43}^{o}k_{32}^{o}k_{25}^{o}},$$

$$k_{52}^{o}=\frac{k_{12}^{o}k_{25}^{o}k_{56}^{o}k_{61}^{o}}{k_{21}^{o}k_{16}^{o}k_{65}^{o}}.$$

For completeness, we have performed other simulations which are not reported here. Because the net charge transferred between $C_6 \rightarrow C_2$ is 1.0 (13), and the stoichiometry for hSGLT1 is 2 Na^+:1 glucose, simulations were performed with 1 net charge transferred from $C_3 \rightarrow C_6$. These were performed with different ϵ_{34} and ϵ_{46} values, but under the constraint $\epsilon_{34}+\epsilon_{46}=1$. Our finding was that the results are qualitatively similar, that is, the voltage dependence of the transitions between C_3 and C_6 does not affect the conclusions that the intracellular release of Na^+ and glucose is uncoordinated.

Simulations with TM0. The full structure of vSGLT including the TM0 helix was constructed using MODELLER, version 9.15 (30), along with the galactose-bound crystal structure (PDB ID code 3DH4) and the well-resolved TM0 helix from the apo structure (PDB ID code 2XQ2). Subsequence models are referred to as M4. The structural alignment of TM5 and TM8 of 2XQ2 onto 3DH4 results in a good superposition of the TM0 backbone atoms allowing us to unambiguously assign the residues in TM0, which were deposited as alanine residues in 3DH4 (Fig. S9 *A* and *B*). The first unassigned alanine in the 3DH4 structure aligns with F10 in the 2XQ2 structure, and the resulting assignment is identical to the one predicted by Mazier et al. (9). One hundred models were generated, and the MODELLER DOPE score, along with visual inspections of side-chain rotamers of TM0, led to the final model selection used for all subsequent simulations. The protein along with the galactose substrate and the crystallographic Na^+ were oriented with respect to the *z* axis using Orientation of Proteins in Membranes (OPM) (43) and then inserted into a 1-palmitolyl-2-oleoyl-*sn*-glycero-3-phosphatidylethanolamine (POPE) membrane using the CHARMM-GUI Membrane Builder (31). The system was then solvated in a rectangular box 94 × 94 × 102 $Å^3$ in size containing 85,000 atoms and neutralized with 150 mM Na^+ and Cl^-.

Next, the system was minimized with NAMD, version 2.9 (37), using conjugate gradient for 10,000 steps. Coordinates were saved at six different times during the minimization procedure: 2,000 (S0), 6,000 (S1), 7,000 (S2), 8,000 (S3), 9,000 (S4), and 10,000 (S5) steps. Simulations were carried out using the CHARMM36 parameter set [CHARMM22 with CMAP correction for the protein (32, 33), CHARMM force field for pyranose monosaccharides for galactose (35), and CHARMM36 for lipids (34)] using a TIP3P water model (36). Each system (S0–S5) was subjected to gradual heating from 10 to 310 K at a rate of 20 K every 15 ps using temperature reassignment. During the heating phase, the dynamics were carried out in a constant volume/temperature (NVT) ensemble with the Na2 site Na^+, galactose heavy atoms, and protein backbone heavy atoms restrained with a 10 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$ harmonic force constant. The side-chain and lipid head group heavy atoms were restrained with a 5 $kcal{\cdot}mol^{-1}{\cdot}Å^{-2}$ force constant. After reaching 310 K, the force constraints were decreased by one-half, and a fast 25-ps simulation in the NVT ensemble was carried out. Next, we switched to an NPT ensemble using the Langevin piston barostat with a 200-fs piston period and 100-fs piston decay constant to maintain the pressure at 1 bar. Temperature was maintained at 310 K using Langevin dynamics with 1 ps^{-1} damping coefficient. Constraints were gently reduced over the next 1.2 ns, and each system was then allowed to equilibrate for 5 ns without restraints. Finally, each system was simulated for 100 ns or until the Na^+ escaped from the Na2 site (Fig. S9*C*). Bond lengths between hydrogen and heavy atoms were constrained with the SHAKE algorithm, and a 1-fs time step was used for heating and the initial NPT phase, followed by a 2-fs time step for the remainder of the simulations. The particle mesh Ewald summation was applied with an interpolation order of 6 along the grid. The van der Waals interactions were smoothly switched to zero between 11 and 12 Å.

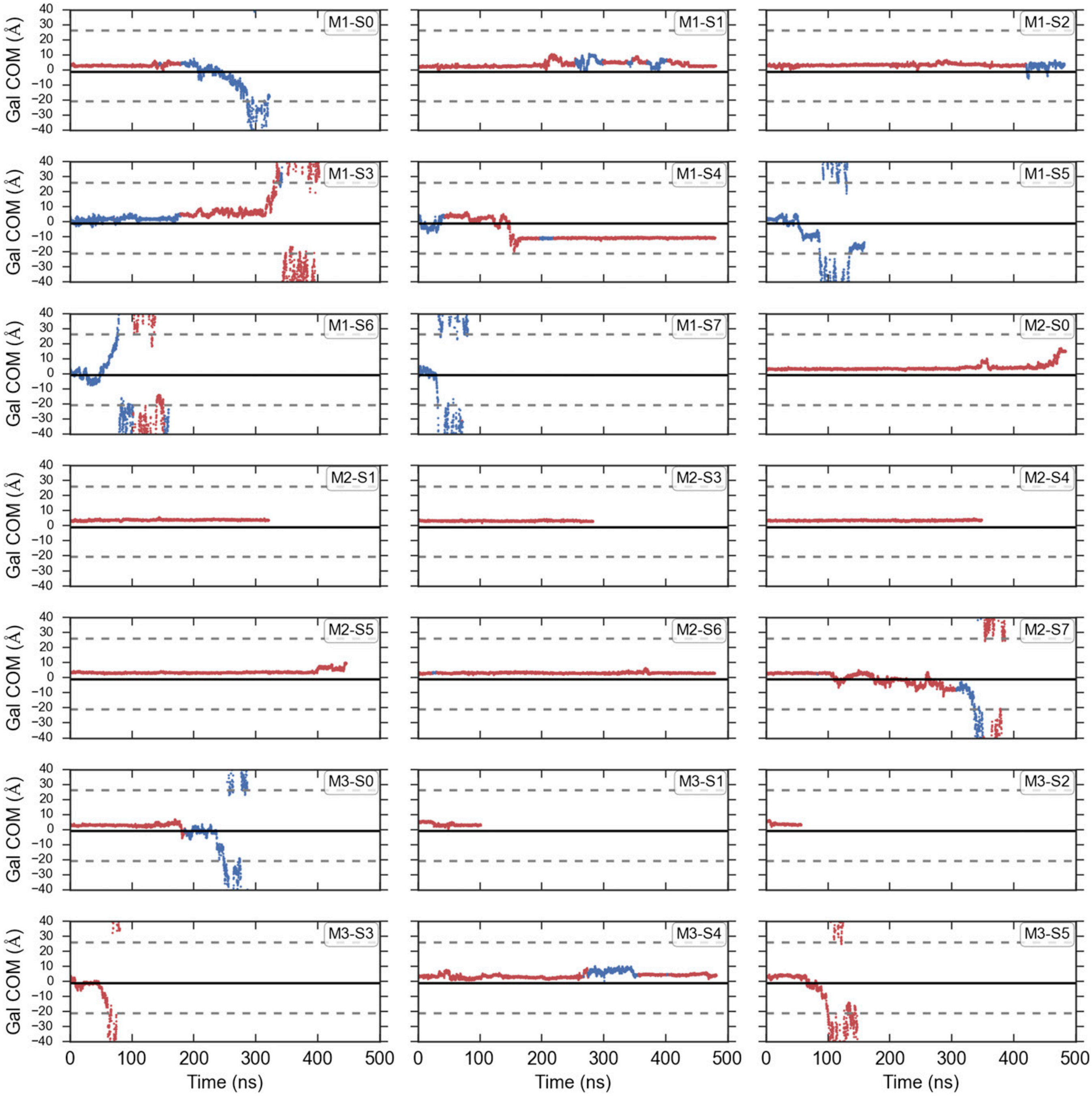

Fig. S1. Galactose escape and Y263 rotamer conformation. For each of the simulations, the center of mass of the galactose molecule is shown projected onto the *z* axis. The points along the trajectory are colored red if Y263 is in the crystallographic rotamer conformation and blue if it is in the alternative rotameric state. The boundaries of the membrane are shown as dashed lines, and the center of mass of the side-chain atoms of Y263 in the crystallographic conformation is shown as a solid black line.

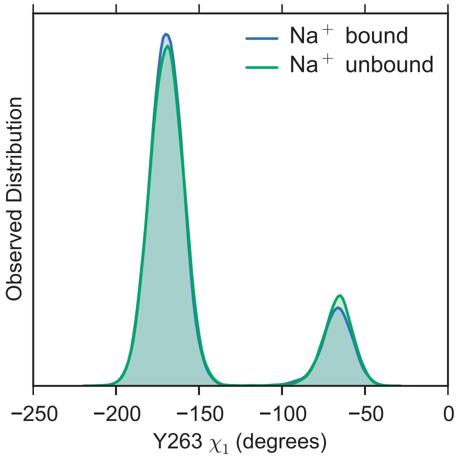

Fig. S2. Rotamer conformations of Y263 with and without Na$^+$. Distribution of the χ_1 dihedral angle of Y263 with Na$^+$ present (blue) or absent (green) from the Na2 site. The Na2 site is categorized as being occupied if any Na$^+$ ion in the simulation box is within 4.5 Å of the either oxygen atom in the carboxyl group of D189. It is difficult to distinguish between the distributions due to the high degree of overlap.

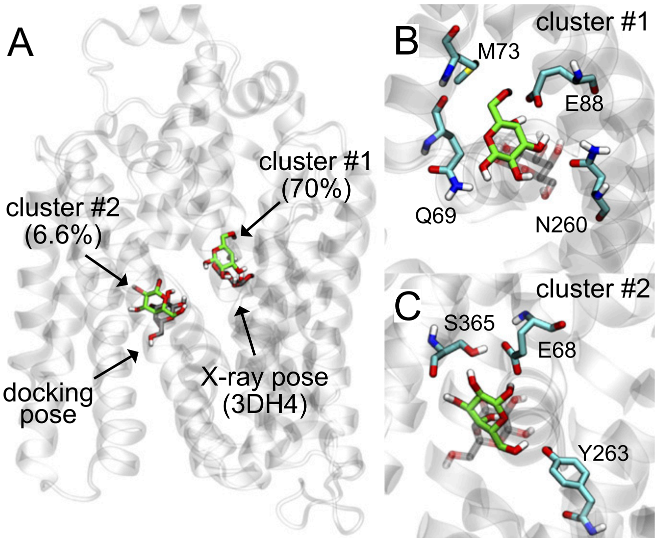

Fig. S3. Clustering analysis of galactose molecules from sugar release simulations. (*A*) The vSGLT structure (PDB ID code 3DH4) with resolved galactose (gray molecule on *Right*) and the generator for the most populated cluster (green), which contains 70% of the data. The Schrödinger Glide XP software package (44) was used to dock (rigid receptor/flexible ligand) a second galactose molecule into the intracellular cavity resulting in the pose on the *Left* (gray). This pose occupies the second most populated cluster (6.6% of the data) with the cluster generator shown in green. Clustering was carried out using MSMBuilder3 in concert with MDTraj (41). The K-centers algorithm was used with a square Euclidean metric and 500 clusters. (*B*) Magnified image of the top cluster generator (green) and resolved galactose molecule from the 3DH4 structure (gray). (*C*) Magnified image of the second most populated cluster generator (green) with docked pose (gray). The docked pose and cluster generator contain three of the residues identified by Li et al. in their recent work (11).

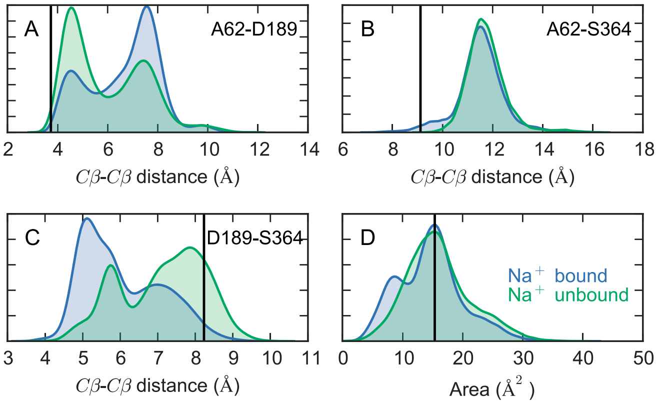

Fig. S4. Residue–residue distance distributions in the Na2 site. Distributions of the Cβ–Cβ distances between (*A*) A62 and D189, (*B*) A62 and S364, and (*C*) D189 and S364, in the presence and absence of Na$^+$ in the Na2 site. (*D*) The area of the triangle formed by the three residues. The black vertical line in each panel is the value observed in the 3DH4 X-ray structure.

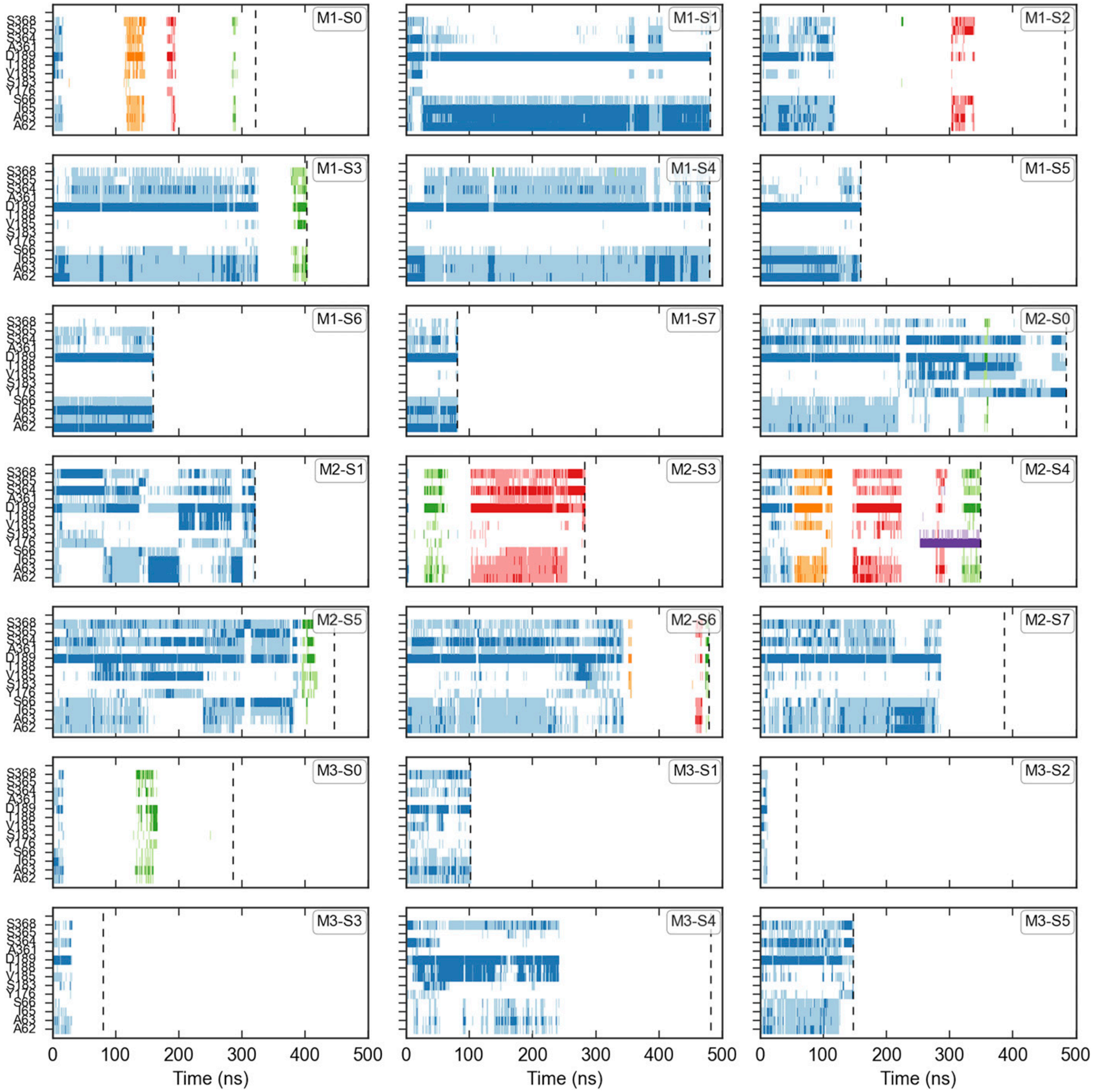

Fig. S5. Na$^+$ interactions with key Na2 binding site residues. The different primary colors represent distinct Na$^+$ ions, where the ion initially placed in the Na2 site is shown in blue. Time points where the sodium is within 3 Å of the residue's backbone carbonyl or terminal side-chain oxygen are shown as dark colors. When the Na$^+$ is between 3 and 4.5 Å from the same set of atoms, the time point is shown as a lighter shade of the same color. The end of the trajectory is denoted by a vertical dashed line.

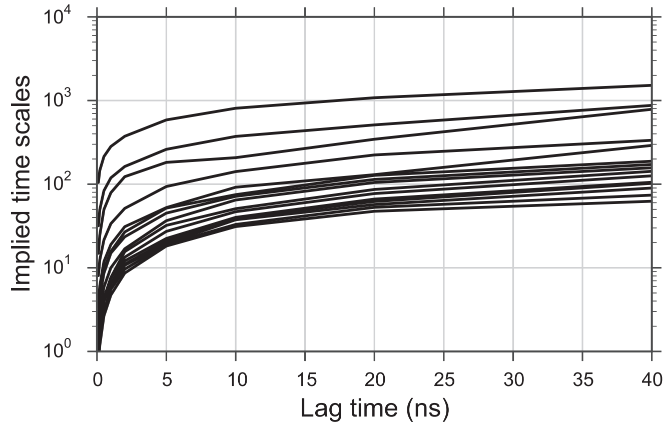

Fig. S6. MSM implied timescales. Implied time scales calculated from the MSM transition probability matrix as a function of lag time.

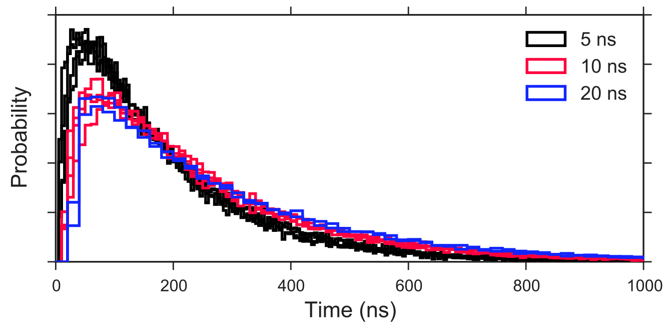

Fig. S7. Na^+ escape first passage time distributions. Na^+ escape time distribution calculated from 30,000 simulated trajectories of the MSM transition matrix for models with varying lag time and hyperparameters. Models are selected based on well-behaved implied timescales and GMRQ scoring criteria described in the text.

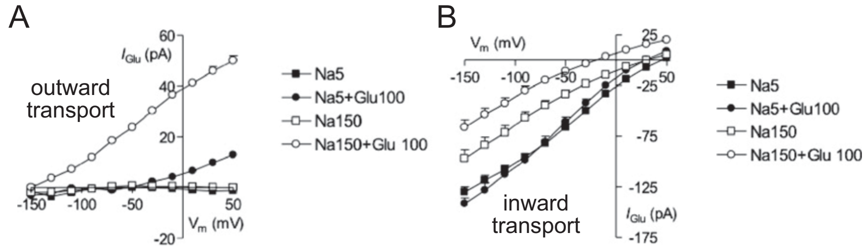

Fig. S8. Voltage dependence of inward and outward currents. (*A*) Outward currents measured in the presence of 10 mM Na^+ in the extracellular solution. Transinhibition of the outward current was tested with four different intracellular solutions: 5 mM Na^+ (black squares), 5 mM Na^+ plus 100 mM glucose (black circles), 150 mM Na^+ (white squares), and 150 mM Na^+ plus 100 mM glucose (white circles). Symbols are means and SEM (n = 9). (*B*) Inward currents measured in the presence 150 mM Na^+ and 100 mM glucose in the extracellular solution. The intracellular solutions and corresponding symbols are the same as those used in *A*. Symbols are means and SEM (n = 9).

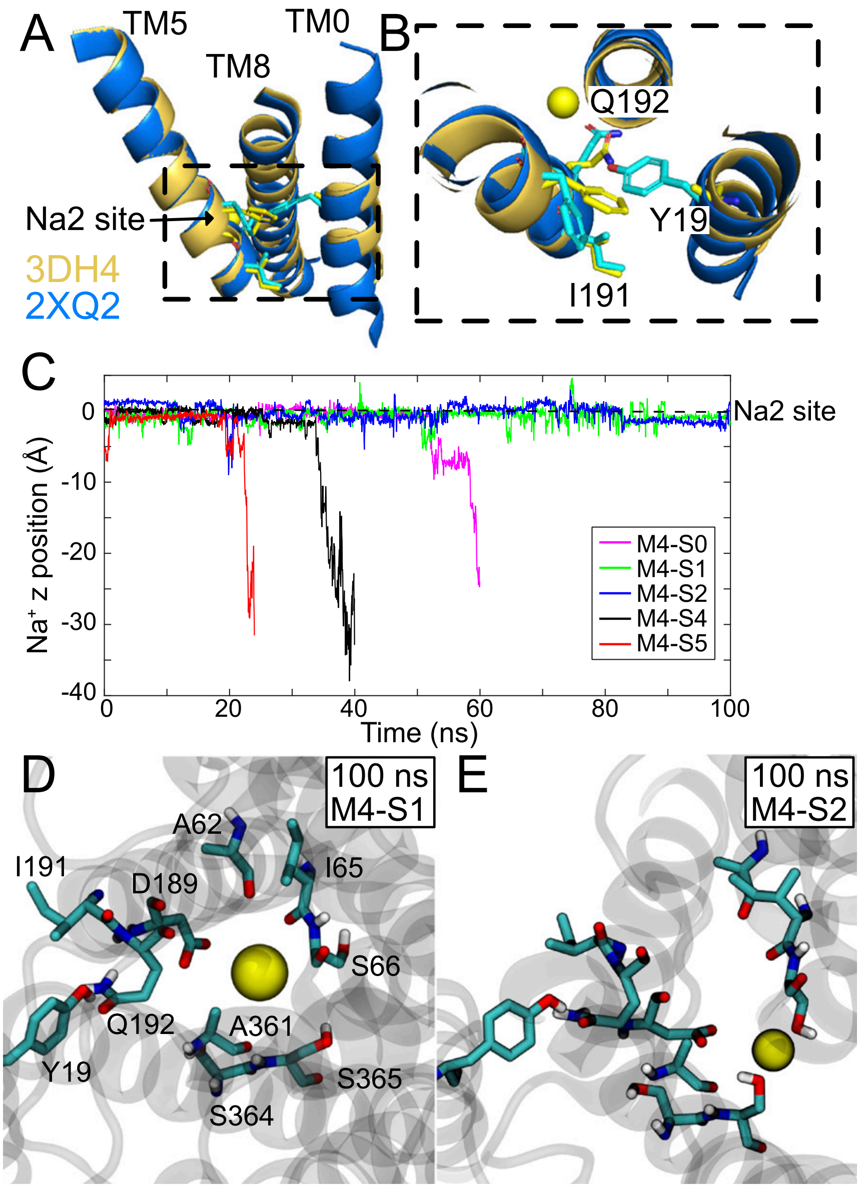

Fig. S9. vSGLT simulations with TM0 helix. (*A*) Superposition of 3DH4 structure (gold) onto the 2XQ2 structure (blue) using only the C_{α} atoms of TM5 and TM8. TM0 backbone from both structures align well. Y19, I191, and Q192 are represented as sticks (cyan from 2XQ2, yellow from 3DH4). (*B*) Magnified view of dashed box in *A*. Y19 in TM0 from 2XQ2 is positioned to interact with Q192 in TM5, as predicted by Mazier et al. (9). Although Q192 adopts a slightly different conformation in 3DH4 and 2XQ2, I191 is identical in both structures. (*C*) Position of bound Na2 Na$^+$ ion in time. Time 0 corresponds to the end of equilibration, and *z* equal zero is the starting position of Na$^+$ in the X-ray site. Release to the cytoplasm was assumed for *z* values less than −10 Å. Results for model M4-S3 is not pictured because the Na$^+$ was lost to the cytoplasm before the end of equilibration. The Na$^+$ in models M4-S1 and M4-S2 remains bound for the entire 100-ns simulation. The coordination of the Na$^+$ in the Na2 site at 100 ns for model M4-S1 (*D*) and model M4-S2 (*E*). Model M4-S1 retains many of the Na$^+$ contacts observed in the 3DH4 X-ray structure, whereas model M2-S2 does not. In all simulations, Y19 remains directly hydrogen bonded to Q192 throughout most of the simulation. Occasionally, a water molecule mediates the interaction, as observed in ref. 9, but we never observe Y19 form a hydrogen bond to the backbone atoms of I191. A list of Na$^+$ exit times is provided in Table S4.

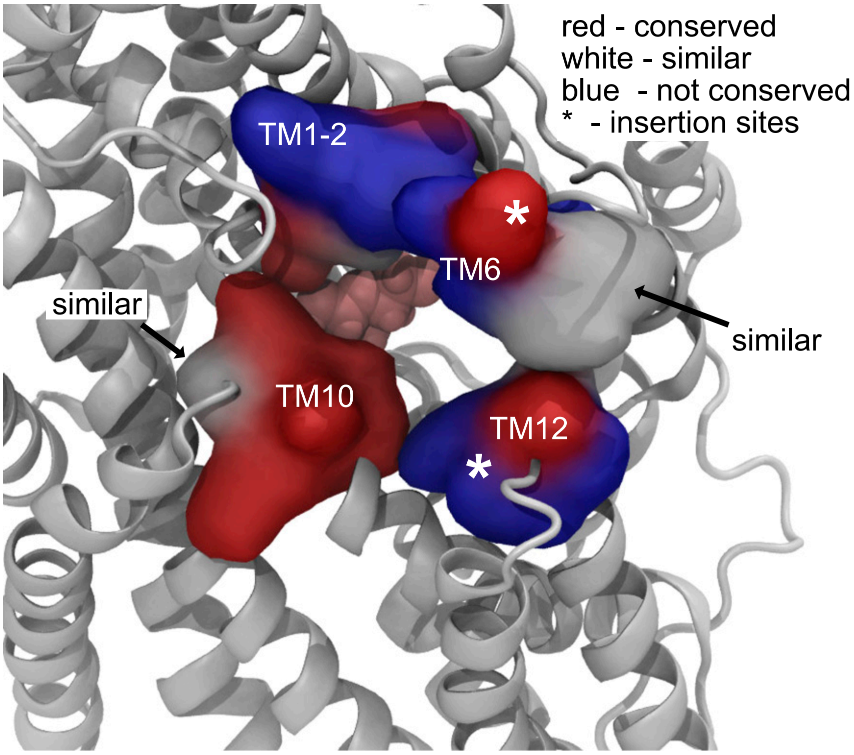

Fig. S10. The outer gate is poorly conserved from vSGLT to hSGLT1. Pictured is a snapshot of vSGLT from a simulation in which the outer gate opens widely. The view is from the extracellular solution, and galactose can be seen in the binding site (pink). The outer gate residues are represented by their molecular surface, and they have been classified into three categories: conserved (red), similar (white), and nonconserved (blue). There are two stretches of amino acids inserted into hSGLT1/2 that are not present in vSGLT (21), and these insertion sites are indicated by asterisks in TM6 and TM12.

Table S1. MSM hyperparameters were tuned via grid search

Hyperparameter	Grid search values
No. of tICA components	1, 2, 3, 4
tICA lag time, ns	1, 2, 5
No. of MSM microstates	25, 50, 100, 200, 400, 800
MSM lag time, ns	0.1, 0.2, 0.5, 1, 2, 5, 10, 20

Table S2. Model parameters for the five-state kinetic model

$k_{12} = 50{,}000\ \mathrm{M}^{-2}\cdot\mathrm{s}^{-1}$	$k_{21} = 300\ \mathrm{s}^{-1}$	$\varepsilon_{12} = 0.3$	$\varepsilon_{21} = 0.3$
$k_{16} = 600\ \mathrm{s}^{-1}$	$k_{61} = 25\ \mathrm{s}^{-1}$	$\varepsilon_{16} = 0.7$	$\varepsilon_{61} = 0.7$
$k_{23} = 45{,}000\ \mathrm{M}^{-1}\cdot\mathrm{s}^{-1}$	$k_{32} = 20\ \mathrm{s}^{-1}$	$\varepsilon_{23} = 0$	$\varepsilon_{32} = 0$
$k_{34} = 50\ \mathrm{s}^{-1}$	$k_{43} = 50\ \mathrm{s}^{-1}$	$\varepsilon_{34} = 0$	$\varepsilon_{43} = 0$
$k_{46} = 10\ \mathrm{s}^{-1}$	$k_{64} = 3{,}750{,}000\ \mathrm{M}^{-3}\cdot\mathrm{s}^{-1}$	$\varepsilon_{46} = 0$	$\varepsilon_{64} = 0$

Table S3. Model parameters for the six-state kinetic model

$k_{12} = 50{,}000\ \mathrm{M}^{-2}\cdot\mathrm{s}^{-1}$	$k_{21} = 300\ \mathrm{s}^{-1}$	$\varepsilon_{12} = 0.3$	$\varepsilon_{21} = 0.3$
$k_{16} = 600\ \mathrm{s}^{-1}$	$k_{61} = 25\ \mathrm{s}^{-1}$	$\varepsilon_{16} = 0.7$	$\varepsilon_{61} = 0.7$
$k_{23} = 45{,}000\ \mathrm{M}^{-1}\cdot\mathrm{s}^{-1}$	$k_{32} = 20\ \mathrm{s}^{-1}$	$\varepsilon_{23} = 0$	$\varepsilon_{32} = 0$
$k_{34} = 50\ \mathrm{s}^{-1}$	$k_{43} = 50\ \mathrm{s}^{-1}$	$\varepsilon_{34} = 0$	$\varepsilon_{43} = 0$
$k_{45} = 800\ \mathrm{s}^{-1}$	$k_{25} = 0.01\ \mathrm{s}^{-1}$	$\varepsilon_{45} = 0$	$\varepsilon_{54} = 0$
$k_{65} = 4{,}500\ \mathrm{M}^{-2}\cdot\mathrm{s}^{-1}$	$k_{56} = 16\ \mathrm{s}^{-1}$	$\varepsilon_{56} = 0$	$\varepsilon_{65} = 0$

Table S4. Sodium escape times with TM0

Simulation	Escape time
M4-S0	60 ns
M4-S1	Entire 100 ns
M4-S2	Entire 100 ns
M4-S3	During equilibration
M4-S4	35 ns
M4-S5	21 ns