Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning

Darko Mitrovic[1], Sarah E. McComas[1,2], Claudia Alleva[1,2], Marta Bonaccorsi[1,2], David Drew[2,#], Lucie Delemotte[1,#]

[1]Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden. [2]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden.

Abstract

Sugar porters represent the largest group of secondary-active transporters. Some members, such as the glucose (GLUT) transporters, are well-known for their role in maintaining blood glucose homeostasis in mammals, with their expression upregulated in many types of cancers. Because only a few sugar porter structures have been determined, mechanistic models have been constructed by piecing together structural states of distantly-related proteins. Current GLUT transport models are predominantly descriptive, and oversimplified. Here, we have combined coevolution analysis and comparative modeling, to predict structures of the entire sugar porter superfamily in each state of the transport cycle. We have analysed the state-specific contacts inferred from coevolving residue pairs, and shown how this information can be used to rapidly generate free-energy landscapes consistent with experimental estimates, as illustrated here for the mammalian fructose transporter GLUT5. By comparing many different sugar porter models and scrutinizing their sequence, we have been able to define the molecular determinants of the transport cycle, which are conserved throughout the sugar porter superfamily. We have also been able to highlight differences leading to the emergence of proton-coupling, validating and extending the previously proposed latch mechanism. Our computational approach is transferable to any transporter, and to other protein families in general.

Introduction

Due to the importance of glucose and other monosaccharides for cell metabolism[1], sugar porters are the largest and widest-spread family of small molecule transporters across all Kingdoms of life[2,3]. In mammals, the sugar porters are referred to as glucose (GLUT) transporters belonging to the Solute Carrier Family 2A[4-6]. Human has 14 different GLUT isoforms, and each isoform has a distinct pattern of tissue distribution, gene regulation, substrate preference and kinetic properties[6-8]. GLUT1, for example, is distributed in a wide range of tissues, including the blood–brain barrier, and is essential for glucose transport into the brain[1,6], whereas GLUT4 is mostly localized in skeletal muscle and adipose tissue, and is the major insulin-stimulated glucose transporter[8]. GLUT5 is the only member specific to fructose and is the major route for its intestinal absoprtion[9,10]. In plants, fungi, bacteria and parasites the sugar porter family has likewise expanded into a large number of different isoforms, providing essential and niche roles in the uptake of D-glucose and other monosaccharides. Plants, for example, express specific sugar porter isoforms for seed, fruit and pollen production[11] and yeast have 20 different hexose transporters with various kinetics[12], which are targeted for biofuel production[13]. Parasites, such as *Plasmodium Falciparum* rely on sugar transport to infect their host and reproduce[14,15], resorting to a promiscuous sugar porter, which can take up a variety of sugars[16]. Understanding the molecular basis for sugar transport is thus a fundamental question in biology, with important medical and biotechnological applications.

Sugar porters belong to the Major Facilitator Superfamily (MFS)[2,3] and are defined by an N-and C-terminal bundle of 6 TMs, which are connected by a cytosolic loop that is made up of three to four intracellular helices (ICH)[2,17]. The sugar porters can be subclassified by a sequence motif[18], which structures have shown corresponds to an intracellular salt-bridge network that selectively stabilizes the outward-facing state[17,19-21]. Constructing a transport cycle requires assembling five different conformational states along the transport cycle: outward-open, outward-occluded, occluded, inward-occluded and inward-open (Fig. 1A,B)[2,17]. Globally, sugar porters operate according to the rocker-switch alternating-access mechanism[2,17]. Although the two bundle are structurally similar, D-glucose is not coordinated evenly, but almost entirely by residues located in the C-terminal bundle[16,22,23]. As a consequence, half helices TM7b and TM10b in the C-terminal bundle are thought to undergo local rearrangements to control access to the sugar binding site from the outside and inside, respectively[2,17,20,24].

Whilst the sugar porter family is structurally the best characterized out of all MFS transporters[2], there is still not a single sugar porter that has structures determined in all of these different states. Indeed, some conformations, such as the occluded and inward-occluded states[16,25,26], have only been experimentally resolved in one isoform. Due to the paucity of structural information, transport models are currently descriptive, lacking rigorous validation. Consequently, it is still unclear if the single snapshots represent physiologically relevant conformations along the transport cycle, and if unique structural features are transferable to more distantly-related members.

Despite low sequence identity, structures of distantly-related sugar porters show a high degree of structural conservation as compared to many other types of MFS transporters[2,16]. Using the 20 available experimental structures spread across all major conformations[2], we hypothesized that we should be able to pinpoint co-evolving residues pairs specific to each state, and then use these contacts for structure prediction. Notably, we reasoned that this approach would be superior to either i) homology modeling, which is intrinsically biased towards the captured structure, regardless of whether the features of the template are actually transferable across isoforms or not, or ii) structures generated by coevolution only, which may settle on a conformational state that satisfies the contacts from each of the multiple states, yet fails to capture a physiological conformation.

Here, we present a novel approach for state-specific structure prediction. In essence, we have used experimental structures determined in each of the different conformations to train a neural network to identify state-specific contacts. After filtering these contacts using coevolution information, we have then used them to apply biases to Alphafold2 models[27] and driven them towards the various conformational states along the functional cycle. Next, we have combined these contacts into collective variables to use in enhanced sampling molecular dynamics (MD) simulations [28,29], and computed the free-energy landscape of the fructose transporter GLUT5[21]. Finally, with a set of conserved state-specific contacts defined, we were able to pinpoint both contacts that govern the conformational cycle across the sugar porter family, and a series of transporter-specific interactions that control conformational cycling of each of the sugar porter subfamilies. In particular, we concentrate on deciphering the molecular and evolutionary determinants of sugar-proton-coupled symport.

Results

Inferring co-evolving conserved residue pairs across all sugar porters

AlphaFold2 structural models of the various sugar porters are thought to collectively capture all major conformational states[30]. However, deterministic prediction of a given state is not possible. Therefore, as expected, principal component analysis (PCA) based on pair-wise interactions reveals that sugar porters cluster within their individual subfamilies rather than according to the conformational state that happens to be predicted by the structure-prediction method (Fig. 1C, D). To be able to rationally build models of individual SPs in each of the states in the transport cycle, we thus need to steer AlphaFold2 generated models, or experimentally captured structures, when applicable, into alternative states. We achieve this by using state-specific contacts derived from coevolution analysis.

To do so, we first generated a representative sequence alignment by aligning ~1000 sequences from each of the evolutionarily-distant sugar porter relatives belonging to mammalian sugar transporters (GLUTs), the bacterial and parasitic transporters (*Pf*HT1, XylE), the plant sugar transporters (STPs), and the fungal hexose transporters (HXT), separately (Figure 1A). These sugar porter family members were selected because they are either part of a major evolutionary branch of the sugar porter family or are functionally distinct and structurally characterized (as is the case for *Pf*HT1 or XylE). We then used the resulting multiple sequence alignments (MSA) as input for Direct Coupling Analysis (DCA) to generate coevolution maps[31]. The coevolution maps generated from separately aligned MSAs were subsequently combined into a global coevolution map, filtering out contacts that were predominantly species-specific (see Methods). As illustrated for the proton-coupled xylose transporter XylE (Figure 2A), there is an extensive overlap between the MSA-derived coevolution contact maps and experimental inward-facing structural contacts. Nevertheless, several coevolving pairs do not correspond to interactions found in the structure. Overall, we estimate that ~10% of the top 500 co-evolving pairs represent contacts not formed in available crystal structures, presumably representing contacts forming in other conformational states (see Methods).

Extracting state-dependent coevolving residue pairs from experimental structures

To determine which coevolving pairs, correspond to contacts that are formed or broken in the 20 structures of different functional states, we trained a convolutional neural network (CNN) to classify the main five conformational states using as input contact maps from experimentally determined structures filtered by coevolution scores (Figure 2B). The CNN architecture was designed to avoid redundancy of contacts between neighboring residue while allowing residue pairs to be in contact in several functional states (Methods, Figure 2B). Layer-wise relevance backpropagation (LRP) was then performed[32] on all five output classes separately, leading to the assignment of encouraged and discouraged contacts for each conformation (Figure 2C).

Interestingly, encouraged and discouraged contacts were not confined to inter-bundle contacts and are instead spread throughout the entire structure (Table S1). They also displayed strong state-dependency (Figure 3). Because of the low number of experimental structures available, training of the neural network carried a substantial risk of overfitting. We thus sought to validate the results by making sure that the state-specific contacts identified were indeed in the expected regions. Starting from the outward-facing conformation, we observe the expected encouraged contacts of inter-bundle salt-bridges formed between TM3-TM10b and TM5-TM8 helices. In contrast, we observe a strong signal for the discouraged contacts between TM1 and the extracellular gate TM7b, which are known to come together during rocker-switch transition into the inward-facing states[2,21]. In the occluded structure of *Pf*HT1[16], the extracellular gate TM7b was had moved fully inwards and transition into a broken helix at the point closest to TM1. Furthermore, mutations in TM1 were found to be just as critical for transport as those in TM7b[16], indicating that TM1 and TM7b interactions might be important in driving formation of the occluded state. Indeed, we observe a robust signal for encouraged contacts between TM1 and TM7b, which is present as early as the occluded state, confirming that TM7b and the interactions with TM1 are co-evolving to enable attainment of the occluded conformation. As expected for an intermediate state, the occluded state has the maximal number of encouraged contacts. One additional set of encouraged contacts in the occluded state are located between the intracellular gating helix TM10b and TM4, which are required to drive formation of the occluded state from the inward-facing conformation[2,25].

Interestingly, we observe the formation of less well-known inter-bundle contacts between TM1 and TM5, and between TM3 and TM6 when transitioning from an outward-open to an outward-occluded state, possibly modulating the conformational change of TM1. In transition between outward-facing and inward-facing states, additional contacts are also seen between TM3 and TM11, both at the extracellular region, in which case they are encouraged in the inward-facing states, and at the intracellular region, in which case they are encouraged in the outward-facing ones. These previously unreported contacts appear particularly important for transporter function as they are at the basis of the rocker switch motion passing through the occluded state. Taken together, the state-specific coevolution analysis is able to replicate the main structural transitions expected from comparing individual static structures, as well uncovering new contacts to be evaluated (Table S2).

Generating sugar porter structures for each major conformation

To expand the available structural information for sugar porters, AlphaFold2 models were driven into all five conformational states in RosettaMP[33] using the derived encouraged and discouraged state-dependent contacts as attractive and repulsive biases, respectively (Figure 2C, Methods). A straightforward assessment of the quality of the models obtained was difficult to perform, however, as few experimental structures are available. Nevertheless, for an overall approximation, we calculated the Cα root mean squared deviation (RMSD) to the phylogenetically closest available experimental structure. As shown in Figure 4, the mean RMSD difference has an acceptable value of ~2.5 Å, with the RMSD distribution spanning a 95% confidence interval of 1.5 to 3.2 Å. In addition, to evaluate the stability of these models, 10 ns molecular dynamics simulations of these models embedded in a model membrane bilayer were then performed, which further relaxed the structures, bringing their overall RMSD to the closest homolog to ~2.1 Å with the RMSD distribution spanning a 95% confidence interval of 1.1 to 2.7 Å. The RosettaMP modeling pipeline thus seems to bring the models close to their local free energy minimum, but relaxation by MD simulations further improves the quality of the models by relaxing them even more. To evaluate the cause of the RMSD reduction, we compared the GLUT5 outward-open state model before and after MD relaxation to the crystal structure of the same transporter solved in the same state (PDB ID 4YBQ), which showed that the model quality was improved

from a 2.15 Å Cα RMSD to a 1.49 Å Cα RMSD, mainly via the relaxation of TM1 (Figure 4C). It seems that, in this case, flexible helices require an explicit solvent model to attain a more physiological conformation.

As a next assessment of quality, we re-analyzed the STP10 model against the inward-open STP10 crystal structure (PDB ID 7AAR)[34], whose structure was not present in our initial training set as its structure was released after training of our model. The inward-open state model of STP10 previously had a calculated Cα RMSD of 4.3 Å away from its closest homologue with known structure, XylE (PDB 4QIQ)[25]. However, the actual Cα RMSD to the experimental STP10 inward-facing structure difference is only 2.2 Å, which was further reduced to 1.8 Å after MD simulations. Inspection of the STP10 structure reveals a unique helical kink in TM10, which is not present in XylE (Figure 4B). An additional structure for inward-open GLUT4[35] was further determined after our initial model training was finalized. Our inward-open GLIUT4 model has a RMSD of 2.3 Å and 1.8 Å before and after MD relaxation, respectively. Reassuringly, our method is thus able to model new structural features absent from the training set. Indeed, the Cα RMSD estimated for HXT and STP members, as well as for models of the inward occluded states, which have fewer structures available to properly assess their accuracy, is generally higher (Figure 4A). Overall, we conclude that the generated sugar porter models are of sufficient quality for linking conformational states of sugar porters.

A free-energy landscape for the fructose transporter GLUT5

Characterizing robust free energy landscapes of conformational cycles of sugar transporters will enable us to better understand the mechanistic basis for sugar transport. Although free energy landscapes for GLUT transporters have been reported previously[36-39], these were generated from a few structures only, and lacked the most recent structure in the occluded conformation[2,16]. Indeed, as co-evolution analysis confirms, the occluded state is an intermediate that has a number of important and specific co-evolved pairs that are critical for linking the outward and inward-facing conformations. In addition, the methods used to generate these landscapes did not consider whether structural features captured in specific structures were transferable to other family members. Here, we decided to generate a free energy landscape for the

fructose transporter GLUT5 using family-wide information in the form of coevolving contacts. We chose GLUT5 because it is the only transporter with experimental structures in both fully outward- and inward-facing conformations[21] and because our recently determined free energy landscape using a more traditional enhanced sampling method can serve as a comparison[40].

The most efficient enhanced sampling techniques require the choice of a low-dimensional collective variable (CVs) set that encompass all degrees of freedom implicated in a conformational transition[28,41]. Given that state-dependent contacts constitute the driving force for a conformational transition, CVs based on those are well suited to distinguish conformational states and to enhance transitions between them. We thus trained a support vector machine (SVM) to distinguish the modelled structures of adjacent states (Figure S2) and extracted therefrom top coevolving contacts, as inferred from their high SVM coefficients. We then designed state-specific CVs as weighted sums of distances between top state-specific coevolving pairs, using as weights the SVM coefficients (See Methods). Finally, we ran accelerated weighted histogram (AWH), an efficient enhanced sampling method natively available in GROMACS, using these CVs as input. Accumulating 250-650 ns of MD simulations enabled an extensive sampling of the conformational space and the estimation of the corresponding free energy landscapes (Figure 5A-C) with satisfactory accuracy (Figure S1). Extracting the free energy along the minimum free energy path linking adjacent states made it possible to estimate a continuous 1d landscape that follows the minimum free energy path linking the outward-facing and the inward-facing states via the three occluded intermediates (Figure 5D, see Methods).

In the absence of substrate, the most favorable conformation for GLUT5 is the outward-facing state, which is consistent with biochemical analysis and the salt-bridge network on the intracellular side stabilizing this state[20,21,42]. The outward-open and outward-occluded states appear to be of comparable free energies, and the barriers between the outward-open and outward-occluded states are fairly low at 4 to 8 kJ/mol, respectively (Figure 5D). This observation is consistent with the fact that the structure of GLUT3 has been determined in both outward-open and outward-occluded states, even in the presence of a bound maltose[23]. The largest energetic barrier of ~34 kJ/mol is located between the outward occluded and the inward-occluded state (Figure 5D). This barrier, presumably arises in part from the breakage of the strictly-conserved salt-bridge

network, and is consistent with the activation barrier of 10 kcal/mol as estimated by GLUT1 kinetics (43 kJ/mol) [43,44](Figure 6E). Notably, the relaxed GLUT5 occluded state does not fall on the largest energetic barrier corresponding to the transition state (Figure 5). However, in the presence of a substrate sugar, GLUT5 passes through a transition state that closely matches the occluded state of *Pf*HT1[40]. Our analysis indicates that coevolution-driven MD simulations are detecting an energetic minimum for an occluded state prior to the transition state which has not been experimentally observed (Figure S3). It is possible that the co-evolution analysis is detecting a pre-transition state, which is likely to be lowly populated in the presence of a substrate sugar. Taken together, we conclude that using evolutionary-based collective variables enables us to obtain sugar porter conformational free energy landscapes using AWH using simulations of the order of hundreds of nanoseconds. This represents an increase in performance by at least an order of magnitude compared to all of previous attempts[38,45,46].

Family-wide state-dependent interactions

Having validated that the top co-evolving pairs can be combined into collective variables able to connect the different conformational states and describe the energetics of transitions between them appropriately, we scrutinize the interactions that went into the construction of these CVs by projecting them onto the models of GLUT5 states (Figure 6). Notably, these interactions are more robustly defined than in the previous analysis shown in Figure 3, since those were based on the analysis of a reduced set of experimentally resolved structures. In addition, we track the functional diversity across the SP family by extracting the types of residue interactions found at these sites (Figure 6).

Starting from the outward-open state, as expected, salt-bridge forming residues appear to stabilize the outward-facing conformations, a network that is maintained through to the occluded state (Figure 6A). In addition, the residue Ser306 located at the back of TM7b has co-evolved to form an interaction to Gln366 interaction in TM9, an interaction seen as Thr/Arg or Lys in the STP members and as between small and/or polar residues in the rest of the superfamily (Figure 6B).

The outward-occluded state is characterized by a bend in the TM7b helix, which occludes sugar exit[2,47]. The contact analysis shows that this state appears to be stabilized by co-evolving pairs between Tyr297 in TM7b and Asn325 and His387 in TM10a, which are peripheral to the sugar binding-site and could therefore be connected to sugar binding (Figure 6C). This interaction remains formed through to the inward occluded open state. In most other GLUT members and in the HXT proteins, Tyr297 and His387 are replaced by Phe residues that are expected to interact with TM10a residues via hydrophobic or pi-pi interactions.

In the occluded state, TM7b transitions from a bent to broken helix, which occludes the sugar-binding site and moves TM7b closer to TM1. Interestingly, Asn40 of TM1 forms a hydrogen bond with the backbone residues at the break-point in TM7b, an interaction that remains formed through to the inward open states (Figure 6D). The Asn40 is conserved in *Pf*HT1 and mutation to alanine was shown to severely impair transport[16]. In most other family members, a hydrogen bond is made possible by the presence of a small polar residue in TM1, highlighting that stabilized TM7b closure is generally connected to interactions in TM1 as early as in the occluded state.

In addition to the most outward-facing salt bridge network (Figure 6A) breaking, rearrangements of TM10a helix and formation of the Tyr374 (TM10)-Thr318 (TM8) hydrogen bond characterizes entry into the inward facing states (Figure 6F). These positions often feature residue pairs capable of interaction either via hydrophobic interactions or H-bonding (Leu/Val-Ala in STPs, Asn/Tyr-Thr/Ser in the rest of the family). The formation of this contact in the inward occluded state is critical determinant for the rocker switch motion, as the resulting rearrangement of TM10a promotes the tilting of the peripheral helices, which ultimately rocks the helical bundles into an inward facing state.

For the intracellular gate to open, the residual Glu401-Arg159 salt bridge (Figure 6E) must completely break. Inspection of the distribution of the minimum distance between Glu401-Arg159 contact across reveals that it increases from an average of 1.4Å in the occluded state, to 3.4Å in the inward occluded state. The weakening of Glu401-Arg159 presumably contributes to releasing TM10b. Entry into the inward-open state, stabilizing the tilted TM10b, features the full breakage of the Glu401-Arg159 interaction, and Glu401 forming a salt bridge with a different partner, namely Arg341

(Figure 6G). The rotation of TM10b shifts the occluding Leu397 residue away from intracellular pathway to the sugar binding site, opening the intracellular gate fully. This state is also stabilized by additional strictly conserved hydrophobic contacts in the TM10-TM7 interface (positions 397-400 in GLUT5, which rearrange from an interaction with TM4 in the outward-facing and occluded state) (Figure 6G and S5).

Coevolving residues support proton-coupling in sugar porters

Having characterized the aspects of the sugar porter transport cycle that are likely conserved across the family, we turn to family-specific features, with a focus on the differences between passive and proton-coupled transporters, such as XylE. The current working model is that an aspartic acid reside in TM1 (Asp27 in XylE) is allosterically coupled to the sugar transport, but does not participate directly in sugar binding[2,36,48,49]. Indeed, the aspartic acid to asparagine mutant has the same sugar-binding affinity as wildtype[50]. Rather, it is thought that the protonation of the TM1 aspartic acid is required for transport as its breaks the outward-facing-specific salt-bridge interaction to an arginine residue Arg133 on TM4. Based on the occluded *Pf*HT1 structure[2], highlighting the coupling between TM1 and TM7b, a latch mechanism for proton-coupling was proposed[2]. Simplistically, breakage of the Asp27-Arg133 salt-bridge enables TM1 to come closer to TM7b and TM7b only comes close enough to TM1 when sugar is present. Formation of the occluded state is catalyzed by sugar binding, but in proton coupled transporters, this is only enabled by the release of the latch (Figure 7A and S4).

The co-evolution analysis is consistent with TM1 and TM7 interactions driving formation of the occluded state. More specifically, we indeed see that the equivalent TM1 Asp and TM4 Arg salt-bridge is broken in occluded states of all known proton-coupled symporters, such as STP10. The details of the conformational rearrangement differ, however. After breaking of the latch, in the occluded state of XylE, Arg133 forms a salt bridge with the neighboring TM6 Glu206 residue (Figure 7A). In STP10, on the other hand, Glu206 is replaced by an alanine, and R133 rearranges by adopting a different rotameric state facing towards TM11, where it appears to be stabilized by an interaction with the phospholipid headgroups (Figure 7B and S4).

Most GLUTs are thought to be passive transporters and do not possess an acidic residue at position 27, consistent with its purported role as proton carrier (Figure 7D).

Intriguingly, however HXTs, which are generally not proton-coupled, do carry an acidic residue in this position. Our analysis reveals that this is explained by the fact that these transporters contain a latch interaction that is maintained throughout the conformational cycle thanks to a large-scale pivot of the first helical bundle (Figure 7C and S4). This conformational change unique to HXTs is enabled by helix breaking residues at position 202 (Pro/Gly), which are replaced by mostly hydrophobic residues in the rest of the family (Figure 7D). GLUT2, a passive transporter that nevertheless possesses an acidic residue at position 27, has a Serine in position 205 which interacts with D27, which appears to help the salt bridge remain formed in the occluded state (Figure S4D).

Discussion

Alphafold2 has made it possible to produce high-quality models of structures associated with any amino acid sequence[27]. Nevertheless, these structures represent an arbitrary conformational state, and do not capture the conformational heterogeneity that enables protein function. Modifying the MSA depth enables to make models of alternate states, though in a non-deterministic manner[30]. In this work, we have used i. AlphaFold models, ii. coevolution analysis encoded by family-wide multiple sequence alignments and iii. state-specific contacts extracted from the ~20 structures to guide enhanced MD simulations. Our work goes beyond the analysis of static structures and enables the separation of state-specific and transporter-specific features, thus firmly establishing the family-wide determinants of the conformational cycle as well as establishing specificities of GLUT5 and of proton-coupled sugar transporters. By introducing coevolution as a measure of predicted interaction, we are able to develop models that go beyond the comparison of structures only. We also minimize information that may be lost during dimensionality reduction, as demonstrated by our validation on newly reported structures.

Using state-dependent coevolution-based contacts, we show that we can create a low-dimensional projection that describes the transition between adjacent states. When derived into a set of collective variables used in enhanced sampling molecular dynamics simulations, we can construct weighted conformational ensembles, or in other words free energy landscapes of the process. Comparing our convergence times (to less than

0.01 kJ/mol within 250-650ns per walker and system) to previous works on similar systems[38,45,46] and given the agreement between the activation barrier estimate from kinetic measurements[43,44] we conclude that the introduction of coevolution data is a powerful approach in obtaining accurate free-energy landscapes efficiently.

The conformational ensembles reveal an overall functional cycle that is largely consistent with the evaluation of experimentally resolved structures, yet also provides details that were previously unknown. Across the larger MFS family of transporters, cavity-closing contacts are predominantly formed between TM1-2 and TM7-8 on the outside and between TM4-5 and TM10-11 on the inside[2,17]. However, the sugar porters, as exemplified by GLUT transporters, coordinate the sugars very asymmetrically with only a single residue in the N-terminal bundle coordinating the sugar. As such, local rearrangements during the transport cycle are assumed to be primarily established by local changes in TM7b and TM10b half-helices, which coordinate the sugar in the C-terminal bundle. Although sugar porters are made up two structurally similar bundles, the asymmetrical rearrangements are more akin to the conformational changes described by rocking-bundle proteins[2,17], which are made up from structurally distinct bundles. Consistently, co-evolution analysis is able to detect the importance of the substrate gating regions for driving conformational changes. In particular, the co-evolution analysis shows that TM7b and its interaction with TM1 have evolved to come together already in the occluded state, *i.e.,*, rather than only interacting in the inward-facing conformatios. This conclusion is in agreement with the occluded structure seen for *Pf*HT1[16], wherein TM7b had moved completely inwards to break over the sugar-binding site.

The requirement for TM1 to interact with TM7b in formation of the occluded state further explains how sugar porters can evolve to be proton-coupled, even if the proton-accepting residue itself does not directly coordinate the sugar, i.e., unlike in other sugar transporters, such as LacY. Simplistically, an aspartic acid residue in TM1 forms a salt-bridge with an arginine residue in TM4, restricting its movement. Protonation of the aspartic acid residue removes this constraint and TM1 can come closer to TM7b, which itself is stabilized inwards when it binds a substrate sugar. Interestingly, an acidic residue in TM6 (Asp206, XylE numbering) is further required to provide a favorable alternative conformation for the unpaired arginine. In the passive GLUT transporter GLUT2 and the yeast hexose transporters, for example, this additional acidic residue is

missing and the salt-bridge is still retained due to the assistance of a nearby serine residue (Ser205, XylE numbering). Nevertheless, STP members are proton-coupled and yet they do not possess an additional acidic residue in TM6. Rather, the arginine residue is able to snorkel to the membrane interface in the unpaired state. Notably, although TM7b and TM1 are important, they are part of a larger interaction network that also includes, for example, how TM7b interacts with TM8 in the outward-open state.

Consistent with the orchestrating role of TM7b in substrate translocation, the co-evolved interaction with TM1 is retained across the entire second half of the transport cycle. In contrast, the inward-facing gating helix TM10b is likely to have a more passive role and a co-evolved interaction with TM4 is only formed in the occluded state. Nevertheless, TM10b dynamics are likely to be important and its possible that TM10b movement is facilitated by the stabilization of TM10a with TM8 on the extracellular side (Figure 6F). The bottom of TM10a harbors a strictly-conserved acidic residue, which is part of the intracellular salt-bridge network. The contact maps initially generated for all sugar porters indicated mostly discouraged contacts between TM4 and TM10b residues in the inward-occluded state (bundle 5 contact in Figure 3). In addition, the GLUT5 models show that whilst most contacts are indeed broken, there is still an encouraged contact between TM4 and TM10b residues in the inward-occluded state (Figure 6F, S5B). It thus appears that, at least in GLUT5, only one of the salt-bridge pairs are fully broken between the occluded to inward-occluded states. Such a result is intriguing, and implies that even the rocker-switch bundle movement might utilize asymmetric rearrangements.

Lastly, GLUT transporters are shown as textbook examples of how small molecule transporters are functional equivalents to enzymes[2]. Key to understanding catalysis is to understand how the transition state is formed during substrate translocation. The transition state, however, is only transiently occupied and therefore difficult to experimentally capture. Nevertheless, the parasite transporter *Pf*HT1 has an unusually very polar TM7b gate, which has made it possible to capture an occluded state with sugar present[16,26] i.e., notably this state is unlikely to represent a transition state for *Pf*HT1. Nonetheless, out of the models we built, *Pf*HT1 has highest number of state-specific contacts in both gates. We speculate that the *Pf*HT1 could represent a pre-

transition state (Figure 5), with further dynamics between the bundles required to access the transition state. In summary, although many questions remain unanswered, our novel approach provides a rational framework for understanding how sugar porters function at the molecular level, and provides the information to re-engineer sugar porters with different characteristics, which might be otherwise inaccessible by traditional forward-evolution approaches. At a conceptual level, our work highlights how even the most simplistic type of transporters have evolved fine-tuned and intricate interactions to achieve substrate translocation.

Methods

Coevolution analysis

To construct the sequence alignment at the basis of coevolution analysis, a representative sequence from each member of the sugar transporter family was taken as a seed for a sequence search against the Uniref90 database[51] using PSIBLAST[52]. In this way, a sufficient number of diverse sequences was consistently found around each member (M_{eff}>1000, where M_{eff} denotes the number of sequences with less than 80% identity). The sequence libraries were then aligned within each subfamily using a stair-shaped guide tree[53] in the MUSCLE alignment algorithm[54]. The resulting MSAs were then used as input for Direct Coupling Analysis (DCA), where the aim is to fit a Potts model with functional form[31]:

$$P(x_i^n|x_{-i}^n) = \frac{1}{Z_i}\exp\left(v_i(x_i^n) + \sum_{j=1}^{L} w_{i,j}\left(x_i^n, x_j^n\right)\right) \quad (1)$$

where the conditional probability represents the information from the entire MSA given the model parameters v_i (representing position-wise information) and $w_{i,j}$ (representing pair-wise information). The notation x_{-i}^n denotes the MSA without accounting for the i:th sequence. Thus, for every sequence, all other sequences are used to estimate the probability of observing the parameters. Parameters were fitted using maximizing the pseudolikelihood of observing a set of N sequences of length L[55]:

$$\mathcal{L}\left(v, w \middle| (x_0^0, \dots, x_L^0) \dots (x_0^N, \dots, x_L^N)\right) = \sum_{n=1}^{N}\sum_{l=0}^{L} \ln P(x_i^n|x_{-i}^n, v, w) \quad (2)$$

Since the information about the pairwise information is contained within the w parameter, the evolutionary couplings $C_{i,j}$ in a coevolution matrix C were calculated according to equation 3, and afterwards standardized to the $\mathrm{N}(0,1)$ distribution.

$C_{i,j} = \sum_{A=0}^{20} \sum_{B=0}^{20} w_{i,j}(A,B)$ (3)

The resulting coevolution maps had an average false positive rate (FPR) of 8.4% for the top 500 coevolving pairs, which corresponds to an average of 42 contacts. The coevolution maps from each sequence cluster were combined into a global coevolution map by calculating the average coevolution score at each position. Since members of the family were picked roughly equidistantly in sequence space (a sequence identity of around 50%, where possible), coevolution maps were weighted equally. In general, this procedure ensures that family-wide features are upweighted, whereas features present in only a few coevolution maps are down-weighted.

Importantly, we did not choose to include all contacts, but merely the top t ones that were determined to be sufficient for distinguishing conformational states. *t* was determined using a procedure based on the available experimental structures. First, experimental contact maps were calculated, where contacts were estimated from distances between pairs of residues filtered by a sigmoid function:

$$M(A,B) = \frac{1}{1+\exp(2r_{A,B}-8)} \qquad (4)$$

where $r_{A,B}$ is the minimum distance between residues A and B.

Second, similarity between contact maps were defined as

$$d(x,y) = \langle C_{tot} \otimes M_x, C_{tot} \otimes M_y \rangle \qquad (5)$$

, which was used to calculate a distance matrix between all experimental structures. Third, UMAP[56] was used to obtain a 1D embedding of all experimental structures from the distance matrix, for which the pair-wise symmetrical KL-divergence[57] between the five distributions was calculated. This KL-divergence was minimized with respect to the parameter t. The described procedure yielded a t of 183.

State-specific structure prediction

A Convolutional Neural Network (CNN) was trained to classify the 5 functional states using contact maps from experimentally determined structures.

The input layer was of the size N^2xC, where N is the number of total number of MSA columns, and C the number of experimentally determined contact maps. The output

layer was a simple vector with 5 nodes, each corresponding to an experimentally characterized state.

The network architecture was designed to filter experimental contact maps by coevolution scores, and to avoid redundancy between adjacent residues involved in contacts. To implement these criteria, the input layer was followed by a filtering layer based on the optimized coevolution maps and by a pooling layer that gathered contacts formed by adjacent residue pairs. Lastly, two convolutions with sigmoid activation were applied to yield the output layer through an intermediate hidden layer. The 30-dimensional hidden layer was included to ensure that one contact could appear combinatorically in different states.

The loss function (equation 6) contained L2-regularization to prevent explosion of the weights.

$$L(e,d) = \sum_{i=0}^{n}(e_i - d_i)^2 + \sum_{i=0}^{n} {e_i}^2 \quad (6)$$

, where e is the value of the output nodes during training and d the target values of the same nodes. Training involved thorough regularization and constriction e.g. by the aforementioned pooling but as the training set only contained contact maps from 36 individual chains, many of which originating from the same deposition, overfitting the network was unavoidable[58]. The resulting model should thus not be used to make predictions.

To identify encouraged and discouraged contacts for each state, we performed layer-wise relevance backpropagation (LRP)[32] on all 5 output classes separately. Moreover, these scores are visualized for each feature contact pair with >0.1 relevance in figure 3, where highly scoring pairs in the pooling layer are highlighted with thick bands.

The contacts that scored more than 0.1 through LRP were used to apply attractive and repulsive biases to guide starting models towards the desired functional state. Specifically, initial structures for each family member were downloaded from the AlphaFold2 database website. Then, MultiConstraint (repulsive) and AmbiguousConstraint (attractive) bias functions were applied to all heavy atom distance pairs in the RosettaMP minimization scheme with implicit solvent and membrane[33]. For repulsive MultiConstraint, the built-in Rosetta fade function was applied. For attractive AmbiguousConstraint, simple flat harmonics were applied. The

force constants were deliberately chosen as weak (< 20 Rosetta standard units) in relation to the native forces of the all atom energy function called membrane_highres_Menv_smooth[59]. The weight of the added constraints was 0.1 as not to overshadow the natural energetics of the protein system. The protein-membrane topology was predicted using TOPCONS[60], which was accounted for using the AddMembraneMover. To fold the protein with the modified energy function, energy minimization with a Monte-Carlo component was applied using the provided fastrelax algorithm with 5 repeats. The optimization was conducted in cartesian space (using a pro_close weight of 0. and cart_bonded weight of 0.5 as recommended for cartesian minimization), rather than in Z-matrix form such that the constraints could be applied correctly.

Additionally, to assess the quality of the predicted structures, the RMSD towards the structure of the closest available relative in the same conformational state was calculated. The closest available relative was determined by the BLOSUM62 distance[61] matrix, after filtering for the appropriate conformational state.

Molecular Dynamics simulations

All Molecular Dynamics (MD) simulations were carried out in GROMACS2021[62]. The simulation systems containing the predicted structures of GLUT5 in the outward open, outward occluded, inward occluded and inward open states were prepared using the CHARMM-GUI membrane builder[63]. The systems contained the protein, embedded in a POPC bilayer plunged in a 0.1M KCl solution. The initial PBC box was 85x85x94 $Å^3$, ensuring at least 12.5 Å of water molecules between the protein and the PBC box end at least 10 lipid molecules between each PBC copy of the protein. The force field used was CHARMM36m for protein and lipids[64], and TIP3P[65] for water. The models were equilibrated using the default CHARMM-GUI scheme with one minimization step, and 6 100ps restraint cycles with gradually released restraints in the NPT ensemble, followed by a production simulation of 10ns. The simulations were carried out using a 2-fs time step. The target temperature and pressure were set to 303.15K and 1 bar respectively and maintained by a Nose-Hoover thermostat[66] (coupling separately protein, lipids and solvent) and a Parrinello-Rahman barostat[67] with semi-isotropic

coupling (p=5.0, compressibility 4.5*10-5). Hydrogen bonds were constrained using the linear constraint solver (LINCS)[68], and long-range electrostatics were accounted for using the particle mesh Ewald (PME) method beyond the 12Å electrostatic cutoff[69]. A neighbor-list cutoff was used for vdw interactions with r_{vdw}=12Å and a switching function starting at 10Å.

Collective variable determination

A support vector machine (SVM)[70] was trained on all predicted structures to separate adjacent states (outward occluded and outward open, outward occluded and inward occluded, and inward occluded and inward open). We avoided training on the occluded state given that only one sugar porter, pfHT1, was resolved in that state, possibly representing a functional outlier. To also keep track of the species-specific features of each model, we performed PCA[71] on the same training set (represented in the y-axis of figure S2, supplementary information). Indeed, the same procedure on the alphafold[27] input structures showed that the highest variance was in species-specifity, not conformational states. Ideally, we sought to preserve the species-specific features while only switching conformational states.

The highest SVM coefficients (> 0.193, as determined by the first gap in the histogram of coefficients) were divided into two separate components based on their sign. This yielded two collective variables, with CV_1 describing contacts specific to state 1 and CV_2 describing contacts specific to state 2:

$$CV_1 = \frac{1}{\sum_{i=0}^{N} \max(0, c_i)} \sum_{i=0}^{N} \max(0, c_i) * x_i \qquad (7)$$

$$CV_2 = \frac{1}{\sum_{i=0}^{N} \min(0, c_i)} \sum_{i=0}^{N} -\min(0, c_i) * x_i \qquad (8)$$

where x_i is the minimum distance between the two residues, and c_i the value of the coefficient from the SVM of contact i. The normalization factor was included to facilitate interpretability and to avoid assertion failures within the molecular dynamics code used. Given CV construction, a high value of CV_j thus corresponds to a low amount contacts specific to state j.

Accelerated Weight Histogram simulations

To reconstruct the complete conformational cycle of GLUT5, enhanced sampling simulations were run using the accelerated weighted histogram (AWH) method as implemented in GROMACS2021[62].

Three distinct AWH simulations were run to model each transition, using as CVs the relationships between distances described in eq. 7 and 8. These CVs were coupled to a reference coordinate $\xi(x)$ using a harmonic restraint function, using different force constants for each process (Table 1). The target distribution was chosen as uniform with a free-energy cutoff of 40-60 kJ/mol (Table 1) to avoid sampling of regions of high free energies. The free energy estimate was updated each 100 steps, gathering data from every previous tenth step.

During the initial phase, the update bias size is first held constant (using a diffusion constant, as specified in Table 1, and an initial error of 10 kJ/mol), and then divided by three each time the CVs covers the entire target region. After the number of visits at each point grows larger than the histogram size at that point, the initial phase is exited, after which the update size is continuously and exponentially decreased according to an exp-linear setting. All AWH simulations were run with 4 walkers in parallel with 2 starting in each conformational state at the extremes of each process.

Convergence of the AWH calculations was determined according to three criteria: the free energy landscape was stable over time, the coordinate distribution along the minimum free energy path reached a standard deviation of less than 0.1 from the mean coordinate distribution, and the changes in free energy estimate were below 0.5kJ/mol. The simulations were extended by 25 ns after these three criteria were met to ensure that these properties held true over time. The regions of high free energy (above the process-specific cutoff, see table 1) of the free energy landscape were excluded from the convergence analysis. To assess the convergence graphically, the 2-norm every 100ps was calculated (Figure S1).

Since four independent walkers were used a coverage diameter of 4Å was introduced, which forces each walker to explore an area of 4Å around every point before sharing the bias with all other walkers. In practice, this results in walkers overlapping in phase space, which provides each point with a free energy estimate based on multiple walkers.

Effectively, this provides an inherent CV quality control for the method as different walkers should produce consistent free energy estimates.

Seeing as all three processes (outward opening, rocker switch and inward opening) were run independently, we needed a method to combine the results into one free energy landscape. To this end, we calculated the minimum free energy paths between the deepest basins in the free energy landscapes by the following procedure. First, placing 60 points along a random path in the landscapes and relaxing the points by steepest decent on the landscape. Second, the position of the points was updated by forcing equidistance. The procedure was repeated until reaching convergence with a tolerance of 0.01Å. Due to the randomness in choosing the initial path and risk of ending up in a local minimum this procedure was repeated 100 times, where the minimum energy path was determined as the one with the lowest barrier. With the minimum free energy paths for each landscape calculated, the 2D free energy landscape was projected onto the path by Boltzmann averaging all the closest points to each point along the path. The 1D free energy landscapes where then stitched together by cutting the paths off where the minimum RMSD between to adjacent processes was found, yielding one concatenated free energy landscape for all three processes.

Acknowledgements This work was funded by the Knut and Alice Wallenberg Foundation (D.D) and the Science for Life Laboratory (D.D and L.D), the Göran Gustafsson Foundation (D.D and L.D) and the Swedish Research Council (VR 2019-02433 to D.D and L.D). The MD simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) on Beskow at the PDC Center for High Performance Computing (PDC-HPC).

Data availability The data necessary to reproduce this work is openly available at https://osf.io/qdmpx/

Author contributions Conceptualization DM, DD, LD ; Data Curation DM ; Formal Analysis DM ; Funding Acquisition DD, LD ; Investigation DM SM CA MB DD LD ; Methodology DM ; Project Administration DD, LD ; Resources DD, LD ; Software DM ; Supervision DD, LD; Validation DM SM CA MB DD LD ; Visualization DM, SM ; Writing – Original Draft Preparation DM DD LD ; Writing – Review & Editing DM SM CA MB DD LD

Table 1. Input parameters used for the AWH simulations. The force constant determines the resolution of the free energy surface and the bias towards the reference coordinate. The cover diameter determines the radius in CV space that has to be covered before a walker shares the bias at that point. The free-energy cutoff determined at which level sampling is deemed uninteresting. The convergence time is the simulation time per walker that was needed to achieve convergence according to the stated criteria.

	Diffusion Constant nm2ps-1		Force constant kJ mol-1nm-2		Cover diameter nm		Free-energy cutoff kJ mol-1	Convergence time ns/walker
	CV1	CV2	CV1	CV2	CV1	CV2		
Outward Opening	0.0005	0.001	10000	50000	0.4	0.4	60	374
Rocker switch	0.001	0.001	50000	50000	0.4	0.4	60	653
Inward Opening	0.0005	0.0005	30000	30000	0.4	0.4	40	86

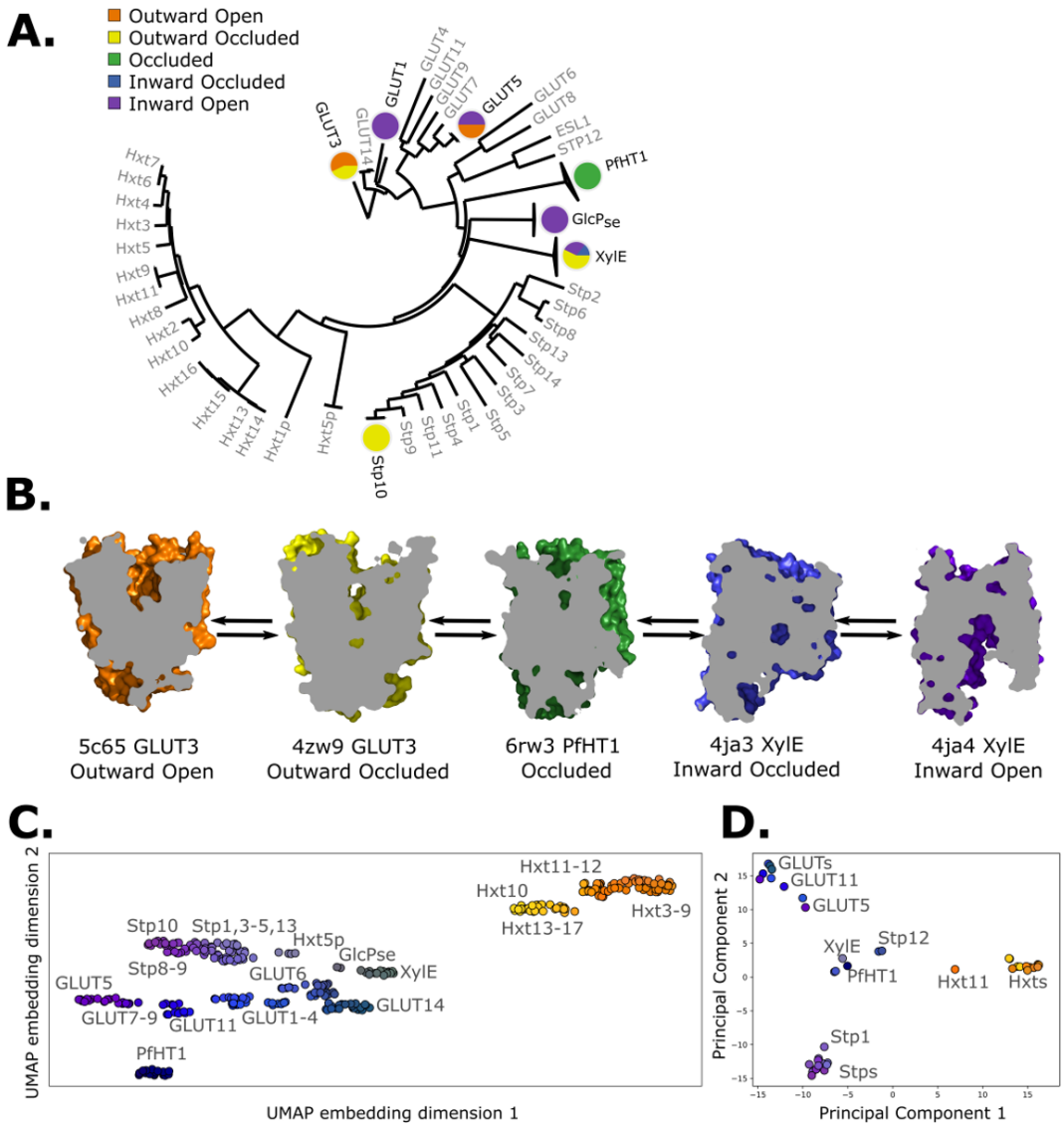

Figure 1. **Available structural and evolutionary information A.** The phylogenetic tree of the sugar transporter family, based on an multiple sequence alignment calculated by the MUSCLE algorithm. Tree generated using iTOL. The available experimental structures for each branch are highlighted as colored circles in the proportions that they appear in. The subfamilies represented in this tree are the mammalian sugar transporters (GLUTs), the bacterial and parasitic outliers (pfHT1, XylE and $GlcP_{se}$), the plant sugar transporters (STPs), and the fungal hexose transporters (HXTs). Sequences were retrieved from UniProtKB **B.** Clipped surfaces of representative structures of each functional state arranged according to the conformational cycle of sugar transporters. Note the alternating access to the intracellular and extracellular solvents throughout the cycle. **C.** A 2D embedding of the sugar transporter family in sequence space, using as similarity measure scores from the BLOSUM62-derived distance matrix. The embedding displays the phylogenetic distance between different branches of the family and SP are labeled as dots, colored according to their phylogenetic proximity. **D.** PCA projection of the available AlphaFold2 models in the sugar transporter family using as features residue-residue distances. SPs cluster according to phylogenetic proximity rather than conformational state

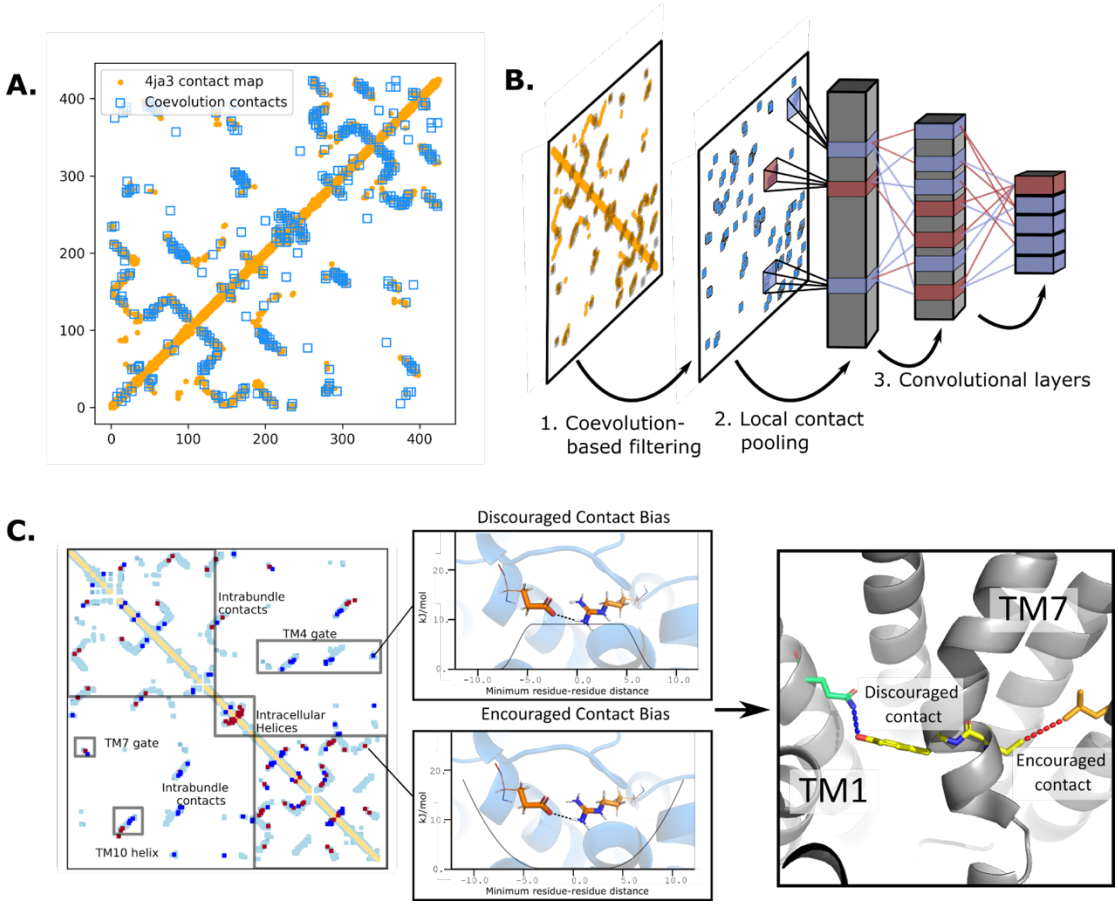

Figure 2. The state-specific structure prediction method. **A.** An overlay of the top 200 coevolving residue pairs (blue) and the experimental contact map of the bacterial XylE structure 4ja3 in an inward occluded state (orange). The secondary structure contacts from the coevolutionary analysis were omitted **B.** The architecture of the neural network used to train to predict conformational states from contacts between residue pairs. The first step consists in filtering out the low-coevolving contacts by the coevolution-based filtering, after which a restricted convolutional layer works to pool together close contacts in 4x4 grids, to account for the fact that contacts between adjacent positions often serve similar functions. Then, two unrestrained convolutions connect the aforementioned layer to the hidden layer and to the 5-dimensional output layer, each node representing a single conformational state. **C.** The bias application. The highly encouraged and discouraged contacts are translated into Ambiguous Constraint and Multi Constraint type biases in the RosettaMP fastrelax module, with the functional forms displayed in the middle panel. Using as input Alphafold2 models of each member in the family, the biases are applied and the energy is minimized in a Monte-Carlo fashion, repeated to generate 100 models. The final structure is chosen as the model with the lowest energy score value.

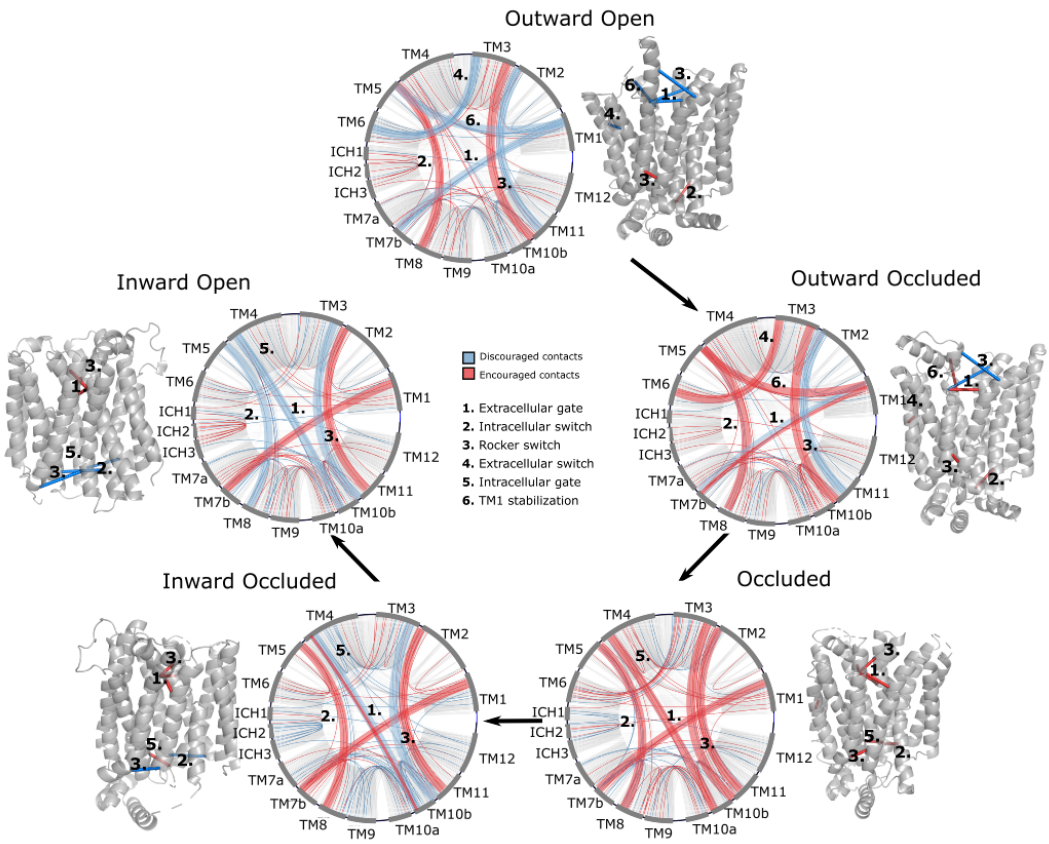

Figure 3. Network representations of the state-specific contact maps from the LRP analysis of the trained neural network. Nodes are labelled by the helix they are a part of, and the edges are colored by their sign in the LRP - blue represents discouraged contacts, and red represents encouraged contacts. Consensus contact maps of all states are shown in light grey in the background. Residue bundles that are encouraged or discouraged in a concerted manner (as revealed by their high importance in the pooling hidden layer of the neural network, Table S2) are highlighted with thick lines.

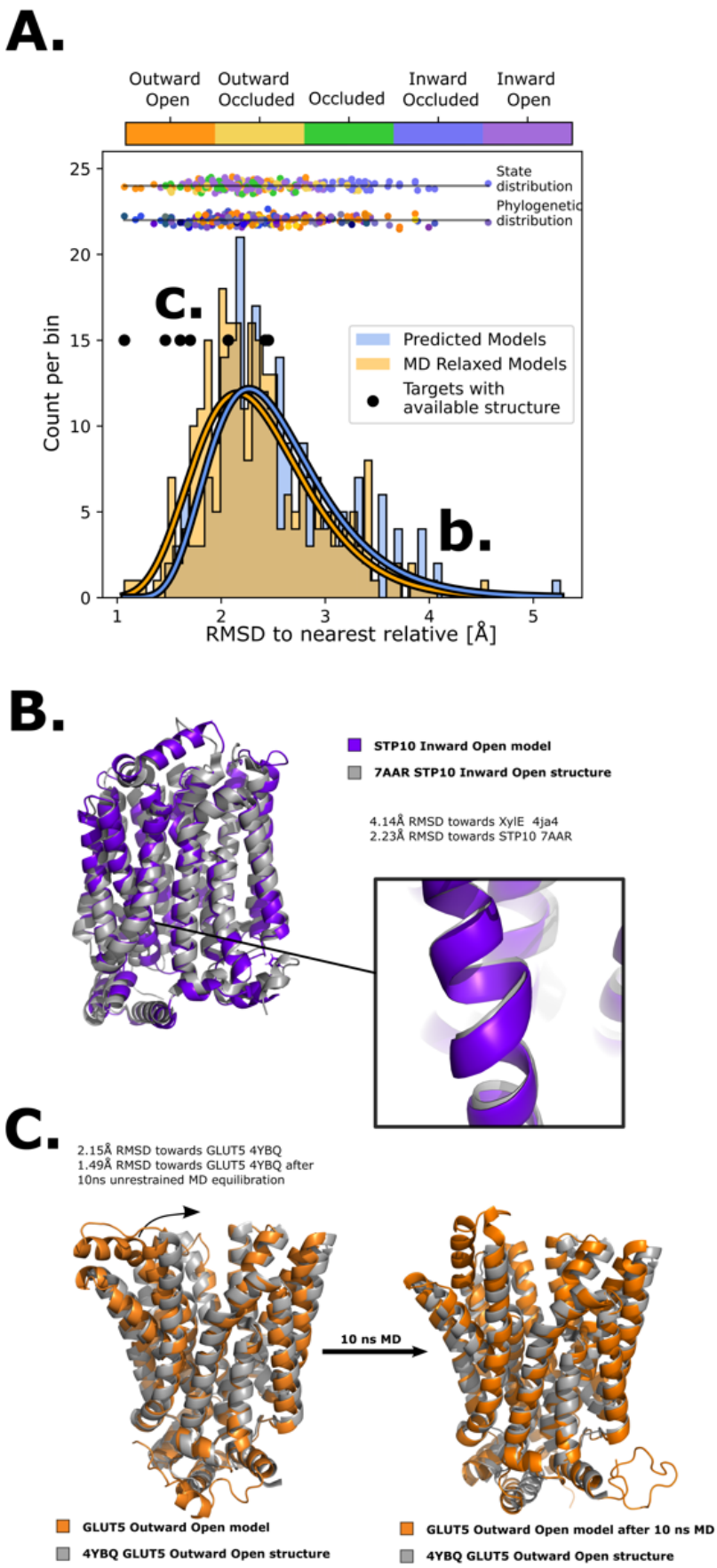
A.
Outward Open
Outward Occluded
Occluded
Inward Occluded
Inward Open
State distribution
Phylogenetic distribution
Count per bin
RMSD to nearest relative [Å]
Predicted Models
MD Relaxed Models
Targets with available structure
b.
c.
B.
STP10 Inward Open model
7AAR STP10 Inward Open structure
4.14Å RMSD towards XylE 4ja4
2.23Å RMSD towards STP10 7AAR
C.
2.15Å RMSD towards GLUT5 4YBQ
1.49Å RMSD towards GLUT5 4YBQ after 10ns unrestrained MD equilibration
10 ns MD
GLUT5 Outward Open model
4YBQ GLUT5 Outward Open structure
GLUT5 Outward Open model after 10 ns MD
4YBQ GLUT5 Outward Open structure

Figure 4. Quality of structure prediction of state-specific models of SPs A. Histogram of RMSD of the structural models (Rosetta, blue and MD-relaxed, orange) from their closest relative in the same conformational state. The distributions of the MD-relaxed structures colored according to state and phylogeny (see color definition in figure 1C) are shown above the histogram. Additionally, the targets with available experimental structures are indicated with black dots. **B.** Alignment between STP10 inward open model and the newly solved 7AAR STP10 structure in the same conformational state. A helical kink not present in any experimentally determined structure so far is shown as an example of a species-specific feature that is captured by our structure-prediction protocol. **C.** (left) Alignment between the GLUT5 outward-open model and an experimental structure of the same isoform in the same state (PDB ID 4YBQ). (right) Alignment between the MD-relaxed GLUT5 outward-open model and PDB 4YBQ.

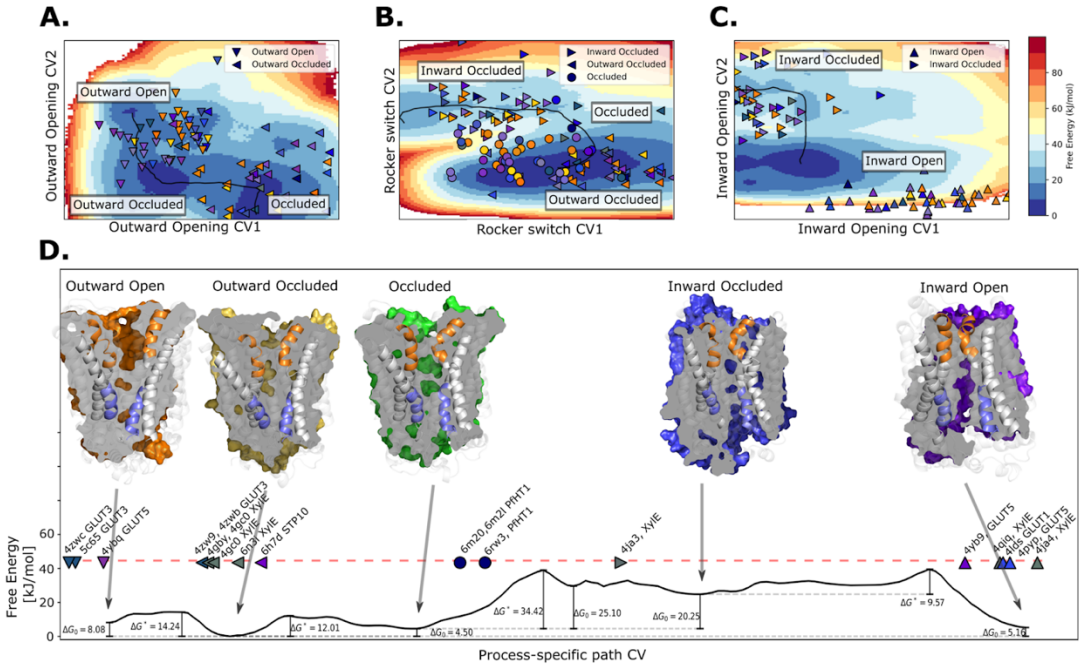

Figure 5. The entire conformational cycle is captured by accelerated weight histogram simulations. A. Free energy landscape of the outward opening process. The most populated free energy basins are labeled according to annotation based on visual inspection and RMSD calculations of snapshots from the basins to the available experimentally determined structures. Projections of models of the outward open and outward occluded models of the various SPs are shown as symbols colored according to figure 1C. The minimum free energy path connecting the outward open and occluded states is highlighted in black. Note that many outward occluded models fall in the basin representing the occluded state, most likely because of the small structural difference between a bent and broken TM7b helix. **B.** Free Energy landscape of the rocker switch process. Representation as in A. Note that although not trained on the occluded models, models mostly either fall in the barrier region or in the GLUT5 occluded basin. **C.** Free Energy landscape of the inward opening process. Representation as in A. **D.** 1-D free energy landscape of the full conformational change, determined by stitching together minimum free energy paths in panels A-C. The concatenation points were found by determining where along the respective minimum energy surfaces the maximum structural overlap (according to the RMSD metric) was found. Representative structural models of each of the most populated basins for are found at the top of the panel, and the projected positions of the experimental structures are shown along the dashed red line.

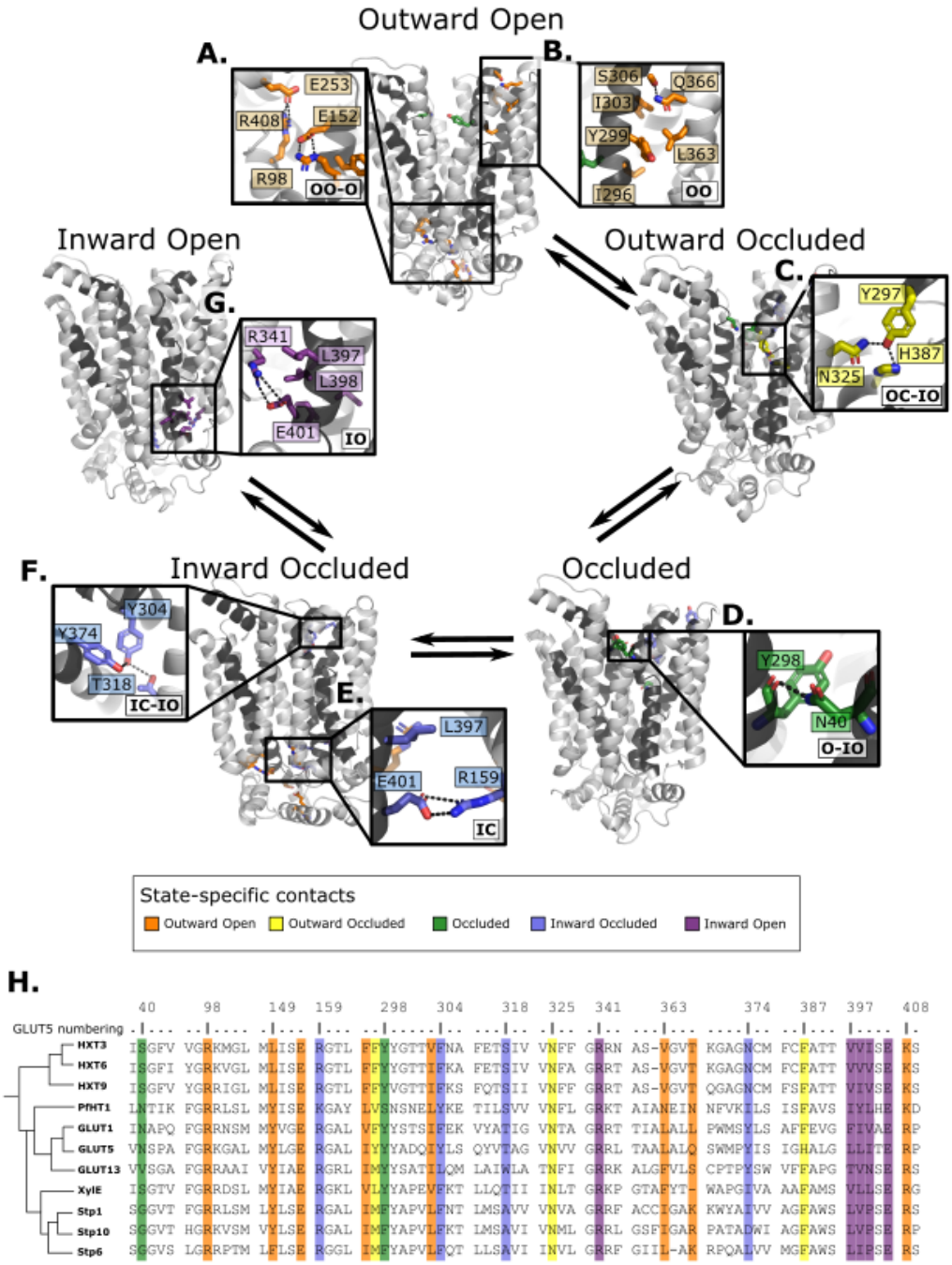

Figure 6. Summary of the structural determinants responsible for the cycling between adjacent conformational states. Contacts are characteristic of the first conformation they appear in, but they can be maintained more or less throughout the cycle depending on the family members. The snapshots are of the GLUT5 models used as a representative member. The bottom right of every panel contains information about which states the contacts are present in. (Outward Open/closed (OO/OC), occluded (O) or Inward Open/Closed (IO/IC)) **A.** Salt bridge network at the intracellular side. This network is intact in all outward facing states, but breaks during the rocker switch motion (except for the E401/R159 contact, see panel E). **B.** TM7b-TM9 interface responsible for stabilizing the outward-open conformation of TM7b, of which the hydrophobic contacts are only present in the outward open state. **C.** TM7b contact network

responsible for promoting both the bent and broken conformation of TM7b. The position of N325 and H387 necessitates the rotation of TM7b which occludes the extracellular gate. **D.** Backbone contact formed between TM7b and TM1, which is only possible when TM7b is completely broken. **E.** TM10b-TM4 interactions that are the last occluding contacts to break before the intracellular gate opens. **F.** TM10a-TM8 contacts responsible for stabilizing the new inter-bundle angle present in the inward facing states. **G.** Salt bridge and hydrophobic nexus responsible for stabilizing the inward-open conformation of TM10b, which fully unblocks the binding site from the intracellular side. **H.** MSA of some representative members at positions shown throughout panels A-G. Since model training was performed on all predicted models using highly coevolving residue pairs as features, the type of interaction present in different subfamilies can be tracked in this MSA.

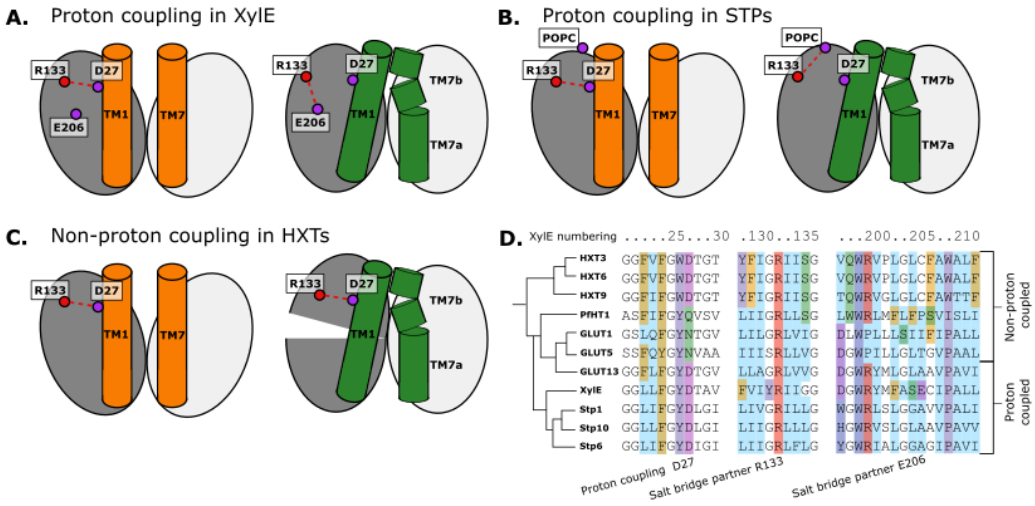

Figure 7. Structural and Evolutionary basis for proton coupling in the sugar transporter family. A. Proposed cartoon model for proton coupling in XylE, showing the highlighted salt bridge network rearrangements between the outward open (orange) and occluded (green) conformational states. Note how these rearrangements facilitate the closure of the extracellular gate. **B.** Proposed cartoon model for proton coupling in the STPs, showing the highlighted salt bridge network rearrangements between the outward open (orange) and occluded (green) conformational states. Note how these rearrangements facilitate the closure of the extracellular gate, and how R133 interacts with another latent salt bridge partner than in XylE. **C.** Proposed cartoon model explaining why HXTs generally lack proton coupling, showing the highlighted salt bridge network rearrangements between the outward open (orange) and occluded (green) conformational states. Note how the crucial R133-D27 contact can stay formed even in the occluded state due to a tilt of the entire helical bundle. **D.** MSA of a selected subset of proton-coupled and non-proton coupled sugar transporters, with residues labeled according to the XylE numbering.

References

1 Koepsell, H. Glucose transporters in brain in health and disease. *Pflugers Arch* **472**, 1299-1343 (2020). https://doi.org:10.1007/s00424-020-02441-x

2 Drew, D., North, R. A., Nagarathinam, K. & Tanabe, M. Structures and General Transport Mechanisms by the Major Facilitator Superfamily (MFS). *Chem Rev* (2021). https://doi.org:10.1021/acs.chemrev.0c00983

3 Reddy, V. S., Shlykov, M. A., Castillo, R., Sun, E. I. & Saier, M. H., Jr. The major facilitator superfamily (MFS) revisited. *FEBS J* **279**, 2022-2035 (2012). https://doi.org:10.1111/j.1742-4658.2012.08588.x

4 Wang, W. W., Gallo, L., Jadhav, A., Hawkins, R. & Parker, C. G. The Druggability of Solute Carriers. *J Med Chem* **63**, 3834-3867 (2020). https://doi.org:10.1021/acs.jmedchem.9b01237

5 Cesar-Razquin, A. *et al.* A Call for Systematic Research on Solute Carriers. *Cell* **162**, 478-487 (2015). https://doi.org:10.1016/j.cell.2015.07.022

6 Holman, G. D. Structure, function and regulation of mammalian glucose transporters of the SLC2 family. *Pflugers Arch* **472**, 1155-1175 (2020). https://doi.org:10.1007/s00424-020-02411-3

7 Mueckler, M. & Thorens, B. The SLC2 (GLUT) family of membrane transporters. *Mol Aspects Med* **34**, 121-138 (2013). https://doi.org:10.1016/j.mam.2012.07.001

8 Huang, S. & Czech, M. P. The GLUT4 glucose transporter. *Cell Metab* **5**, 237-252 (2007). https://doi.org:10.1016/j.cmet.2007.03.006

9 Rand, E. B., Depaoli, A. M., Davidson, N. O., Bell, G. I. & Burant, C. F. Sequence, tissue distribution, and functional characterization of the rat fructose transporter GLUT5. *Am J Physiol* **264**, G1169-1176 (1993).

10 Douard, V. & Ferraris, R. P. Regulation of the fructose transporter GLUT5 in health and disease. *Am J Physiol Endocrinol Metab* **295**, E227-237 (2008). https://doi.org:10.1152/ajpendo.90245.2008

11 Nino-Gonzalez, M., Novo-Uzal, E., Richardson, D. N., Barros, P. M. & Duque, P. More Transporters, More Substrates: The Arabidopsis Major Facilitator Superfamily Revisited. *Mol Plant* **12**, 1182-1202 (2019). https://doi.org:10.1016/j.molp.2019.07.003

12 Boles, E. & Hollenberg, C. P. The molecular genetics of hexose transport in yeasts. *FEMS Microbiol Rev* **21**, 85-111 (1997). https://doi.org:10.1111/j.1574-6976.1997.tb00346.x

13 Wang, M., Yu, C. & Zhao, H. Identification of an important motif that controls the activity and specificity of sugar transporters. *Biotechnol Bioeng* **113**, 1460-1467 (2016). https://doi.org:10.1002/bit.25926

14 Woodrow, C. J., Penny, J. I. & Krishna, S. Intraerythrocytic Plasmodium falciparum expresses a high affinity facilitative hexose transporter. *J Biol Chem* **274**, 7272-7277 (1999).

15 Blume, M. *et al.* A constitutive pan-hexose permease for the Plasmodium life cycle and transgenic models for screening of antimalarial sugar analogs. *FASEB J* **25**, 1218-1229 (2011). https://doi.org:10.1096/fj.10-173278

16 Qureshi, A. A. *et al.* The molecular basis for sugar import in malaria parasites. *Nature* **578**, 321-325 (2020). https://doi.org:10.1038/s41586-020-1963-z

17 Drew, D. & Boudker, O. Shared Molecular Mechanisms of Membrane Transporters. *Annu Rev Biochem* **85**, 543-572 (2016). https://doi.org:10.1146/annurev-biochem-060815-014520

18 Maiden, M. C., Davis, E. O., Baldwin, S. A., Moore, D. C. & Henderson, P. J. Mammalian and bacterial sugar transport proteins are homologous. *Nature* **325**, 641-643 (1987). https://doi.org:10.1038/325641a0

19 Quistgaard, E. M., Low, C., Moberg, P., Tresaugues, L. & Nordlund, P. Structural basis for substrate transport in the GLUT-homology family of monosaccharide transporters. *Nat Struct Mol Biol* **20**, 766-768 (2013). https://doi.org:10.1038/nsmb.2569

20 Deng, D. *et al.* Molecular basis of ligand recognition and transport by glucose transporters. *Nature* **526**, 391-396 (2015). https://doi.org:10.1038/nature14655

21 Nomura, N. *et al.* Structure and mechanism of the mammalian fructose transporter GLUT5. *Nature* **526**, 397-401 (2015). https://doi.org:10.1038/nature14909

22 Deng, D. *et al.* Crystal structure of the human glucose transporter GLUT1. *Nature* **510**, 121-125 (2014). https://doi.org:10.1038/nature13306

23 Deng, D. *et al.* Molecular basis of ligand recognition and transport by glucose transporters. *Nature* (2015). https://doi.org:10.1038/nature14655

24 Nomura N, V. G., Kang HJ, Shimamura T, Nomura Y, Sonoda Y, Hussien S.A, 	Qureshi A, Coincon M, Sato Y, Abe H, Nakada-Nakura Y, Hino T, Arakawa T, Kusano-	Arai O, Iwanari H, Unno H, Murata T, Kobayashi T, Hamakubo T, Kasahara M, Iwata 	S, Drew D. . Structure and mechanism of the mammalian fructose transporter GLUT5. *Nature* **in press** (2015).

25 Wisedchaisri, G., Park, M. S., Iadanza, M. G., Zheng, H. & Gonen, T. Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE. *Nat Commun* **5**, 4521 (2014). https://doi.org:10.1038/ncomms5521

26 Jiang, X. *et al.* Structural Basis for Blocking Sugar Uptake into the Malaria Parasite Plasmodium falciparum. *Cell* **183**, 258-268 e212 (2020). https://doi.org:10.1016/j.cell.2020.08.015

27 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2

28 Harpole, T. J. & Delemotte, L. Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations. *Biochim Biophys Acta Biomembr* **1860**, 909-926 (2018). https://doi.org:10.1016/j.bbamem.2017.10.033

29 Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *J Chem Phys* **151**, 070902 (2019). https://doi.org:10.1063/1.5109531

30 Del Alamo, D., Sala, D., McHaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11** (2022). https://doi.org:10.7554/eLife.75751

31 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* **108**, E1293-1301 (2011). https://doi.org:10.1073/pnas.1111471108

32 Bach, S. *et al.* On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **10**, e0130140 (2015). https://doi.org:10.1371/journal.pone.0130140

33 Alford, R. F. *et al.* An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLoS Comput Biol* **11**, e1004398 (2015). https://doi.org:10.1371/journal.pcbi.1004398

34 Bavnhoj, L., Paulsen, P. A., Flores-Canales, J. C., Schiott, B. & Pedersen, B. P. Molecular mechanism of sugar transport in plants unveiled by structures of glucose/H(+) symporter STP10. *Nat Plants* **7**, 1409-1419 (2021). https://doi.org:10.1038/s41477-021-00992-0

35 Yuan, Y. *et al.* Cryo-EM structure of human glucose transporter GLUT4. *Nat Commun* **13**, 2671 (2022). https://doi.org:10.1038/s41467-022-30235-5

36 Ke, M., Yuan, Y., Jiang, X., Yan, N. & Gong, H. Molecular determinants for the thermodynamic and functional divergence of uniporter GLUT1 and proton symporter XylE. *PLoS Comput Biol* **13**, e1005603 (2017). https://doi.org:10.1371/journal.pcbi.1005603

37 Park, M. S. Molecular Dynamics Simulations of the Human Glucose Transporter GLUT1. *PLoS One* **10**, e0125361 (2015). https://doi.org:10.1371/journal.pone.0125361

38 Galochkina, T., Ng Fuk Chong, M., Challali, L., Abbar, S. & Etchebest, C. New insights into GluT1 mechanics during glucose transfer. *Sci Rep* **9**, 998 (2019). https://doi.org:10.1038/s41598-018-37367-z

39 Chen, L. Y. & Phelix, C. F. Extracellular gating of glucose transport through GLUT 1. *Biochem Biophys Res Commun* **511**, 573-578 (2019). https://doi.org:10.1016/j.bbrc.2019.02.067

40 McComas S.E, M. D., Alleva C, Bonaccorsi M, Drew D, Delemotte L. Determinants of sugar-induced influx in the mammalian fructose transporter GLUT5. *bioRxiv* (2022). https://doi.org:https://doi.org/10.1101/2022.06.17.495601

41 Lindahl, V., Lidmar, J. & Hess, B. Riemann metric approach to optimal sampling of multidimensional free-energy landscapes. *Phys Rev E* **98**, 023312 (2018). https://doi.org:10.1103/PhysRevE.98.023312

42 Schurmann, A. *et al.* Role of conserved arginine and glutamate residues on the cytosolic surface of glucose transporters for transporter function. *Biochemistry* **36**, 12897-12902 (1997). https://doi.org:10.1021/bi971173c

43 Lowe, A. G. & Walmsley, A. R. The kinetics of glucose transport in human red blood cells. *Biochim Biophys Acta* **857**, 146-154 (1986).

44 Ezaki, O. & Kono, T. Effects of temperature on basal and insulin-stimulated glucose transport activities in fat cells. Further support for the translocation hypothesis of insulin action. *J Biol Chem* **257**, 14306-14310 (1982).

45 Selvam, B., Mittal, S. & Shukla, D. Free Energy Landscape of the Complete Transport Cycle in a Key Bacterial Transporter. *ACS Cent Sci* **4**, 1146-1154 (2018). https://doi.org:10.1021/acscentsci.8b00330

46 Takemoto, M., Lee, Y., Ishitani, R. & Nureki, O. Free Energy Landscape for the Entire Transport Cycle of Triose-Phosphate/Phosphate Translocator. *Structure* **26**, 1284-1296 e1284 (2018). https://doi.org:10.1016/j.str.2018.05.012

47 Sun, L. *et al.* Crystal structure of a bacterial homologue of glucose transporters GLUT1-4. *Nature* **490**, 361-366 (2012). https://doi.org:10.1038/nature11524

48 Santos Seica, A. F. *et al.* Asp22 drives the protonation state of the Staphylococcus epidermidis glucose/H(+) symporter. *J Biol Chem* (2020). https://doi.org:10.1074/jbc.RA120.014069

49 Jia, R. *et al.* Hydrogen-deuterium exchange mass spectrometry captures distinct dynamics upon substrate and inhibitor binding to a transporter. *Nat Commun* **11**, 6162 (2020). https://doi.org:10.1038/s41467-020-20032-3

50 Madej, M. G., Sun, L., Yan, N. & Kaback, H. R. Functional architecture of MFS D-glucose transporters. *Proc Natl Acad Sci U S A* **111**, E719-727 (2014). https://doi.org:10.1073/pnas.1400336111

51 Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015). https://doi.org:10.1093/bioinformatics/btu739

52 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997). https://doi.org:10.1093/nar/25.17.3389

53 Boyce, K., Sievers, F. & Higgins, D. G. Simple chained guide trees give high-quality protein multiple sequence alignments. *Proc Natl Acad Sci U S A* **111**, 10556-10561 (2014). https://doi.org:10.1073/pnas.1405628111

54 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004). https://doi.org:10.1186/1471-2105-5-113

55 Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* **110**, 15674-15679 (2013). https://doi.org:10.1073/pnas.1314045110

56 McInnes L, H. J., Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *The Journal of Open Source Software* **3** (2018). https://doi.org:10.21105/joss.00861

57 S. Kullback, R. A. L. On Information and Sufficiency. *The Annals of Mathematical Statistics.* **22** (1951).

58 Hawkins, D. M. The problem of overfitting. *J Chem Inf Comput Sci* **44**, 1-12 (2004). https://doi.org:10.1021/ci0342472

59 Alford, R. F., Fleming, P. J., Fleming, K. G. & Gray, J. J. Protein Structure Prediction and Design in a Biologically Realistic Implicit Membrane. *Biophys J* **118**, 2042-2055 (2020). https://doi.org:10.1016/j.bpj.2020.03.006

60 Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res* **37**, W465-468 (2009). https://doi.org:10.1093/nar/gkp363

61 Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992). https://doi.org:10.1073/pnas.89.22.10915

62 M. J. Abraham, T. M., R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19-25 (2015).

63 Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *J Comput Chem* **29**, 1859-1865 (2008). https://doi.org:10.1002/jcc.20945

64 Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71-73 (2017). https://doi.org:10.1038/nmeth.4067

65 Jorgensen W.L, C. J., Madura J.D. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79** (1983). https://doi.org:https://doi.org/10.1063/1.445869

66 S, N. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **81** (1984). https://doi.org:https://doi.org/10.1063/1.447334

67 Parrinello, M. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52** (1981). https://doi.org:https://doi.org/10.1063/1.328693

68 Hess, B. P-LINCS: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **4**, 116-122 (2008).

69 Darden, T. Y. D. P., L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **98**, 10089-10092 (1993).

70 Cortes C, V. V. Support-vector networks. *Machine Learning* **20**, 273–297 (1995). https://doi.org:10.1007/BF00994018

71 Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics* **2**, 559-572 (2010). https://doi.org:doi/abs/10.1080/14786440109462720