1

2

3

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

6

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

9 [1]Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of
10 Technology, Stockholm, Sweden. [2]Department of Biochemistry and Biophysics,
11 Science for Life Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden.

12

1 **Abstract**

2

3 **Sugar porters represent the largest group of secondary-active transporters. Some**

4 **members, such as the glucose (GLUT) transporters, are well-known for their role**

5 **in maintaining blood glucose homeostasis in mammals, with their expression**

6 **upregulated in many types of cancers. Because only a few sugar porter structures**

7 **have been determined, mechanistic models have been constructed by piecing**

8 **together structural states of distantly-related proteins. Current GLUT transport**

9 **models are predominantly descriptive, and oversimplified. Here, we have**

10 **combined coevolution analysis and comparative modeling, to predict structures of**

11 **the entire sugar porter superfamily in each state of the transport cycle. We have**

12 **analysed the state-specific contacts inferred from coevolving residue pairs, and**

13 **shown how this information can be used to rapidly generate free-energy**

14 **landscapes consistent with experimental estimates, as illustrated here for the**

15 **mammalian fructose transporter GLUT5. By comparing many different sugar**

16 **porter models and scrutinizing their sequence, we have been able to define the**

17 **molecular determinants of the transport cycle, which are conserved throughout**

18 **the sugar porter superfamily. We have also been able to highlight differences**

19 **leading to the emergence of proton-coupling, validating and extending the**

20 **previously proposed latch mechanism. Our computational approach is**

21 **transferable to any transporter, and to other protein families in general.**

22

1 Introduction

2 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.

22 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].

3

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

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

20

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

34

4

1   **Results**

2

3   *Inferring co-evolving conserved residue pairs across all sugar porters*

4   AlphaFold2 structural models of the various sugar porters are thought to collectively

5   capture all major conformational states[30]. However, deterministic prediction of a given

6   state is not possible. Therefore, as expected, principal component analysis (PCA) based

7   on pair-wise interactions reveals that sugar porters cluster within their individual

8   subfamilies rather than according to the conformational state that happens to be

9   predicted by the structure-prediction method (Fig. 1C, D). To be able to rationally build

10   models of individual SPs in each of the states in the transport cycle, we thus need to

11   steer AlphaFold2 generated models, or experimentally captured structures, when

12   applicable, into alternative states. We achieve this by using state-specific contacts

13   derived from coevolution analysis.

14   To do so, we first generated a representative sequence alignment by aligning ~1000

15   sequences from each of the evolutionarily-distant sugar porter relatives belonging to

16   mammalian sugar transporters (GLUTs), the bacterial and parasitic transporters

17   (*Pf*HT1, XylE), the plant sugar transporters (STPs), and the fungal hexose transporters

18   (HXT), separately (Figure 1A**)**. These sugar porter family members were selected

19   because they are either part of a major evolutionary branch of the sugar porter family

20   or are functionally distinct and structurally characterized (as is the case for *Pf*HT1 or

21   XylE). We then used the resulting multiple sequence alignments (MSA) as input for

22   Direct Coupling Analysis (DCA) to generate coevolution maps[31]. The coevolution

23   maps generated from separately aligned MSAs were subsequently combined into a

24   global coevolution map, filtering out contacts that were predominantly species-specific

25   (see Methods). As illustrated for the proton-coupled xylose transporter XylE (Figure

26   2A), there is an extensive overlap between the MSA-derived coevolution contact maps

27   and experimental inward-facing structural contacts. Nevertheless, several coevolving

28   pairs do not correspond to interactions found in the structure. Overall, we estimate that

29   ~10% of the top 500 co-evolving pairs represent contacts not formed in available crystal

30   structures, presumably representing contacts forming in other conformational states

31   (see Methods).

32

1   *Extracting state-dependent coevolving residue pairs from experimental structures*

2   To determine which coevolving pairs, correspond to contacts that are formed or broken

3   in the 20 structures of different functional states, we trained a convolutional neural

4   network (CNN) to classify the main five conformational states using as input contact

5   maps from experimentally determined structures filtered by coevolution scores (Figure

6   2B). The CNN architecture was designed to avoid redundancy of contacts between

7   neighboring residue while allowing residue pairs to be in contact in several functional

8   states (Methods, Figure 2B). Layer-wise relevance backpropagation (LRP) was then

9   performed[32] on all five output classes separately, leading to the assignment of

10   encouraged and discouraged contacts for each conformation (Figure 2C).

11   Interestingly, encouraged and discouraged contacts were not confined to inter-bundle

12   contacts and are instead spread throughout the entire structure (Table S1). They also

13   displayed strong state-dependency (Figure 3). Because of the low number of

14   experimental structures available, training of the neural network carried a substantial

15   risk of overfitting. We thus sought to validate the results by making sure that the state-

16   specific contacts identified were indeed in the expected regions. Starting from the

17   outward-facing conformation, we observe the expected encouraged contacts of inter-

18   bundle salt-bridges formed between TM3-TM10b and TM5-TM8 helices. In contrast,

19   we observe a strong signal for the discouraged contacts between TM1 and the

20   extracellular gate TM7b, which are known to come together during rocker-switch

21   transition into the inward-facing states[2,21]. In the occluded structure of *Pf*HT1[16], the

22   extracellular gate TM7b was had moved fully inwards and transition into a broken helix

23   at the point closest to TM1. Furthermore, mutations in TM1 were found to be just as

24   critical for transport as those in TM7b[16], indicating that TM1 and TM7b interactions

25   might be important in driving formation of the occluded state. Indeed, we observe a

26   robust signal for encouraged contacts between TM1 and TM7b, which is present as

27   early as the occluded state, confirming that TM7b and the interactions with TM1 are

28   co-evolving to enable attainment of the occluded conformation. As expected for an

29   intermediate state, the occluded state has the maximal number of encouraged contacts.

30   One additional set of encouraged contacts in the occluded state are located between the

31   intracellular gating helix TM10b and TM4, which are required to drive formation of the

32   occluded state from the inward-facing conformation[2,25].

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

13

14 *Generating sugar porter structures for each major conformation*
15 To expand the available structural information for sugar porters, AlphaFold2 models
16 were driven into all five conformational states in RosettaMP[33] using the derived
17 encouraged and discouraged state-dependent contacts as attractive and repulsive biases,
18 respectively (Figure 2C, Methods). A straightforward assessment of the quality of the
19 models obtained was difficult to perform, however, as few experimental structures are
20 available. Nevertheless, for an overall approximation, we calculated the Cα root mean
21 squared deviation (RMSD) to the phylogenetically closest available experimental
22 structure. As shown in Figure 4, the mean RMSD difference has an acceptable value of
23 ~2.5 Å, with the RMSD distribution spanning a 95% confidence interval of 1.5 to 3.2
24 Å. In addition, to evaluate the stability of these models, 10 ns molecular dynamics
25 simulations of these models embedded in a model membrane bilayer were then
26 performed, which further relaxed the structures, bringing their overall RMSD to the
27 closest homolog to ~2.1 Å with the RMSD distribution spanning a 95% confidence
28 interval of 1.1 to 2.7 Å. The RosettaMP modeling pipeline thus seems to bring the
29 models close to their local free energy minimum, but relaxation by MD simulations
30 further improves the quality of the models by relaxing them even more. To evaluate the
31 cause of the RMSD reduction, we compared the GLUT5 outward-open state model
32 before and after MD relaxation to the crystal structure of the same transporter solved in
33 the same state (PDB ID 4YBQ), which showed that the model quality was improved

1 from a 2.15 Å Cα RMSD to a 1.49 Å Cα RMSD, mainly via the relaxation of TM1

2 (Figure 4C). It seems that, in this case, flexible helices require an explicit solvent model

3 to attain a more physiological conformation.

4 As a next assessment of quality, we re-analyzed the STP10 model against the inward-

5 open STP10 crystal structure (PDB ID 7AAR)[34], whose structure was not present in our

6 initial training set as its structure was released after training of our model. The inward-

7 open state model of STP10 previously had a calculated Cα RMSD of 4.3 Å away from

8 its closest homologue with known structure, XylE (PDB 4QIQ)[25]. However, the actual

9 Cα RMSD to the experimental STP10 inward-facing structure difference is only 2.2 Å,

10 which was further reduced to 1.8 Å after MD simulations. Inspection of the STP10

11 structure reveals a unique helical kink in TM10, which is not present in XylE (Figure

12 4B). An additional structure for inward-open GLUT4[35] was further determined after

13 our initial model training was finalized. Our inward-open GLIUT4 model has a RMSD

14 of 2.3 Å and 1.8 Å before and after MD relaxation, respectively. Reassuringly, our

15 method is thus able to model new structural features absent from the training set.

16 Indeed, the Cα RMSD estimated for HXT and STP members, as well as for models of

17 the inward occluded states, which have fewer structures available to properly assess

18 their accuracy, is generally higher (Figure 4A). Overall, we conclude that the generated

19 sugar porter models are of sufficient quality for linking conformational states of sugar

20 porters.

21

22 *A free-energy landscape for the fructose transporter GLUT5*

23 Characterizing robust free energy landscapes of conformational cycles of sugar

24 transporters will enable us to better understand the mechanistic basis for sugar transport.

25 Although free energy landscapes for GLUT transporters have been reported

26 previously[36-39], these were generated from a few structures only, and lacked the most

27 recent structure in the occluded conformation[2,16]. Indeed, as co-evolution analysis

28 confirms, the occluded state is an intermediate that has a number of important and

29 specific co-evolved pairs that are critical for linking the outward and inward-facing

30 conformations. In addition, the methods used to generate these landscapes did not

31 consider whether structural features captured in specific structures were transferable to

32 other family members. Here, we decided to generate a free energy landscape for the

1 fructose transporter GLUT5 using family-wide information in the form of coevolving

2 contacts. We chose GLUT5 because it is the only transporter with experimental

3 structures in both fully outward- and inward-facing conformations[21] and because our

4 recently determined free energy landscape using a more traditional enhanced sampling

5 method can serve as a comparison[40].

6 The most efficient enhanced sampling techniques require the choice of a low-

7 dimensional collective variable (CVs) set that encompass all degrees of freedom

8 implicated in a conformational transition[28,41]. Given that state-dependent contacts

9 constitute the driving force for a conformational transition, CVs based on those are well

10 suited to distinguish conformational states and to enhance transitions between them.

11 We thus trained a support vector machine (SVM) to distinguish the modelled structures

12 of adjacent states (Figure S2) and extracted therefrom top coevolving contacts, as

13 inferred from their high SVM coefficients. We then designed state-specific CVs as

14 weighted sums of distances between top state-specific coevolving pairs, using as

15 weights the SVM coefficients (See Methods). Finally, we ran accelerated weighted

16 histogram (AWH), an efficient enhanced sampling method natively available in

17 GROMACS, using these CVs as input. Accumulating 250-650 ns of MD simulations

18 enabled an extensive sampling of the conformational space and the estimation of the

19 corresponding free energy landscapes (Figure 5A-C) with satisfactory accuracy (Figure

20 S1). Extracting the free energy along the minimum free energy path linking adjacent

21 states made it possible to estimate a continuous 1d landscape that follows the minimum

22 free energy path linking the outward-facing and the inward-facing states via the three

23 occluded intermediates (Figure 5D, see Methods).

24 In the absence of substrate, the most favorable conformation for GLUT5 is the outward-

25 facing state, which is consistent with biochemical analysis and the salt-bridge network

26 on the intracellular side stabilizing this state[20,21,42]. The outward-open and outward-

27 occluded states appear to be of comparable free energies, and the barriers between the

28 outward-open and outward-occluded states are fairly low at 4 to 8 kJ/mol, respectively

29 (Figure 5D). This observation is consistent with the fact that the structure of GLUT3

30 has been determined in both outward-open and outward-occluded states, even in the

31 presence of a bound maltose[23]. The largest energetic barrier of ~34 kJ/mol is located

32 between the outward occluded and the inward-occluded state (Figure 5D). This barrier,

33 presumably arises in part from the breakage of the strictly-conserved salt-bridge

9

1   network, and is consistent with the activation barrier of 10 kcal/mol as estimated by

2   GLUT1 kinetics (43 kJ/mol) [43,44](Figure 6E). Notably, the relaxed GLUT5 occluded

3   state does not fall on the largest energetic barrier corresponding to the transition state

4   (Figure 5). However, in the presence of a substrate sugar, GLUT5 passes through a

5   transition state that closely matches the occluded state of *Pf*HT1[40]. Our analysis

6   indicates that coevolution-driven MD simulations are detecting an energetic minimum

7   for an occluded state prior to the transition state which has not been experimentally

8   observed (Figure S3). It is possible that the co-evolution analysis is detecting a pre-

9   transition state, which is likely to be lowly populated in the presence of a substrate

10   sugar. Taken together, we conclude that using evolutionary-based collective variables

11   enables us to obtain sugar porter conformational free energy landscapes using AWH

12   using simulations of the order of hundreds of nanoseconds. This represents an increase

13   in performance by at least an order of magnitude compared to all of previous

14   attempts[38,45,46].

15

16   *Family-wide state-dependent interactions*

17   Having validated that the top co-evolving pairs can be combined into collective

18   variables able to connect the different conformational states and describe the energetics

19   of transitions between them appropriately, we scrutinize the interactions that went into

20   the construction of these CVs by projecting them onto the models of GLUT5 states

21   (Figure 6). Notably, these interactions are more robustly defined than in the previous

22   analysis shown in Figure 3, since those were based on the analysis of a reduced set of

23   experimentally resolved structures. In addition, we track the functional diversity across

24   the SP family by extracting the types of residue interactions found at these sites (Figure

25   6).

26   Starting from the outward-open state, as expected, salt-bridge forming residues appear

27   to stabilize the outward-facing conformations, a network that is maintained through to

28   the occluded state (Figure 6A). In addition, the residue Ser306 located at the back of

29   TM7b has co-evolved to form an interaction to Gln366 interaction in TM9, an

30   interaction seen as Thr/Arg or Lys in the STP members and as between small and/or

31   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

11

1   (Figure 6G). The rotation of TM10b shifts the occluding Leu397 residue away from

2   intracellular pathway to the sugar binding site, opening the intracellular gate fully. This

3   state is also stabilized by additional strictly conserved hydrophobic contacts in the

4   TM10-TM7 interface (positions 397-400 in GLUT5, which rearrange from an

5   interaction with TM4 in the outward-facing and occluded state) (Figure 6G and S5).

6

7   *Coevolving residues support proton-coupling in sugar porters*

8   Having characterized the aspects of the sugar porter transport cycle that are likely

9   conserved across the family, we turn to family-specific features, with a focus on the

10   differences between passive and proton-coupled transporters, such as XylE. The current

11   working model is that an aspartic acid reside in TM1 (Asp27 in XylE) is allosterically

12   coupled to the sugar transport, but does not participate directly in sugar binding[2,36,48,49].

13   Indeed, the aspartic acid to asparagine mutant has the same sugar-binding affinity as

14   wildtype[50]. Rather, it is thought that the protonation of the TM1 aspartic acid is required

15   for transport as its breaks the outward-facing-specific salt-bridge interaction to an

16   arginine residue Arg133 on TM4. Based on the occluded *Pf*HT1 structure[2], highlighting

17   the coupling between TM1 and TM7b, a latch mechanism for proton-coupling was

18   proposed[2]. Simplistically, breakage of the Asp27-Arg133 salt-bridge enables TM1 to

19   come closer to TM7b and TM7b only comes close enough to TM1 when sugar is

20   present. Formation of the occluded state is catalyzed by sugar binding, but in proton

21   coupled transporters, this is only enabled by the release of the latch (Figure 7A and S4).

22   The co-evolution analysis is consistent with TM1 and TM7 interactions driving

23   formation of the occluded state. More specifically, we indeed see that the equivalent

24   TM1 Asp and TM4 Arg salt-bridge is broken in occluded states of all known proton-

25   coupled symporters, such as STP10. The details of the conformational rearrangement

26   differ, however. After breaking of the latch, in the occluded state of XylE, Arg133

27   forms a salt bridge with the neighboring TM6 Glu206 residue (Figure 7A). In STP10,

28   on the other hand, Glu206 is replaced by an alanine, and R133 rearranges by adopting

29   a different rotameric state facing towards TM11, where it appears to be stabilized by an

30   interaction with the phospholipid headgroups (Figure 7B and S4).

31   Most GLUTs are thought to be passive transporters and do not possess an acidic residue

32   at position 27, consistent with its purported role as proton carrier (Figure 7D).

12

1   Intriguingly, however HXTs, which are generally not proton-coupled, do carry an

2   acidic residue in this position. Our analysis reveals that this is explained by the fact that

3   these transporters contain a latch interaction that is maintained throughout the

4   conformational cycle thanks to a large-scale pivot of the first helical bundle (Figure 7C

5   and S4). This conformational change unique to HXTs is enabled by helix breaking

6   residues at position 202 (Pro/Gly), which are replaced by mostly hydrophobic residues

7   in the rest of the family (Figure 7D). GLUT2, a passive transporter that nevertheless

8   possesses an acidic residue at position 27, has a Serine in position 205 which interacts

9   with D27, which appears to help the salt bridge remain formed in the occluded state

10 (Figure S4D).

11

12 **Discussion**

13 Alphafold2 has made it possible to produce high-quality models of structures associated

14 with any amino acid sequence[27]. Nevertheless, these structures represent an arbitrary

15 conformational state, and do not capture the conformational heterogeneity that enables

16 protein function. Modifying the MSA depth enables to make models of alternate states,

17 though in a non-deterministic manner[30]. In this work, we have used i. AlphaFold

18 models, ii. coevolution analysis encoded by family-wide multiple sequence alignments

19 and iii. state-specific contacts extracted from the ~20 structures to guide enhanced MD

20 simulations. Our work goes beyond the analysis of static structures and enables the

21 separation of state-specific and transporter-specific features, thus firmly establishing

22 the family-wide determinants of the conformational cycle as well as establishing

23 specificities of GLUT5 and of proton-coupled sugar transporters. By introducing

24 coevolution as a measure of predicted interaction, we are able to develop models that

25 go beyond the comparison of structures only. We also minimize information that may

26 be lost during dimensionality reduction, as demonstrated by our validation on newly

27 reported structures.

28 Using state-dependent coevolution-based contacts, we show that we can create a low-

29 dimensional projection that describes the transition between adjacent states. When

30 derived into a set of collective variables used in enhanced sampling molecular dynamics

31 simulations, we can construct weighted conformational ensembles, or in other words

32 free energy landscapes of the process. Comparing our convergence times (to less than

13

1   0.01 kJ/mol within 250-650ns per walker and system) to previous works on similar
2   systems[38,45,46] and given the agreement between the activation barrier estimate from
3   kinetic measurements[43,44] we conclude that the introduction of coevolution data is a
4   powerful approach in obtaining accurate free-energy landscapes efficiently.

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

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

14

1   missing and the salt-bridge is still retained due to the assistance of a nearby serine

2   residue (Ser205, XylE numbering). Nevertheless, STP members are proton-coupled and

3   yet they do not possess an additional acidic residue in TM6. Rather, the arginine residue

4   is able to snorkel to the membrane interface in the unpaired state. Notably, although

5   TM7b and TM1 are important, they are part of a larger interaction network that also

6   includes, for example, how TM7b interacts with TM8 in the outward-open state.

7

8   Consistent with the orchestrating role of TM7b in substrate translocation, the co-

9   evolved interaction with TM1 is retained across the entire second half of the transport

10 cycle. In contrast, the inward-facing gating helix TM10b is likely to have a more

11 passive role and a co-evolved interaction with TM4 is only formed in the occluded state.

12 Nevertheless, TM10b dynamics are likely to be important and its possible that TM10b

13 movement is facilitated by the stabilization of TM10a with TM8 on the extracellular

14 side (Figure 6F). The bottom of TM10a harbors a strictly-conserved acidic residue,

15 which is part of the intracellular salt-bridge network. The contact maps initially

16 generated for all sugar porters indicated mostly discouraged contacts between TM4 and

17 TM10b residues in the inward-occluded state (bundle 5 contact in Figure 3). In addition,

18 the GLUT5 models show that whilst most contacts are indeed broken, there is still an

19 encouraged contact between TM4 and TM10b residues in the inward-occluded state

20 (Figure 6F, S5B). It thus appears that, at least in GLUT5, only one of the salt-bridge

21 pairs are fully broken between the occluded to inward-occluded states. Such a result is

22 intriguing, and implies that even the rocker-switch bundle movement might utilize

23 asymmetric rearrangements.

24 Lastly, GLUT transporters are shown as textbook examples of how small molecule

25 transporters are functional equivalents to enzymes[2]. Key to understanding catalysis is

26 to understand how the transition state is formed during substrate translocation. The

27 transition state, however, is only transiently occupied and therefore difficult to

28 experimentally capture. Nevertheless, the parasite transporter *Pf*HT1 has an unusually

29 very polar TM7b gate, which has made it possible to capture an occluded state with

30 sugar present[16,26] i.e., notably this state is unlikely to represent a transition state for

31 *Pf*HT1. Nonetheless, out of the models we built, *Pf*HT1 has highest number of state-

32 specific contacts in both gates. We speculate that the *Pf*HT1 could represent a pre-

1 transition state (Figure 5), with further dynamics between the bundles required to access

2 the transition state. In summary, although many questions remain unanswered, our

3 novel approach provides a rational framework for understanding how sugar porters

4 function at the molecular level, and provides the information to re-engineer sugar

5 porters with different characteristics, which might be otherwise inaccessible by

6 traditional forward-evolution approaches. At a conceptual level, our work highlights

7 how even the most simplistic type of transporters have evolved fine-tuned and intricate

8 interactions to achieve substrate translocation.

9

Methods

Coevolution analysis

12 To construct the sequence alignment at the basis of coevolution analysis, a

13 representative sequence from each member of the sugar transporter family was taken

14 as a seed for a sequence search against the Uniref90 database[51] using PSIBLAST[52]. In

15 this way, a sufficient number of diverse sequences was consistently found around each

16 member (M_{eff}>1000, where M_{eff} denotes the number of sequences with less than 80%

17 identity). The sequence libraries were then aligned within each subfamily using a stair-

18 shaped guide tree[53] in the MUSCLE alignment algorithm[54]. The resulting MSAs were

19 then used as input for Direct Coupling Analysis (DCA), where the aim is to fit a Potts

20 model with functional form[31]:

21 $$P(x_i^n|x_{-i}^n) = \frac{1}{Z_i}\exp\left(v_i(x_i^n) + \sum_{j=1}^{L} w_{i,j}\left(x_i^n, x_j^n\right)\right) \quad (1)$$

22 where the conditional probability represents the information from the entire MSA given

23 the model parameters v_i (representing position-wise information) and $w_{i,j}$ (representing

24 pair-wise information). The notation x_{-i}^n denotes the MSA without accounting for the

25 i:th sequence. Thus, for every sequence, all other sequences are used to estimate the

26 probability of observing the parameters. Parameters were fitted using maximizing the

27 pseudolikelihood of observing a set of N sequences of length L[55]:

28 $$\mathcal{L}\left(v,w\middle|(x_0^0,\dots,x_L^0)\dots(x_0^N,\dots,x_L^N)\right) = \sum_{n=1}^{N}\sum_{l=0}^{L}\ln P(x_i^n|x_{-i}^n, v, w) \quad (2)$$

29 Since the information about the pairwise information is contained within the w

30 parameter, the evolutionary couplings $C_{i,j}$ in a coevolution matrix C were calculated

31 according to equation 3, and afterwards standardized to the $N(0,1)$ distribution.

16

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

2   The resulting coevolution maps had an average false positive rate (FPR) of 8.4% for

3   the top 500 coevolving pairs, which corresponds to an average of 42 contacts. The

4   coevolution maps from each sequence cluster were combined into a global coevolution

5   map by calculating the average coevolution score at each position. Since members of

6   the family were picked roughly equidistantly in sequence space (a sequence identity of

7   around 50%, where possible), coevolution maps were weighted equally. In general, this

8   procedure ensures that family-wide features are upweighted, whereas features present

9   in only a few coevolution maps are down-weighted.

10   Importantly, we did not choose to include all contacts, but merely the top t ones that

11   were determined to be sufficient for distinguishing conformational states. *t* was

12   determined using a procedure based on the available experimental structures. First,

13   experimental contact maps were calculated, where contacts were estimated from

14   distances between pairs of residues filtered by a sigmoid function:

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

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

17   Second, similarity between contact maps were defined as

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

19   , which was used to calculate a distance matrix between all experimental structures.

20   Third, UMAP[56] was used to obtain a 1D embedding of all experimental structures from

21   the distance matrix, for which the pair-wise symmetrical KL-divergence[57] between the

22   five distributions was calculated. This KL-divergence was minimized with respect to

23   the parameter t. The described procedure yielded a t of 183.

24

25   **State-specific structure prediction**

26   A Convolutional Neural Network (CNN) was trained to classify the 5 functional states

27   using contact maps from experimentally determined structures.

28   The input layer was of the size $N^2 \times C$, where N is the number of total number of MSA

29   columns, and C the number of experimentally determined contact maps. The output

1 layer was a simple vector with 5 nodes, each corresponding to an experimentally

2 characterized state.

3 The network architecture was designed to filter experimental contact maps by

4 coevolution scores, and to avoid redundancy between adjacent residues involved in

5 contacts. To implement these criteria, the input layer was followed by a filtering layer

6 based on the optimized coevolution maps and by a pooling layer that gathered contacts

7 formed by adjacent residue pairs. Lastly, two convolutions with sigmoid activation

8 were applied to yield the output layer through an intermediate hidden layer. The 30-

9 dimensional hidden layer was included to ensure that one contact could appear

10 combinatorically in different states.

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

12 weights.

13 $L(e, d) = \sum_{i=0}^{n}(e_i - d_i)^2 + \sum_{i=0}^{n} e_i^2$ (6)

14 , where e is the value of the output nodes during training and d the target values of the

15 same nodes. Training involved thorough regularization and constriction e.g. by the

16 aforementioned pooling but as the training set only contained contact maps from 36

17 individual chains, many of which originating from the same deposition, overfitting the

18 network was unavoidable[58]. The resulting model should thus not be used to make

19 predictions.

20 To identify encouraged and discouraged contacts for each state, we performed layer-

21 wise relevance backpropagation (LRP)[32] on all 5 output classes separately. Moreover,

22 these scores are visualized for each feature contact pair with >0.1 relevance in figure 3,

23 where highly scoring pairs in the pooling layer are highlighted with thick bands.

24 The contacts that scored more than 0.1 through LRP were used to apply attractive and

25 repulsive biases to guide starting models towards the desired functional state.

26 Specifically, initial structures for each family member were downloaded from the

27 AlphaFold2 database website. Then, MultiConstraint (repulsive) and

28 AmbiguousConstraint (attractive) bias functions were applied to all heavy atom

29 distance pairs in the RosettaMP minimization scheme with implicit solvent and

30 membrane[33]. For repulsive MultiConstraint, the built-in Rosetta fade function was

31 applied. For attractive AmbiguousConstraint, simple flat harmonics were applied. The

18

1 force constants were deliberately chosen as weak (< 20 Rosetta standard units) in

2 relation to the native forces of the all atom energy function called

3 membrane_highres_Menv_smooth[59]. The weight of the added constraints was 0.1 as

4 not to overshadow the natural energetics of the protein system. The protein-membrane

5 topology was predicted using TOPCONS[60], which was accounted for using the

6 AddMembraneMover. To fold the protein with the modified energy function, energy

7 minimization with a Monte-Carlo component was applied using the provided fastrelax

8 algorithm with 5 repeats. The optimization was conducted in cartesian space (using a

9 pro_close weight of 0. and cart_bonded weight of 0.5 as recommended for cartesian

10 minimization), rather than in Z-matrix form such that the constraints could be applied

11 correctly.

12 Additionally, to assess the quality of the predicted structures, the RMSD towards the

13 structure of the closest available relative in the same conformational state was

14 calculated. The closest available relative was determined by the BLOSUM62 distance[61]

15 matrix, after filtering for the appropriate conformational state.

16

17

18 **Molecular Dynamics simulations**

19 All Molecular Dynamics (MD) simulations were carried out in GROMACS2021[62]. The

20 simulation systems containing the predicted structures of GLUT5 in the outward open,

21 outward occluded, inward occluded and inward open states were prepared using the

22 CHARMM-GUI membrane builder[63]. The systems contained the protein, embedded in

23 a POPC bilayer plunged in a 0.1M KCl solution. The initial PBC box was 85x85x94

24 $Å^3$, ensuring at least 12.5 Å of water molecules between the protein and the PBC box

25 end at least 10 lipid molecules between each PBC copy of the protein. The force field

26 used was CHARMM36m for protein and lipids[64], and TIP3P[65] for water. The models

27 were equilibrated using the default CHARMM-GUI scheme with one minimization

28 step, and 6 100ps restraint cycles with gradually released restraints in the NPT

29 ensemble, followed by a production simulation of 10ns. The simulations were carried

30 out using a 2-fs time step. The target temperature and pressure were set to 303.15K and

31 1 bar respectively and maintained by a Nose-Hoover thermostat[66] (coupling separately

32 protein, lipids and solvent) and a Parrinello-Rahman barostat[67] with semi-isotropic

19

1   coupling (p=5.0, compressibility 4.5*10-5). Hydrogen bonds were constrained using

2   the linear constraint solver (LINCS)[68], and long-range electrostatics were accounted for

3   using the particle mesh Ewald (PME) method beyond the 12Å electrostatic cutoff[69]. A

4   neighbor-list cutoff was used for vdw interactions with r_{vdw}=12Å and a switching

5   function starting at 10Å.

6

Collective variable determination

8   A support vector machine (SVM)[70] was trained on all predicted structures to separate

9   adjacent states (outward occluded and outward open, outward occluded and inward

10   occluded, and inward occluded and inward open). We avoided training on the occluded

11   state given that only one sugar porter, pfHT1, was resolved in that state, possibly

12   representing a functional outlier. To also keep track of the species-specific features of

13   each model, we performed PCA[71] on the same training set (represented in the y-axis of

14   figure S2, supplementary information). Indeed, the same procedure on the alphafold[27]

15   input structures showed that the highest variance was in species-specifity, not

16   conformational states. Ideally, we sought to preserve the species-specific features while

17   only switching conformational states.

18   The highest SVM coefficients (> 0.193, as determined by the first gap in the histogram

19   of coefficients) were divided into two separate components based on their sign. This

20   yielded two collective variables, with CV_1 describing contacts specific to state 1 and

21   CV_2 describing contacts specific to state 2:

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

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

24   where x_i is the minimum distance between the two residues, and c_i the value of the

25   coefficient from the SVM of contact i. The normalization factor was included to

26   facilitate interpretability and to avoid assertion failures within the molecular dynamics

27   code used. Given CV construction, a high value of CV_j thus corresponds to a low

28   amount contacts specific to state j.

29

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.

21

1   Effectively, this provides an inherent CV quality control for the method as different

2   walkers should produce consistent free energy estimates.

3   Seeing as all three processes (outward opening, rocker switch and inward opening) were

4   run independently, we needed a method to combine the results into one free energy

5   landscape. To this end, we calculated the minimum free energy paths between the

6   deepest basins in the free energy landscapes by the following procedure. First, placing

7   60 points along a random path in the landscapes and relaxing the points by steepest

8   decent on the landscape. Second, the position of the points was updated by forcing

9   equidistance. The procedure was repeated until reaching convergence with a tolerance

10   of 0.01Å. Due to the randomness in choosing the initial path and risk of ending up in a

11   local minimum this procedure was repeated 100 times, where the minimum energy path

12   was determined as the one with the lowest barrier. With the minimum free energy paths

13   for each landscape calculated, the 2D free energy landscape was projected onto the path

14   by Boltzmann averaging all the closest points to each point along the path. The 1D free

15   energy landscapes where then stitched together by cutting the paths off where the

16   minimum RMSD between to adjacent processes was found, yielding one concatenated

17   free energy landscape for all three processes.

18

19

7

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

10

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

17
18
19
20
21
22

23

24

25

26

27

28

29

30

31

32

1

2

3

4

5

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

12

	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

13
14

15
16
17
18

19

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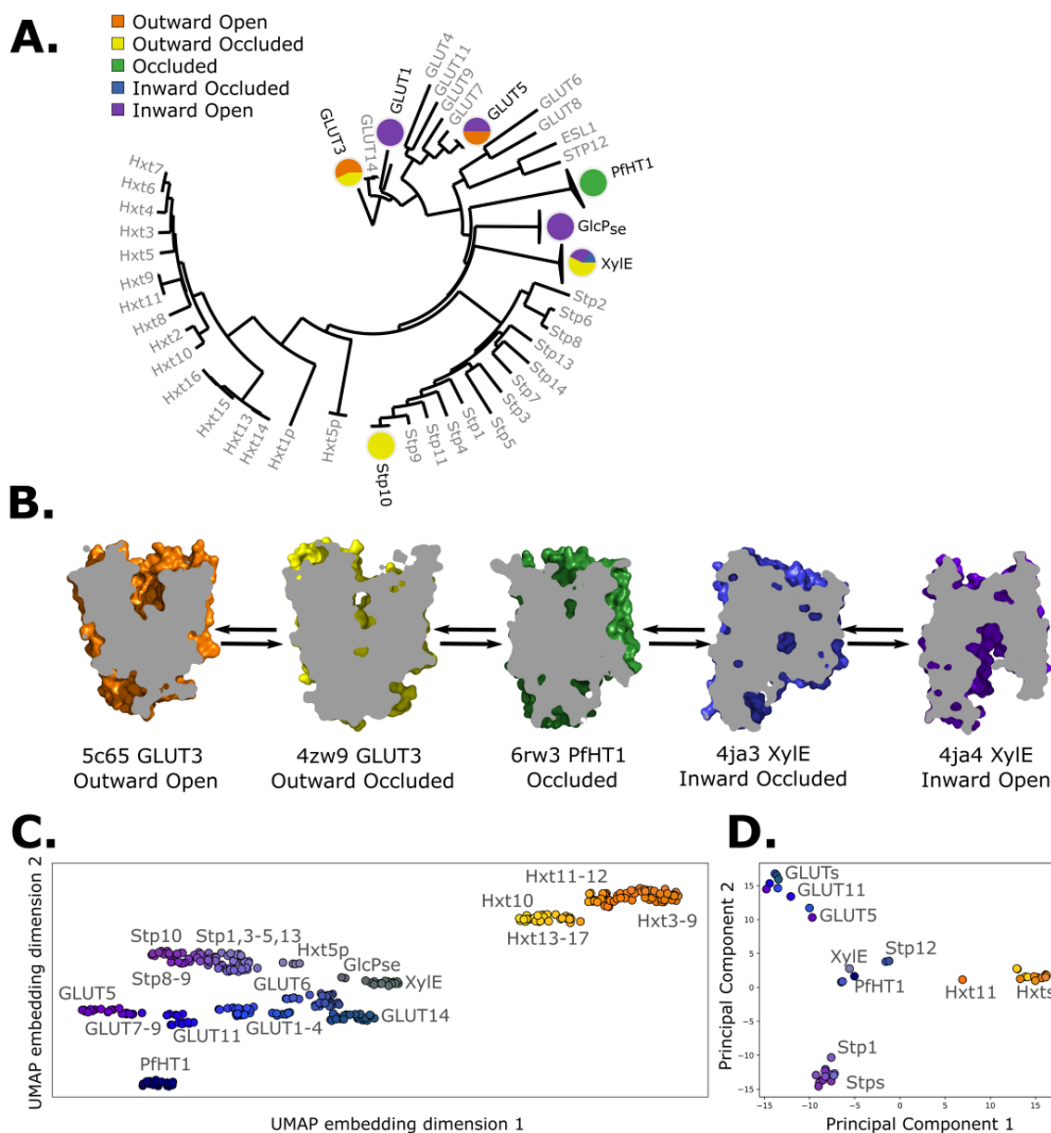
2
3 **Figure 1**. **Available structural and evolutionary information A.** The phylogenetic
4 tree of the sugar transporter family, based on an multiple sequence alignment calculated
5 by the MUSCLE algorithm. Tree generated using iTOL. The available experimental
6 structures for each branch are highlighted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

25
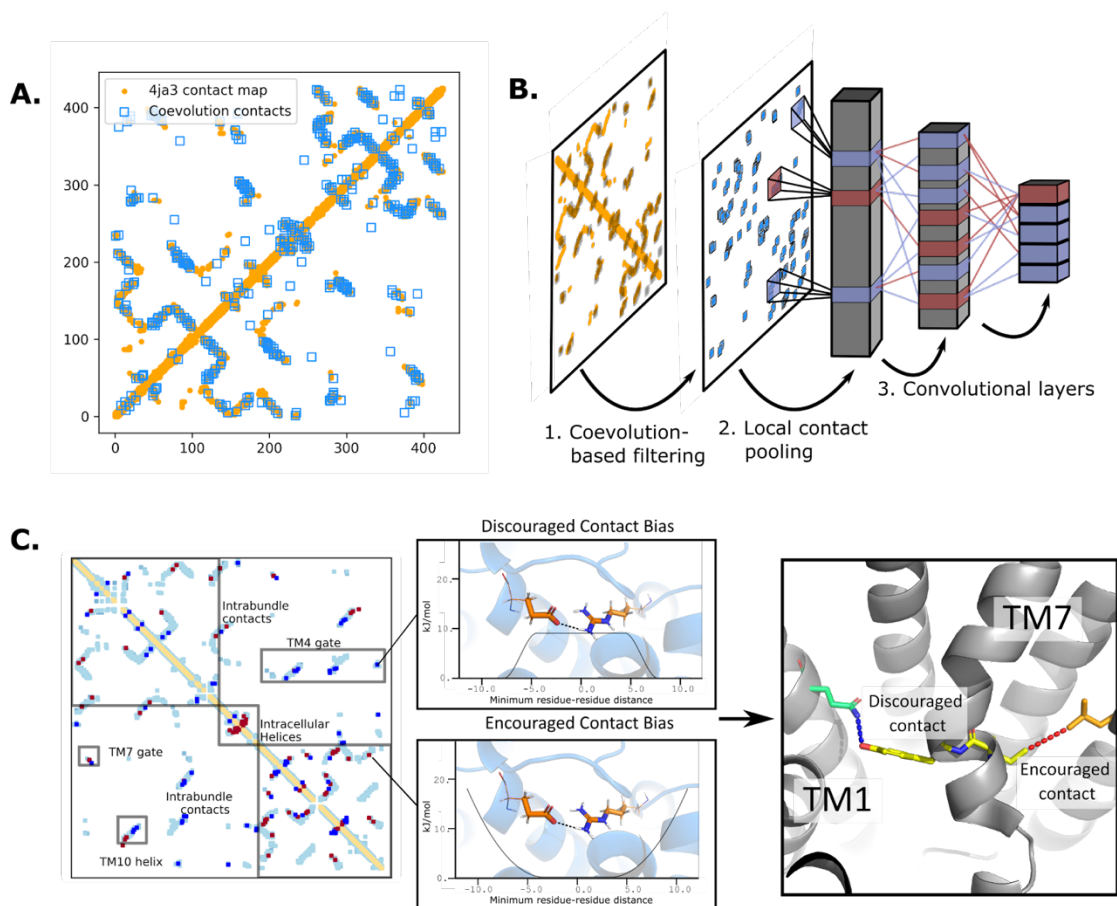
Figure 2. The state-specific structure prediction method. **A.** An overlay of the top 200 coevolving residue pairs (blue) and the experimental contact map of the bacterial XylE structure 4ja3 in an inward occluded state (orange). The secondary structure contacts from the coevolutionary analysis were 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.

1

2

3

4

5
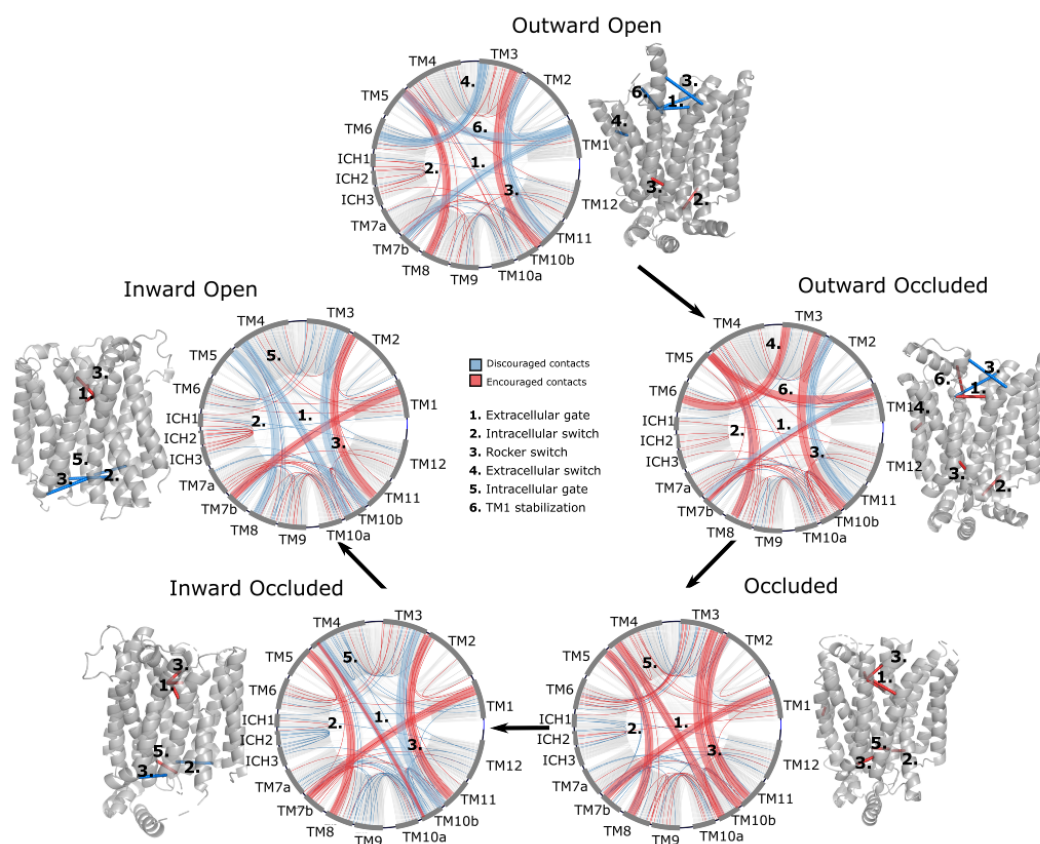
Figure 3. Network representations of the state-specific contact maps from the LRP analysis of the trained neural network. Nodes are labelled by the helix they are a part of, and the edges are colored by their sign in the LRP - blue represents discouraged contacts, and red represents encouraged contacts. 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.

13

14

15

16

17

18

19
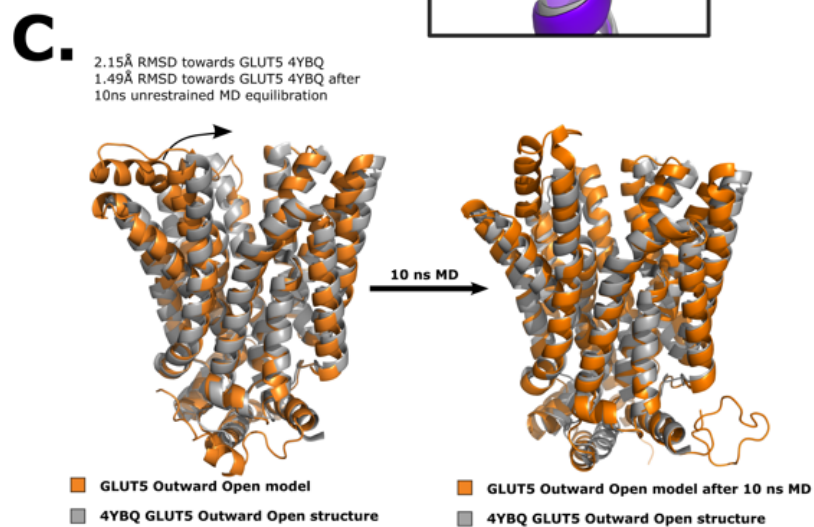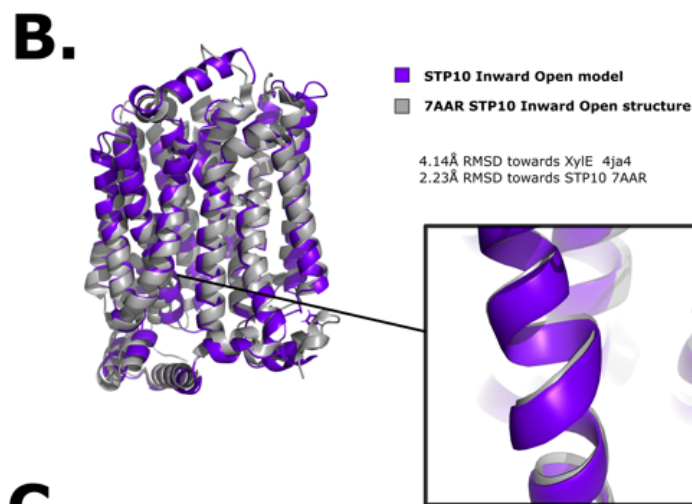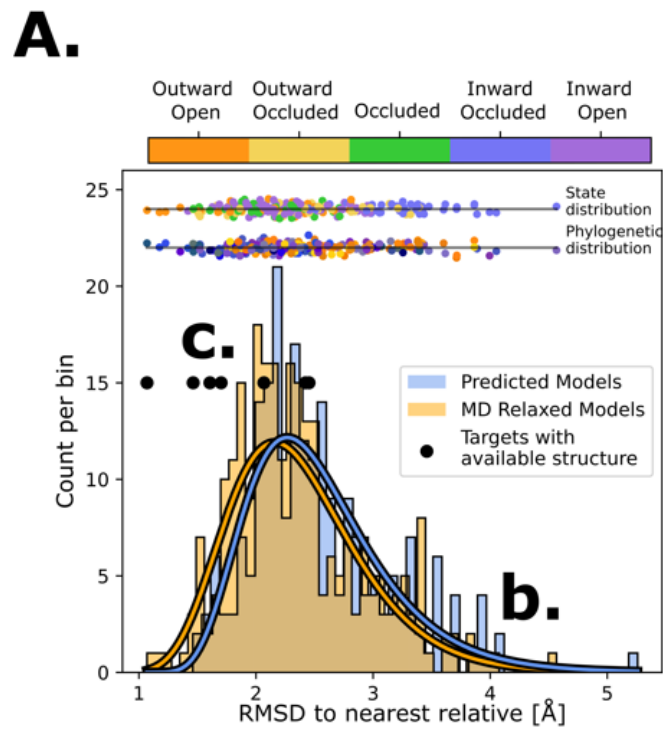
20

27

1
2
3
4
5
6
7 **Figure 4. Quality of structure prediction of state-specific models of SPs A.**
8 Histogram of RMSD of the structural models (Rosetta, blue and MD-relaxed, orange)
9 from their closest relative in the same conformational state. The distributions of the
10 MD-relaxed structures colored according to state and phylogeny (see color definition
11 in figure 1C) are shown above the histogram. Additionally, the targets with available
12 experimental structures are indicated with black dots. **B.** Alignment between STP10
13 inward open model and the newly solved 7AAR STP10 structure in the same
14 conformational state. A helical kink not present in any experimentally determined
15 structure so far is shown as an example of a species-specific feature that is captured by
16 our structure-prediction protocol. **C.** (left) Alignment between the GLUT5 outward-
17 open model and an experimental structure of the same isoform in the same state (PDB
18 ID 4YBQ). (right) Alignment between the MD-relaxed GLUT5 outward-open model
19 and PDB 4YBQ.
20
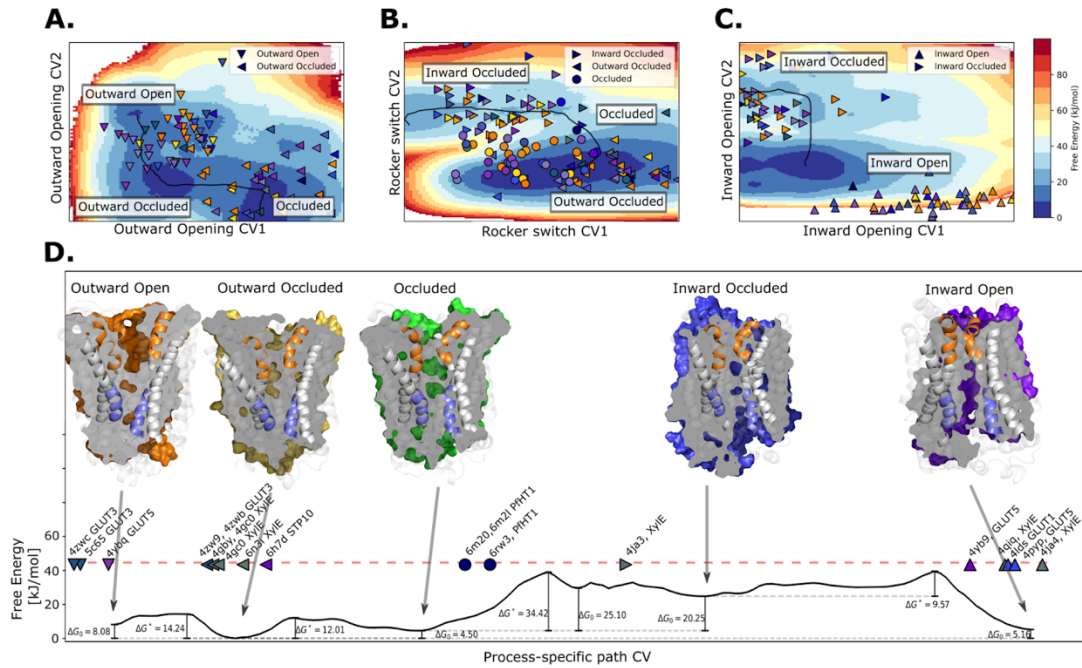
Figure 5. The entire conformational cycle is captured by accelerated weight histogram simulations. A. Free energy landscape of the outward opening process. The most populated free energy basins are labeled according to annotation based on visual inspection and RMSD calculations of snapshots from the 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.

1
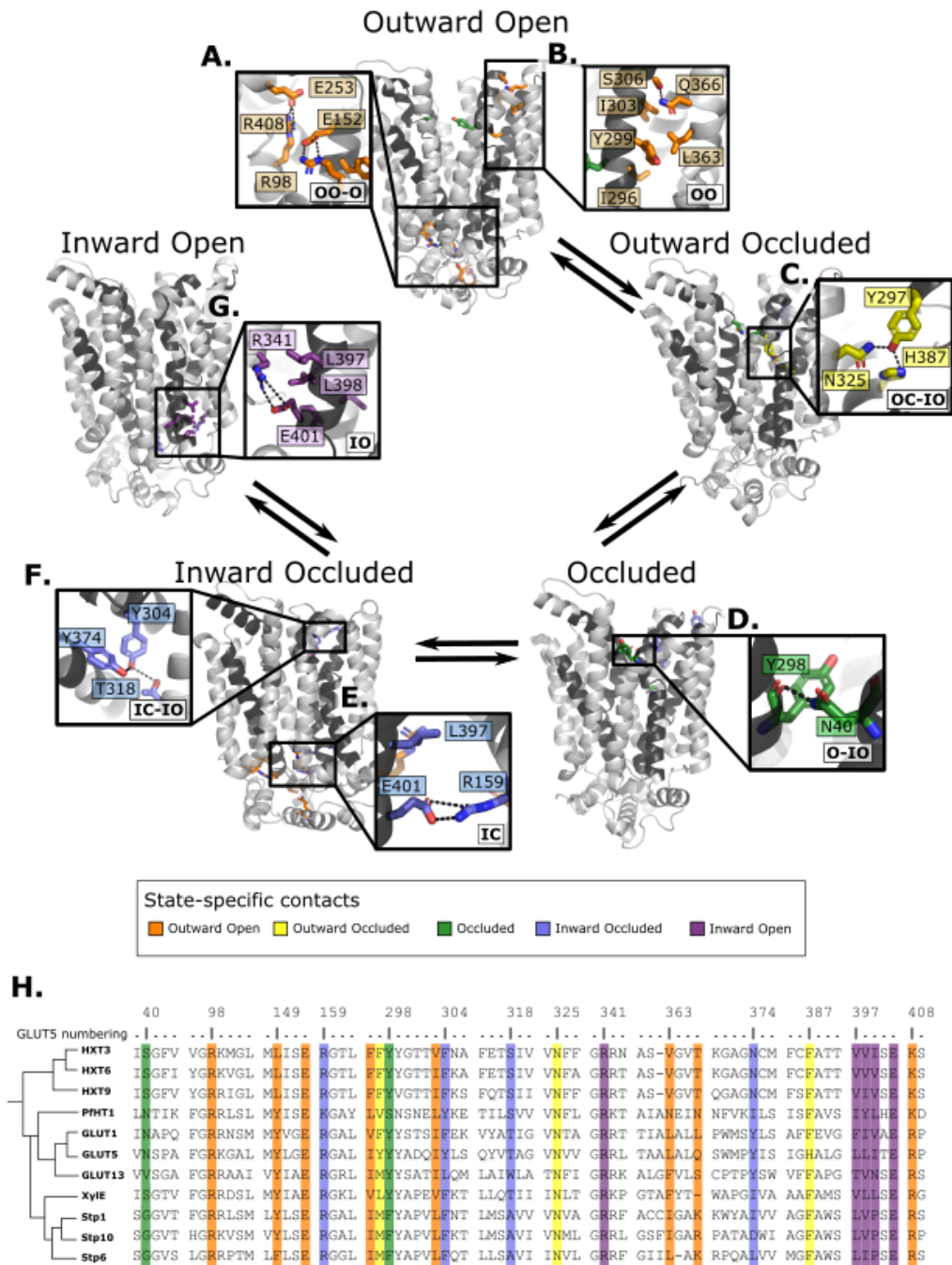
Figure 6. Summary of the structural determinants responsible for the cycling between adjacent conformational states. Contacts are characteristic of the first conformation they appear in, but they can be maintained more or less throughout the cycle depending on the family members. The snapshots are of 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

31

1 responsible for promoting both the bent and broken conformation of TM7b. The
2 position of N325 and H387 necessitates the rotation of TM7b which occludes the
3 extracellular gate. **D.** Backbone contact formed between TM7b and TM1, which is only
4 possible when TM7b is completely broken. **E.** TM10b-TM4 interactions that are the
5 last occluding contacts to break before the intracellular gate opens. **F.** TM10a-TM8
6 contacts responsible for stabilizing the new inter-bundle angle present in the inward
7 facing states. **G.** Salt bridge and hydrophobic nexus responsible for stabilizing the
8 inward-open conformation of TM10b, which fully unblocks the binding site from the
9 intracellular side. **H.** MSA of some representative members at positions shown
10 throughout panels A-G. Since model training was performed on all predicted models
11 using highly coevolving residue pairs as features, the type of interaction present in
12 different subfamilies can be tracked in this MSA.
13
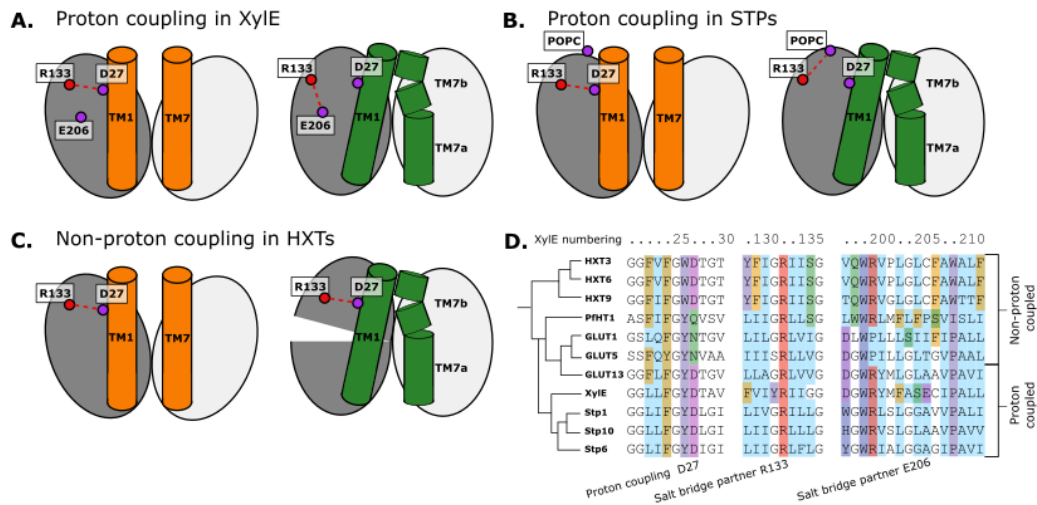
Figure 7. Structural and Evolutionary basis for proton coupling in the sugar transporter family. A. Proposed cartoon model for proton coupling in XylE, showing the highlighted salt bridge network rearrangements between the outward open (orange) and occluded (green) conformational states. Note how these 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.

33

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