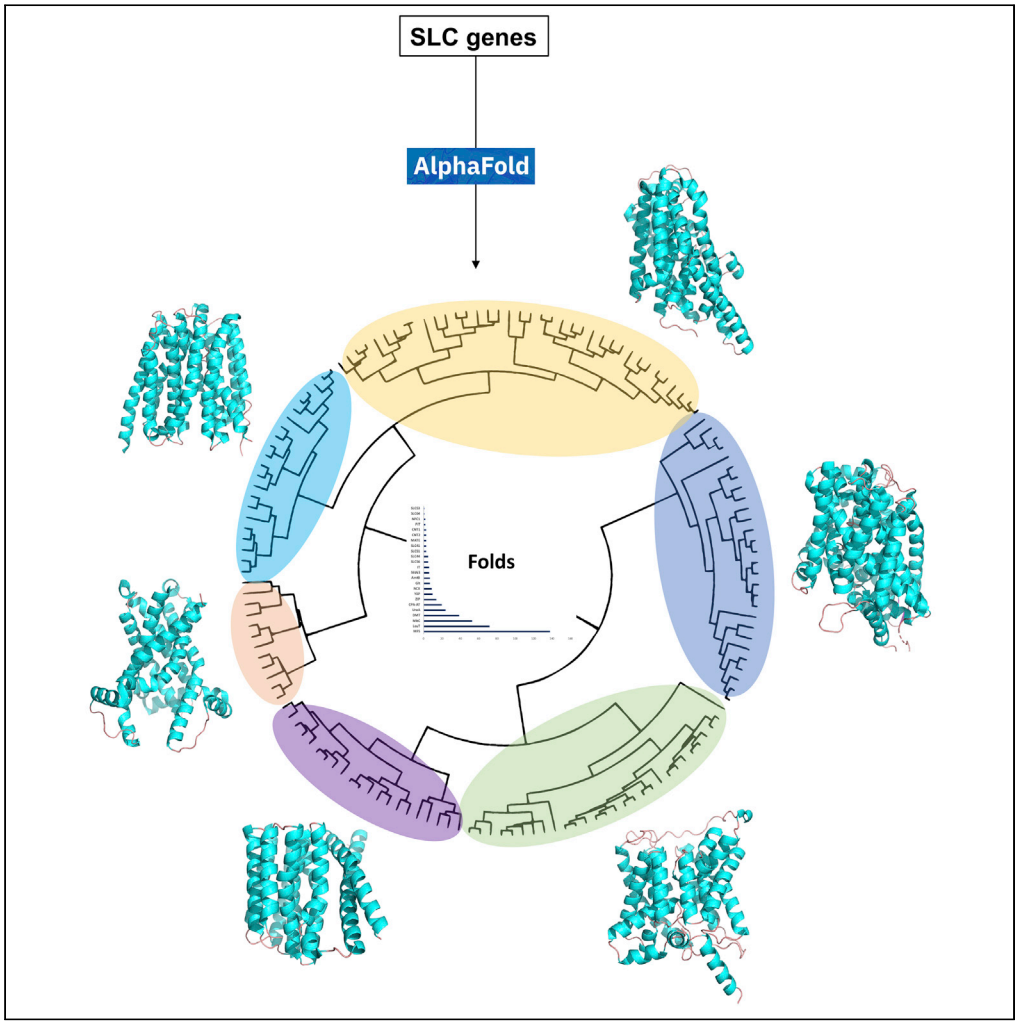
iScience

CellPress
OPEN ACCESS

Article

A structure and evolutionary-based classification of solute carriers



Evandro Ferrada, Giulio Superti-Furga

eferrada@cemm.oeaw.ac.at (E.F.)
gsuperti@cemm.oeaw.ac.at (G.S.-F.)

Highlights

A classification of SLC genes based on structure and evolutionary information

Experimental and AlphaFold models reveal the structural diversity of solute carriers

New associations between orphan genes and members of characterized SLC families

A previously unidentified SLC gene member of the nucleoside-sugar transport family

Ferrada & Superti-Furga, iScience 25, 105096 October 21, 2022
https://doi.org/10.1016/j.isci.2022.105096

iScience

CellPress
OPEN ACCESS

Article

A structure and evolutionary-based classification of solute carriers

Evandro Ferrada[1,3,*] and Giulio Superti-Furga[1,2,*]

SUMMARY

Solute carriers are an operationally defined diverse family of membrane proteins involved in the transport of nutrients, metabolites, xenobiotics, and drugs. Here, we provide an integrative classification of solute carriers by combining evolutionary information with proteome-wide structure models recently made available through the AlphaFold resource. Analyses of orthologous relations among 455 protein-coding genes currently classified as human solute carriers, over the fully sequenced genomes of 2,100 species, suggest no more than approximately 180 independent evolutionary origins. Structural comparative analyses provided further insight revealing a total of 24 structurally distinct transmembrane folds, increasing by approximately 40% the number of previously described SLC structural folds. In addition, a structural comparative analysis identified a new human solute carrier member and revealed details of noncanonical ones. Our analyses uncover new ancestral relations between solute carrier genes, provide insights into the evolution of remote homologs and a platform to test hypotheses of functional deorphanization.

INTRODUCTION

Biological systems require separation from the environment via a semipermeable membrane to contain their genomic and metabolic identity, create energy through chemical gradients, and protect from potentially toxic environmental conditions (Zhang et al., 2019). Cells exchange a large diversity of essential and nonessential compounds with their environment. Because not all chemicals can diffuse through the cell membrane at rates required by cellular physiology, organisms have evolved membrane proteins that facilitate the transport and regulation of nutrient concentrations. Among the large diversity of membrane proteins, those composed of all-alpha transmembrane (TM) domains are the largest class, encompassing almost a third of characterized superfamilies (Lomize et al., 2012). Among the membrane proteins that belong to this all-alpha class, solute carriers (SLCs) are predominantly represented and relatively less studied (Giacomini et al., 2010; Hediger et al., 2013; César-Razquin et al., 2015; Wang et al., 2019). Current classifications of human SLCs, based on sequence similarity and sparse functional annotations, span from 430 to nearly 450 members (Meixner et al., 2020; Tweedie et al., 2021), which represent approximately 9% of human membrane proteins (Dobson et al., 2015b). Based on these criteria, SLCs are currently grouped into 66 families, ranging between 1 and 53 genes per family (Perland and Fredriksson, 2017; Saier et al., 2021). SLCs transport diverse substrates such as amino acids, sugars, complex metabolites including neurotransmitters, vitamins, cofactors, and trace metals such as copper and zinc. The main transport mechanisms associated with SLCs are often called secondary active or facilitative, because of the use of energy already stored in the chemical gradient across membranes (Drew and Boudker, 2016). Moreover, due to their essential transport functions and metabolic relevance, SLCs are associated with several human diseases, including many inborn errors of metabolism and neurodevelopmental disorders (Zhang et al., 2019). SLC transporters are also the target of prominent drugs (Wang et al., 2019). For instance, the SLC family 6 member A4 (i.e., SLC6A4), or serotonin transporter, is the target of currently approved drugs for the treatment of major depressive and anxiety disorders (Rask-Andersen et al., 2013).

Although essential for understanding the function and diversity of SLC families, we know little about their structural diversity. Up until the beginning of 2022, approximately 25 experimentally determined human SLC structures were known, and according to our estimations, accurate models based on homologous genes were feasible for only ~30% of all human SLCs. Importantly, the large sequence divergence between SLC families forbids the transferability of structural and functional information to orphan genes, those with

[1]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, 1090 Vienna, Austria

[2]Center for Physiology and Pharmacology, Medical University of Vienna, Schwarzspanierstrasse 17A, 1090 Vienna, Austria

[3]Lead contact

*Correspondence: eferrada@cemm.oeaw.ac.at (E.F.), gsuperti@cemm.oeaw.ac.at (G.S.-F.)

https://doi.org/10.1016/j.isci.2022.105096

uncharacterized substrate or function, which encompass a third of human SLCs (Meixner et al., 2020). Yet, the development of techniques for the detection of remote homologs has led to the identification of an increasing number of new and uncharacterized SLC genes over the recent years (Gyimesi and Hediger, 2022; Hediger et al., 2013; Meixner et al., 2020; Perland and Fredriksson, 2017; Xie et al., 2021).

Pioneering work on the experimentally determined three-dimensional structure of SLC transporters has opened new perspectives on our understanding of the transport mechanism as well as drug engagement of some of the most biologically and pharmacologically important members of the supergroup (Cotrim et al., 2019; Navratna and Gouaux, 2019; Shi 2013; Yan, 2017). The recent availability of a large number of structure models achieved by the development of artificial intelligence methods heralds yet another dramatic shift in our understanding of structural genomics of many proteins, including transporters (Jumper et al., 2021; Tunyasuvunakool et al., 2021). Here, we use recent data of structure models for the complete proteome of human and several other species to explore the structural diversity of SLC genes. Because two proteins might adopt the same fold without necessarily sharing a common ancestor, we complement structural analyses with evolutionary associations between orthologous genes across a large set of representative fully sequenced genomes (Altenhoff et al., 2021). Our analyses show that human SLC genes have no more than approximately 180 independent evolutionary origins. A structure pairwise analysis using known experimental and AlphaFold structure models revealed relations between remote homologs, suggesting that several orphan SLCs are structurally closely related to established SLC families. Importantly, our analysis expanded the number of described transmembrane folds for SLC proteins from 14 to 24 and related orphan genes to functionally characterized families. We uncovered a new member of SLC family 35, that we call SLC35G7. Finally, our study provides clues about noncanonical SLC families.

RESULTS

Sequence domain diversity and the evolutionary origins of solute carrier families

We collected sequence information for 455 SLC human genes including previous annotations (Meixner et al., 2020); and new entries compiled from the HGNC (Povey et al., 2001) and TCDB (Saier et al., 2021) databases (Table S1). Previous classifications of SLC genes have relied almost exclusively on Pfam annotations (Perland and Fredriksson, 2017). In our updated list of SLC human genes, we found a total of 83 Pfam families. Among these, 64 families map to transmembrane segments, while 19 identified extracellular or intracellular conserved accessory domains (Table S2).

One drawback of relying on Pfam for a classification of SLCs is that a single gene can be associated with multiple Pfam families. While most human SLC genes (70%) map to a single Pfam family, others show complex domain architectures with up to 6 Pfam domains per gene. Examples include SLC genes with multiple internal repeats, such as most members of SLC family 25 or mitochondrial carriers (MitC), in which a single Pfam domain repeats three times to form the transmembrane fold (Ruprecht and Kunji, 2020). Moreover, a significant fraction of SLC genes (12%) shows Pfam domains accessory to the transmembrane segments (Table S3). Although accessory domains might not necessarily be critical for the function of SLC transporters, they reflect evolutionary diversification resulting from indels and/or the gain/loss of domains often following gene duplication. This is particularly the case for some SLC families with variable (e.g., SLC families 7, 12, 21, 22, 39, and 55) and/or multiple (e.g., SLC family 9) number of accessory domains across family members (Table S3).

Most Pfam families show remote sequence similarity with other families and have been grouped into related Pfam "clans". Among transmembrane Pfam families, our annotation identified 11 clans, which include 50 out of the 66 SLC families, ranging from 1 to 21 SLC families per clan (Table S4). Larger clans include the MFS and APC, with 21 and 12 SLC families, respectively. Other five clans reveal homologous relations between SLC families: NhaA (SLC families 9 and 10), MviN_MATE (families 47 and 62), IT (P_HUMAN and families 13 and 53), and DMT (families 35, 39, and 57). New human SLC genes added to the most recent update include a new Pfam family (i.e., PF04193, PQ loop repeat), which belongs to a clan also found in SLC families 50 and 54 (Tables S1 and S4).

SCL families of common ancestry

A second drawback of relying only on the Pfam classification is that remote homology (e.g., clan membership) does not necessarily imply a common evolutionary origin. The availability of full genome sequences for a large set of representative species allows us to track the evolutionary history of protein families in a

Table 1. Summary of SLC genes, orphans, and families sharing common ancestors

SLC Families	SLC Members
orphans, 22	MFSD9, MFSD14A, MFSD14B, SLC22A18
orphan, 32, 38	TMEM104; SLC32A1; SLC38A7; SLC38A8
orphan, 59	MFSD12; SLC59A1; SLC59A2
Orphans	MFSD11; UNC93A
17, 37	SLC17A9; SLC17A5; SLC37A2; SLC37A4; SLC37A3; SLC37A1
17, 63	SLC17A6; SLC17A7; SLC17A2; SLC17A3; SLC17A1; SLC17A4; SLC17A8; SLC63A1; SLC63A3; SLC63A2
8, 24	SLC8B1; SLC24A4; SLC24A2; SLC24A3; SLC24A5
36, 38	SLC36A1; SLC36A4; SLC36A3; SLC36A2; SLC38A5; SLC38A1; SLC38A2; SLC38A4; SLC38A6; SLC38A3

List of SLC genes from different SLC families and/or orphans sharing a common ancestor.

systematic manner (Dessimoz et al., 2005). The presence orabsence of orthologous genes in species at increasing evolutionary distances provides information about their common origin and relative diversification. We used a current classification of orthologous genes to estimate the common ancestry between extant human SLC genes. We found that human SLCs can be traced back to at most 181 common ancestors, meaning that all extant SLC human genes most likely originated approximately 180 independent times. Given that the estimation of this number of events depends on several parameters, as well as on the current available information, it represents only an approximate upper bound to the actual number of common ancestors of SLCs genes. For instance, a larger number of genomic sequences could provide higher confidence in identifying common ancestors, reducing the estimated number of independent origins. Conversely, one can consider the total number of folds as an approximate lower bound estimation of the number of independent origins (see below). The 181 groups of SLCs related by common ancestry (ie., orthologous groups) vary from containing 1 to 15 individual human SLC genes, most of which belong to the same SLC family. The largest groups are SLC family 6 (15 genes), SLC family 2 (14 genes), and SLC family 16 (13 genes) (Table S5). Larger SLC families appear to have several independent origins. SLC family 25, for instance, is composed of 3 large groups of orthologous genes, plus few groups of more recent origin.

Notably, evolutionary related groups also provide insights on the common origin of members of different families (Table 1). For instance, members of SLC family 17 (vesicular glutamate transporter) have most likely emerged as part of two independent events, one associated with SLC family 37 (sugar-phosphate/phosphate exchanger), and another associated with SLC family 63 (sphingosine-phosphate transporter) (Table 1). Moreover, orthologous groups including orphan genes can provide evidence of recent functional diversification, likely conserved structural features, and mechanisms of transport. Our analyses identified three orthologous groups containing at least one orphan gene (Table S5). Members of SLC families 32 and 38 (amino acid transporters) are closely related to transmembrane protein 104 (TMEM104), a protein belonging to the amino acid/auxin permease family and likely to be also an amino acid transporter. Similarly, orphans MF14A, MF14B, and MFSD9 share a recent common ancestor with SLC22A18, a transporter of organic cations based on a proton efflux antiport mechanism (Reece et al., 1998). Finally, MFSD12, a cysteine transporter present in melanosomes and lysosomes (Adelmann et al., 2020), is closely related to SLC59A2, a Na^+-dependent lysophosphatidylcholine symporter in red blood cells and platelets (Vu et al., 2017). The full list of groups of SLC genes sharing common ancestors is listed in Table S5.

Structural diversity of solute carriers

A drawback of a classification based solely on sequence information is that proteins that belong to the same family, or share common ancestry, might still have substantial structural differences. Conversely, even in the absence of homology, protein families can share the same fold. To explore the structural diversity of SLCs, we used protein structure information of the canonical isoforms of 455 SLC human genes made recently available by the AlphaFold resource (Jumper et al., 2021). Inspection of the quality of these structure models based on both residue level information and structure disorder predictions showed that 83% of the models have the recommended quality to make accurate interpretations based on relative backbone

position (STAR Methods, Table S6). To provide a frame of reference for our comparative analysis, we added to the 455 SLCs models, 49 experimentally solved structures of SLCs (Table S7). This list of SLC structures was compiled from the literature and encompasses 22 human and a total of 35 eukaryotic SLC structures. The total 504 models were structurally aligned in an all-against-all manner, and a measure of structural (dis) similarity was determined for all pairs (STAR Methods). A hierarchical clustering analysis based on the group average method revealed several clearly distinct groups of SLC families identified as clusters of conserved overall structure and a much larger group composed of structurally heterogeneous families (STAR Methods, Figure 1). The overall structure of the clusters is robust to the systematic removal of SLC family members (STAR Methods). Moreover, 90% (44/49) in our set of experimentally known structures included in the analysis, co-localized with the clusters populated by structure models of their respective SLC families (Figure 1, leaves highlighted in green background).

Our analysis provides details on families with previously completely unknown structures (e.g., SLC families 11, 20, 34, and 39), only recently solved (ie., SLC family 44), and SLC families of currently known experimental structures that lack annotation (e.g., SLC families 6 and 13) as well as details on the structural relation between family members. Most common SLC folds, including the MFS and LeuT, appear clearly defined with nested clusters of distinct SLC families. For instance, the LeuT fold includes SLC families 5, 6, 7, 11, 12, 32, 36, and 38, each with a low degree of structural dissimilarity between members of the same family. Similarly, the MFS fold, the largest group of SLC genes, includes all members of the MFS Pfam clan, such as SLC families 2, 15–19, 22, and SLCO. Several members of the MFS fold populate separate but closely related clusters. Structural comparisons of members from these clusters show the presence of indels and accessory domains in the transmembrane fold that most likely originated in the common ancestor of paralogous genes, and have expanded into functionally similar, but structurally distinct members.

Similar to the MFS and LeuT folds, SLC family 25 (mitochondrial carrier, MitC) forms a single well-defined cluster with a degree of similarity comparable to the one observed for LeuT and MFS clusters. Two additional examples are SLC families 35 (nucleoside-sugar transporter) and 57 (NiPA-like magnesium transporter), with no previous structure information, which notably merge into a single nested cluster, in agreement with their classification under the same Pfam clan (i.e., DMT) (Figure 1 and Table S2).

We also found another three minor clusters including SLC families with characterized structures, such as SLC families 9 and 10, the Glt fold represented by SLC family 1, and a third cluster clearly composed of SLC families 4, 23, and 26. SLC families 9 and 10 were known to share the same fold (Hu et al., 2011; Hunte et al., 2005). Interestingly, although SLC families 4, 23, and 26 belong to the APC Pfam clan, our analysis reveals clear structural dissimilarities with respect to the main LeuT cluster, which includes most APC clan members.

Our analysis also revealed a larger, structurally heterogeneous group of smaller clusters with no common Pfam clan memberships. Common to all these families is the large average structural heterogeneity, as evidenced by the long branches in the dendrogram (Figure 1). The existence of such smaller clusters may suggest that their members have diverged substantially, beyond detection of similarity, or have evolved independently. They could also be structurally too complex (e.g., form multimeric complexes) to be predicted and classified with enough accuracy. For instance, groups including SLC families 30 and 51 with respect to 47, as well as 8 and 24 with respect to 13, are on average as dissimilar as comparing LeuT fold members from members of the MitC fold (i.e., SLC family 25). Yet some of these small, nested clusters, reveal structural similarities anticipated by associations found through orthology and/or Pfam clan analyses, such as SLC families 8 and 24; 47 and 62; and between SLC family 13 (Na + -sulfate/carboxylate cotransporter) and the OCA gene known to transport tyrosine and control pH in melanocytes (Lee et al., 1995).

Finally, our analysis uncovered SLC genes that might function as part of multimeric complexes or might have been misclassified as SLCs (Figure 1, branches highlighted with an '*'). For instance, SLC families 3 and 7 (or cationic amino acid transporter family) are known to function as a heterodimer. SLC family 7, part of the LeuT fold cluster, is known to harbor the transmembrane transport domain, and accordingly, SLC family 3, in fact an ancillar protein, falls into a well-defined, separate cluster. In addition, SLC 27 (or fatty acid transporter family) reveals a conflicting pattern (Figure 1). Members of this family, whose mechanism of function remains uncharacterized (Ohkuni et al., 2013), seem to have only one to two transmembrane

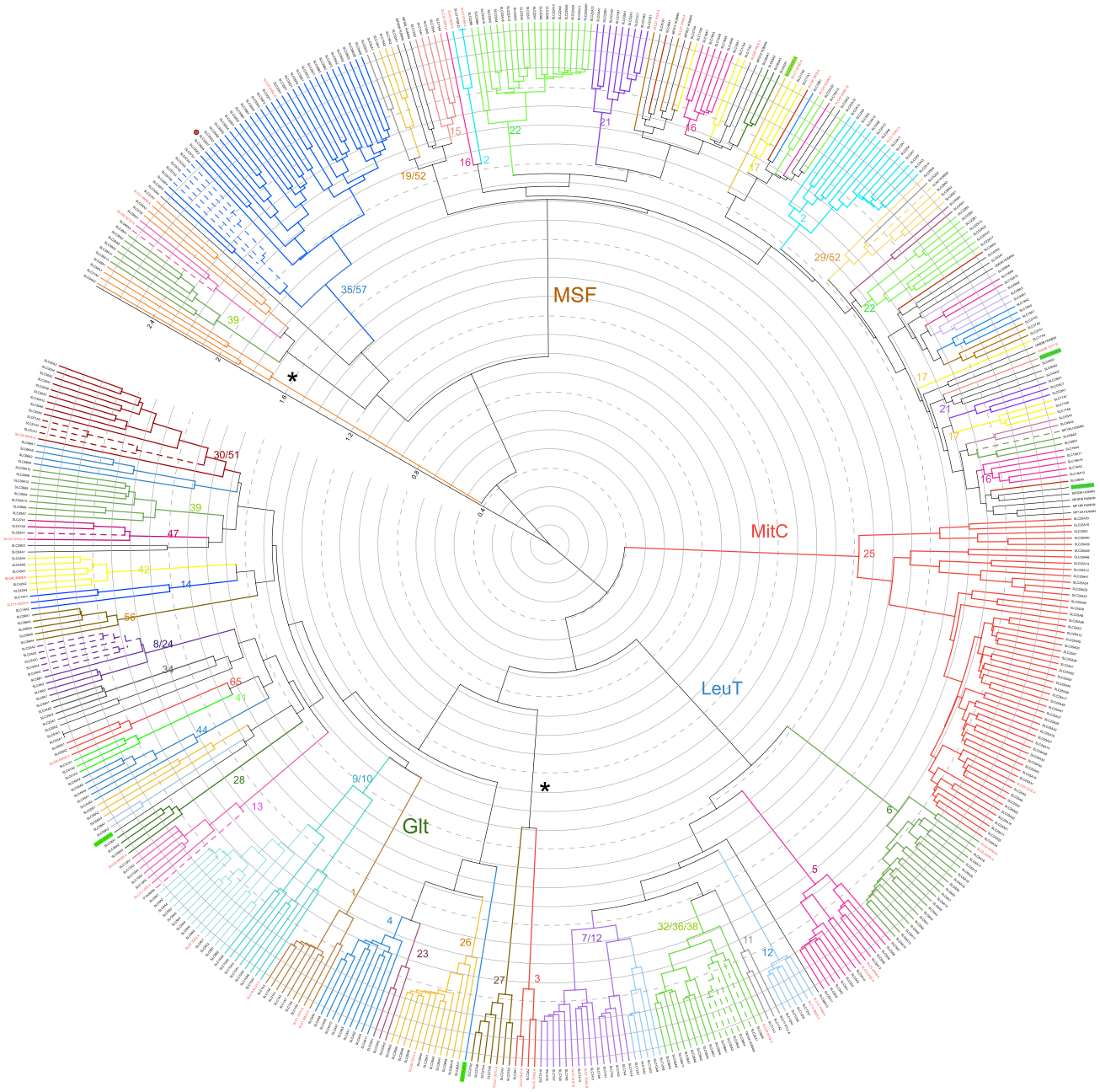


Figure 1. Hierarchical clustering based on structural similarity between structure models of human solute carrier proteins

Structure models for 450 human SLCs were structurally aligned against each other and against models for the entire human proteome. A set of 49 experimentally solved structures were included (leaves labeled in red, and red with a green background for those that did not exactly cluster with their respective structure models). A dissimilarity matrix was built and used for a hierarchical clustering analysis based on Euclidean distance and the group average method. Branch colors were assigned to members of the same SLC families. Dashed branches were used to highlight groups of nested families suggestive of common ancestry. A scan over the entire human proteome identified new putative SLC genes close to the 35 family (highlighted with a red dot). Stars, highlight noncanonical SLCs. For an interactive, online version of this figure see: https://itol.embl.de/tree/9114179215460162903139 1. See also Tables S1 and S7.

domain(s), suggesting that they either function as a multimer, coupled to an additional transport subunit, resembling the heteromeric complex of the SLC families 3 and 7; or simply facilitates transport of lipids by attaching partially to the membrane. Comparable signatures of dissimilarity are revealed in the case of SLC families 48 (heme transporter family) and 54 (mitochondrial pyruvate carriers). Both families have small

domains with significant structural similarity with members of the SLC families 31 and 39 (copper and metal ion transporter families, respectively). Structural alignments between these families, however, only cover half of the transmembrane fold, suggesting that they might form multimers, such as yeast orthologs of members of the SLC family 54 (Tavoulari et al., 2019), or simply do not perform a transport function at all.

A previously unidentified member of SLC family 35

In addition to providing a description of the structural diversity of the SLC superfamily, structure models can be used to find putative new SLC members. In order to search for new SLC members that may have been missed in previous sequence-based searches, we used structural alignments to scan all 455 SLC models described above, against the full set of human structure models (Tunyasuvunakool et al., 2021) (STAR Methods). We were able to identify a new SLC gene with a large structural similarity to members of the nucleoside-sugar transporter, SLC family 35 (Figure 1, red dot). Due to its similarity to the SLC35G members, we suggest renaming this protein SLC35G7. This protein, previously annotated by UniProt as TMEM144, has 12 transmembrane segments, which is similar to the average 11 TM segments among members of the SLC family 35. The gene sequence belongs to Pfam family TMEM144 (PF07857), which was not previously represented among SLC 35 members, yet it was identified as a remote homolog of the DMT Pfam clan (CL0184). The gene has homologous sequences through 230 eukaryotes and close orthologs in 139 species (Dessimoz et al., 2005). This suggests that its correct annotation was missed due to remote homology. Evidence of expression at the protein level for SLC35G7 has been reported (Prentice et al., 2011). Moreover, the close homology to the SLC35G subcluster suggests that SLC35G7 is a sugar transporter, as supported by homology annotations (Fiegler et al., 1999; Jones et al., 2014).

Diversity of SLC transmembrane folds

The degree of structure dissimilarity between most of the clusters discussed in the previous sections suggests that they are amenable to a discrete classification. Two important issues arise here. First, the intrinsic flexibility of proteins imposes a limitation for methods seeking to identify similarity between static models of protein structures. Second, it has been widely recognized that the structural space of proteins is continuous, and therefore a discrete classification of folds might not necessarily fully apply to a group of proteins with a diverse degree of sequence divergence. Nonetheless, we wanted to quantify differences indicative of distinct SLC folds. To reduce noise from predicted structure models, we extracted the models' transmembrane (TM) domains (STAR Methods). Using these TM domains, we repeated our clustering analysis on a subset of 436 models (60 SLC families) with TM domains spanning at least 6 TM segments. The 20 SLC genes discarded from the analysis are: SLC3A1, SLC3A2, SLC27A1, SLC27A2, SLC27A3, SLC27A4, SLC27A5, SLC27A6, SLC31A1, SLC31A2, SLC48A1, SLC51B, SLC54A1, SLC54A2, SLC54A3, SLC55A1, SLC55A2, SLC55A3, SLC58A1, and SLC58A2. As expected, we found a similar, however, much more clearly defined distribution of clusters with reduced intrafamily structural diversity (Figure 2).

An important question to address is how robust the clustering is to the addition or removal of SLC genes. To tackle this question, we studied the resulting number of clusters in resampled sets of SLC models, ensuring that at least 1 member of each family remained in each sample (STAR Methods). Similarly, we repeated the same analysis using samples of conformational variants generated for each model (STAR Methods). In both cases, the overall conclusions described below remained unchanged.

Our analysis, based only on TM domains, identified several of the major clusters of known folds described before (e.g., MFS, MitC, and LeuT) (cf. Figures 1 and Figure 2). In addition, a comparison of the clustering analyses reveals few clusters that show systematic changes in their relative position in the dendrogram. For instance, and most notably, the cluster including SLC families 4, 23, and 26, which was previously found in isolation, now falls closer to members of the LeuT fold. This rearrangement reflects the remote homology between SLC families 4, 23, and 26 and the LeuT fold, as also supported by the Pfam classification, which identifies these families as part of the APC Pfam clan. In addition, the clustering analysis focused on TM domains reveals structural similarities between pairs of genes for which there was no previous evidence based on either the Pfam clan classification or orthology. For instance, the structural similarity between SLC families 14 and 42 provides support to their similar molecular functions (i.e., Urea transporter and Rh ammonium transporter, respectively) and their unique channel-like mechanism of transport among SLCs (Levin et al., 2009). A similar example was found for SLC families 34 and 20, which also have closely related molecular functions (Type II and Type III Na^+-cotransporter family, respectively) and showed TM domains that are structurally similar to SLC family 13 (Figure 2).

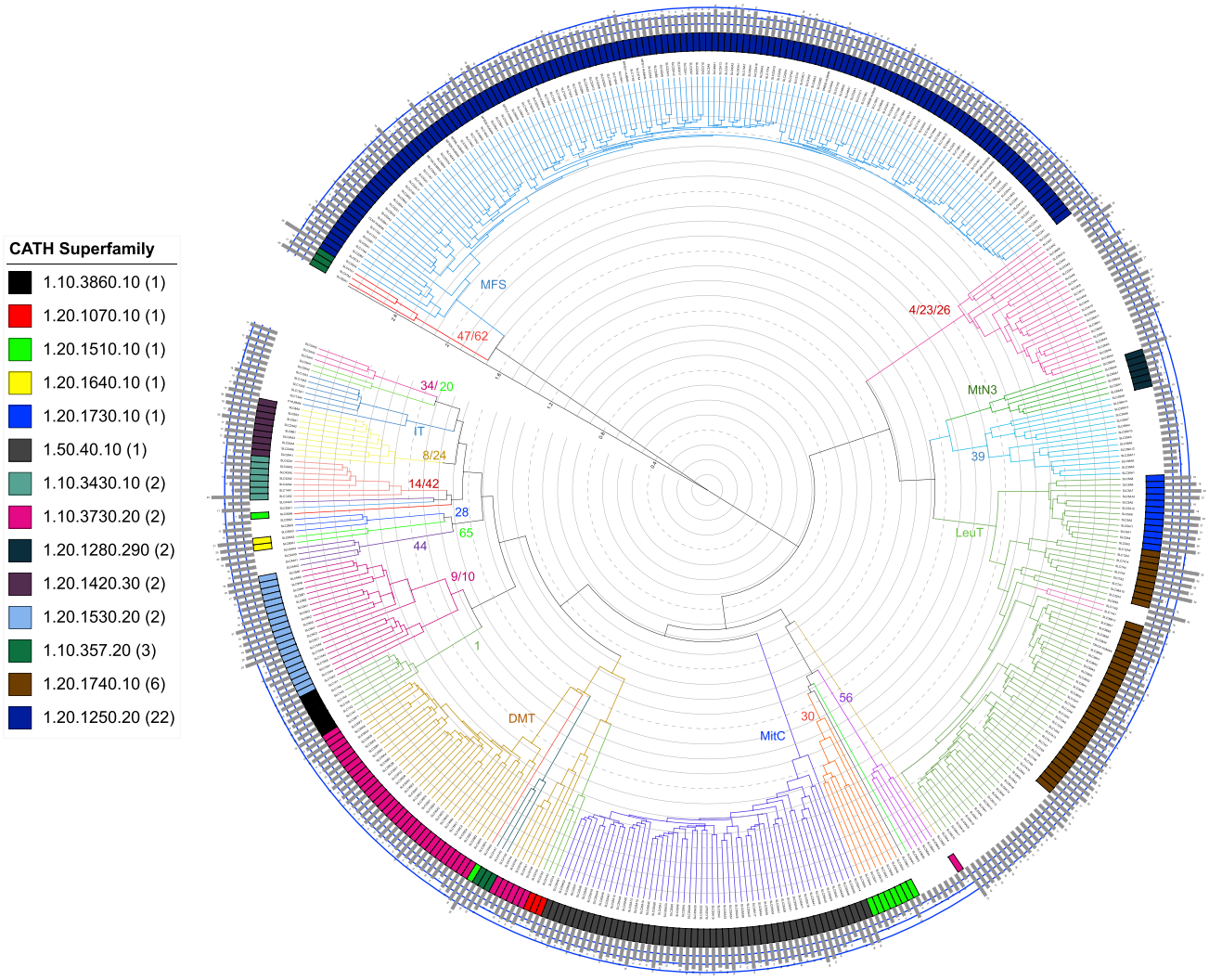


Figure 2. Hierarchical clustering based on structural similarity between transmembrane domains extracted from structure models of human solute carrier proteins

The transmembrane (TM) domains of 436 human SLCs, with at least 6 transmembrane (TM) segments, were structurally aligned against each other. The dendrogram summarizes a hierarchical clustering analysis based on pairwise structural similarity (STAR Methods). Branch colors and branch labels highlight clusters of relatively independent structural similarity. Colored rectangles (and legend) represent 14 CATH superfamilies annotated over the TM domains of 45 of the 60 SLC families in the dendrogram. Bars on the dendrogram tips show the number of TM segments per SLC gene, and the three blue lines indicate a scale highlighting 5, 10, and 15 TM segments. For an interactive, online version of this figure see: https://itol.embl.de/tree/9114169130992416613701 20. See also Tables S8 and S9.

There are several experimentally characterized SLC structures whose folds have been annotated through classifications such as CATH (Sillitoe et al., 2021). To identify previously unknown or described folds, we contrasted our clustering analysis with annotated CATH homologous superfamilies on the TM domains of the subset of 436 human SLC genes. According to CATH, the TM domains of 45 SLC families can be classified into 14 homologous superfamilies (Figure 2). Importantly, some SLC families with known experimental structures have not been assigned a CATH superfamily even though their structures have been known (e.g., SLC family 6), or because they have been experimentally solved only recently (e.g., SLC13 and SLC44). The 14 superfamilies with at least one CATH annotation span three out of the five different architectures within the main alpha class (1.10: orthogonal bundle; 1.20: up-down bundle; 1.50: alpha/alpha barrel). The most common architecture is the up-down bundle (1.20), which is observed in 38 SLC families, including the MFS and LeuT folds (Figure 2).

As expected, different CATH homologous superfamilies can map to SLC families with the same overall structure fold. For instance, SLC family 5 is associated with the CATH superfamily sodium-glucose cotransporter (1.20.1730.10), while most SLC families adopting the LeuT fold map to the amino acid/polyamine transporter I superfamily (1.20.1740.10) (Figure 2). Conversely, the same homologous superfamily can lead to larger structural dissimilarity as observed in SLC families 41 versus 47/62. Remote homology between these families is detected by the CATH classification, but not by Pfam, which only classifies families 47 and 62 as part of the same clan (CL0222). Yet, SLC families identified as part of the same Pfam clan (DMT, CL0184), such as 35, 57, and 39, are also observed in distinctly different clusters, with SLC family 39 closely related to members of the APC clan (Figure 2).

In order to include proteins not yet classified by CATH, we sought to identify new SLC folds as structural dissimilar clusters with relatively distinct topologies (STAR Methods). The degree of structural dissimilarity is simply equivalent to the average distance between pairs of SLC structures in a cluster (Figure 2; STAR Methods). In order to classify a cluster as structurally dissimilar, we use as a reference, the average dissimilarity observed between members of well-characterized structure folds (i.e., LeuT, MFS, and MitC clusters). In addition to structural dissimilarity, we inspected the topology or number and relative distribution of transmembrane segments per fold (Table S9). Although our analyses revealed mostly conserved transmembrane topologies, we observed some folds with substantial variations. In particular, members of the NhaA fold, including SLC families 9 and 10, show a large diversity of topologies, mostly clustered around 9 and 14 transmembrane segments. Other folds showing fewer discrepancies are NCX, PiT, and ZIP folds (Table 2). Such discrepancies might result from relaxed selection, evolutionary diversification, and are expected to be more pronounced in mutationally robust folds. Our analyses revealed 24 relatively independent clusters (Table 2; Table S8). These clusters represent a 40% (10 out of 24) increase in the number of structure folds previously characterized from experimentally solved SLC structures in the CATH protein classification database (Sillitoe et al., 2021) (Figure 2).

Overall, we identified several SLC families of previously unknown structures and/or with putative new SLC folds (11, 28, 30, 39, 41, 44, 51, 53, 56, 4/23/26, 34/20, and 47/62) (Figure 2; Table 2 and S8). Most importantly, our analyses highlight the relative (dis)similarity between SLC families, providing an overview of the structural diversity of human SLC genes.

DISCUSSION

We sought to use currently available structure models for most human genes, as well as sequence and evolutionary information, to provide an updated classification of human SLC genes (Figure 2, Table S5 and S8). Overall, our analyses revealed a complex landscape for the structure and functional diversification of SLC families.

Analyses based on orthology suggest that extant human SLC genes emerged not more than approximately 180 independent times, and most of them have orthologs in all representative species of eukaryotes. Orthology also revealed notable evolutionary relationships between SLC genes, including several cases of members from different SLC families, which might help deorphanization efforts. Like orthologous relations, patterns of sequence similarity of protein families and superfamilies can support evidence of common ancestry and detect homology despite large degrees of sequence divergence. Annotation of remote homologs using the CATH classification revealed that the TM domain of 45 SLC families map to 14 CATH homologous superfamilies, a proxy for structure folds. Large-scale comparisons of structure models for the entire human proteome revealed at least 24 relatively independent clusters with levels of structural dissimilarity large enough as those observed between well-characterized structure folds. Thus, our analyses, based on AlphaFold structure models, revealed a 40% increase in the number of structure folds previously characterized from experimentally solved SLC structures. Furthermore, we used wide proteome structural comparisons to scan for new SLC genes and found a new member of the nucleoside-sugar transporter family (i.e., SLC35G7).

Our analyses are not free of drawbacks. We used predicted models of protein structure, whose quality can suffer from the presence of disordered regions and can be influenced by protein-protein interactions. There can be errors in the modeling of these structures that escape quality control based on a residue-level measure of accuracy. Such errors might particularly affect the clustering of multidomain SLCs through systematic deviations in secondary structure. Similarly, the presence of large, disordered regions can introduce noise in pairwise comparisons. We have tried to minimize the impact of disorder by focusing on TM domains, which do show larger degrees of conservation and lower average error per residue, compared to loops. The formation of multimeric complexes, however, remains a major challenge for

Table 2. Summary of fold classification

Fold	Numer of SLC genes	TM segments	SLC Families/or orphans
MFS	138	12	MFSD6, UN93B, MFS6L, MFSD8, MF13A, MF14B, CLN3, MFS12, MFSD9, MF14A, MFSD1, MFS11, UN93A, SLC2, SLC15, SLC16, SLC17, SLC18, SLC18B1, SLC19, SLCO5, SLCO4C1, SLCO1B3, SLCO3, SLCO1C1, SLCO4, SLCO1, SLCO1B7, SLCO2B1, SLCO1B1, SLCO6, SLCO2, SLC22B3, SLC22B1, SLC22, SLC22B2, SLC22, SLC22B4, SLC22, SLC22B5, SLC22, SLC29, SLC33, SLC37, SLC40, SLC43, SLC45, SLC46, SLC49, SLC52, SLC59, SLC60, SLC61, SLC63
LeuT	72	12	TM104, SLC5, SLC6, SLC7, SLC11, SLC12, SLC32, SLC36, SLC38
MitC	53	6	SLC25
DMT	39	10	SLC35E2, SLC35F4, SLC35F5, SLC35E1, SLC35, SLC35D3, SLC35G2, SLC35F1, SLC35F3, SLC35E2B, SLC35, SLC35B3, SLC35G1, SLC35B2, SLC35C1, SLC35F2, SLC35E4, SLC35F6, SLC35D1, SLC35G6, SLC35C2, SLC35G7, SLC35G4, SLC35G3, SLC35, SLC35G5, SLC35, SLC35D2, SLC35B1, SLC35B4, SLC35D4, SLC35, SLC35E3, SLC57
UraA	24	12	SLC4, SLC23, SLC26
NhaA	20	9	SLC9C1, SLC9C2, SLC9, SLC9B2, SLC9B1, SLC10
ZIP	14	6,8	SLC39
YiiP	10	6	SLC30
NCX	9	9,11	SLC8, SLC8B1, SLC24
Glt	7	8	SLC1
AmtB	7	12	SLC14, SLC42
MtN3	6	7	SLC50, SLC66
IT	6	13	P_HUMAN, SLC13
SLC56	5	3	SLC56
SLC44	5	9	SLC44
SLC51	3	7	SLC51
SLC41	3	10	SLC41
MATE	3	13	SLC47, SLC62
CNT2	3	10	SLC34
CNT1	3	11	SLC28
PiT	2	10,12	SLC20
NPC1	2	13	SLC65
SLC64	1	6	SLC64
SLC53	1	8	SLC53

A total of 436 SLC genes were classified into 24 structure folds according to structure dissimilarity and topology criteria.

protein structure prediction. Several SLC families might be forming homo- or heteromeric complexes, in particular those with few TM segments. Thus, in our analyses of the number of distinct clusters of structure folds, we discarded 20 genes from 8 of such SLC families (STAR Methods). Similarly, the use of a conservative measure of structural similarity (STAR Methods), the visual inspection of many clusters,

and the co-clustering of known experimental structures with their respective SLC families, as well as their conserved topology suggest that error should not affect our main observations.

Yet another important limitation of our methods is to capture unbiased similarities from static models of protein structure. Such models do not capture dynamic properties, which might vary among different SLC families, as evidenced by transport mechanisms of different degrees of dissimilarity between conformers of the same fold (Lezon and Bahar, 2012). In the context of structure fold classification, we have tried to address this issue by using a resampling analysis. A similar approach could be applied to large-scale comparisons of structures, such as the one used to find the new member of SLC family 35. Comparative structural analyses using flexible alignment methods are currently under development and might offer a solution to this problem. Similarly, the formation of homomultimers might lead to structural similarities that our current approach fails to detect. Recent development of AlphaFold2 for the modeling of protein-protein interaction will likely help to address this issue.

Despite these limitations, our analyses provide an updated and integrative classification of SLC genes. Detecting close evolutionary relations between SLC genes can provide clues for understanding the functions of yet uncharacterized families (Table 1). Previously, this strategy has been applied using remote homology methods and Pfam clans, which have identified 4 main groups of common origin. These groups relate family members of the MFS clan (alpha group and MFS fold) and the APC clan (beta group, and LeuT fold), in addition to SLC families 8/24 (delta) and 9/10 (gamma) (Fredriksson et al., 2008). Our analyses confirmed these previous observations based on Pfam clans and revealed new ones. For instance, families 32, 36, and 38, all amino acid transporters with a LeuT fold, are nested and highly similar. Families 8 (Na+/Ca2+ exchanger family) and 24 (Na+/(Ca2+, K+) exchanger family) or families 9 (Na+/H+ exchanger) and 10 (Na+/bile salt cotransporter) are structurally as close as the average intra-family degree of similarity across other SLCs. In addition to these relations, we observed several other pairs of families with high structural similarity. These include families 30 and 51; 47 and 62; 13 and the OCA2 gene (P protein); 7 and 12; 53 and MFS13A; 29 and CLN3; and 35 and 57. These closely nested relations suggest common ancestry despite undetected sequence homology. Moreover, close associations can provide clues of similar mechanisms of transport. This may be true, for example, for families 14 (urea transporter family) and 42 (Rh ammonium transporter), which have been shown experimentally to have a similar channel-like mechanism of transport (Levin et al., 2009). Finally, closely related orthologs, as those found in large and functionally diverse families (e.g., 2, 16, and 25), or between families (e.g., 17 and 37 and 17 and 63), are ideal candidates for the study of the evolution of substrate specificities from common ancestors.

The insights brought by recent developments in protein structure predictions are hard to overstate (Jumper et al., 2021; Callaway 2020). The impact and possible applications of these recent advancements are still expanding in scope, from drug and protein design to the prediction of protein complexes and their dynamics (Del Alamo et al., 2022). In the context of SLC biology, these advances are likely to provide hypotheses for understanding mechanisms of molecular function, deorphanization, and druggability. Future analyses based on flexible structural alignments, multimers, and the detailed characterization of the fold architecture of these clusters, as well as additional functional information from experimental studies, will further refine the SLC classification and serve as a platform for the study of function, mechanisms of transport, substrate specificity, and deorphanization.

Limitations of the study

We identified three important limitations of our work. First, we used predicted models of protein structure, whose quality can suffer from the presence of disordered regions or from the influence of protein-protein interactions. Errors in the modeling of these structures might escape quality control based on a residue-level measure of accuracy. Such errors might particularly affect the clustering of multidomain SLCs through systematic deviations in secondary structure. Similarly, the presence of large, disordered regions can introduce noise in pairwise comparisons. A second, and similar limitation of our study, is to capture unbiased similarities from static models of protein structure. Such models do not capture dynamic properties, which might vary among different SLC families, as evidenced by different transport mechanisms. Finally, SLC families might be forming homo- or heteromeric complexes, particularly those with few TM segments. In the main text, we discuss ways in which we have addressed the impact of these caveats on our conclusions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Sequence data annotation
  - Orthology and domain annotation
  - Structure data and quality control
  - Pairwise, genome-wide structural comparison and definition of transmembrane segments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Transmembrane segment topology

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105096.

ACKNOWLEDGMENTS

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreements No 101034439. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. G.S.F. was supported by the Austrian Academy of Sciences. We thank the G.S.F. laboratory for fruitful discussions. We are grateful to Tabea Wiedmer, Barbara Steuer, Ulrich Goldmann, and Enrico Girardi for critically reading the manuscript.

AUTHOR CONTRIBUTIONS

E.F. and G.S.-F. designed the study plan. E.F. performed analysis, prepared the figures and tables. E.F. and G.S.-F. analyzed and interpreted the data. E.F. and G.S.-F. wrote the manuscript.

DECLARATION OF INTERESTS

G.S.-F. is a cofounder and own shares of Solgate GmbH, an SLC-focused company. E.F. declares no competing interests.

Received: June 10, 2022
Revised: July 22, 2022
Accepted: September 4, 2022
Published: October 21, 2022

REFERENCES

Adelmann, C.H., Traunbauer, A.K., Chen, B., Condon, K.J., Chan, S.H., Kunchok, T., Lewis, C.A., and Sabatini, D.M. (2020). MFSD12 mediates the import of cysteine into melanosomes and lysosomes. Nature *588*, 699–704. https://doi.org/10.1038/s41586-020-2937-x.

Altenhoff, A.M., Train, C.M., Gilbert, K.J., Mediratta, I., Mendes de Farias, T., Moi, D., Nevers, Y., Radoykova, H.S., Rossier, V., Warwick Vesztrocy, A., and Dessimoz, C. (2021). OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. *49*, D373–D379. https://doi.org/10.1093/nar/gkaa1007.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Bolar, K., Bolar, M.K., and LazyData, T.R.U.E. (2019). Package 'STAT.'. R Package Version, pp. 2–5.

Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. *49*, D437–D451. https://doi.org/10.1093/nar/gkaa1038.

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature *588*, 203–204. https://doi.org/10.1038/d41586-020-03348-4.

César-Razquin, A., Snijder, B., Frappier-Brinton, T., Isserlin, R., Gyimesi, G., Bai, X., et al. (2015). A call for systematic research on solute carriers. Cell *162*, 478–487. https://doi.org/10.1016/j.cell.2015.07.022.

Cotrim, C.A., Jarrott, R.J., Martin, J.L., and Drew, D. (2019). A structural overview of the zinc transporters in the cation diffusion facilitator family. Acta Crystallogr. D Struct. Biol. *75*, 357–367. https://doi.org/10.1107/S2059798319003814.

Del Alamo, D., Sala, D., Mchaourab, H.S., and Meiler, J. (2022). Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife *11*, e75751. https://doi.org/10.7554/eLife.75751.

Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., and Gonnet, G.H.

(2005). OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In RECOMB Workshop on Comparative Genomics (Springer), pp. 61–72. https://doi.org/10.1007/11554714_6.

Dobson, L., Reményi, I., and Tusnády, G.E. (2015a). CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res. *43*, W408–W412. https://doi.org/10.1093/nar/gkv451.

Dobson, L., Reményi, I., and Tusnády, G.E. (2015b). The human transmembrane proteome. Biol. Direct *10*, 31. https://doi.org/10.1186/s13062-015-0061-x.

Drew, D., and Boudker, O. (2016). Shared molecular mechanisms of membrane transporters. Annu. Rev. Biochem. *85*, 543–572. https://doi.org/10.1146/annurev-biochem-060815-014520.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

Erdős, G., Pajkos, M., and Dosztányi, Z. (2021). IUPred3: prediction of protein disorder enhanced with unambiguous experimental annotation and visualization of evolutionary conservation. Nucleic Acids Res. *49*, W297–W303. https://doi.org/10.1093/nar/gkab408.

Fiegler, H., Bassias, J., Jankovic, I., and Brückner, R. (1999). Identification of a gene in Staphylococcus xylosus encoding a novel glucose uptake protein. J. Bacteriol. *181*, 4929–4936. https://doi.org/10.1128/JB.181.16.4929-4936.1999.

Fredriksson, R., Nordström, K.J.V., Stephansson, O., Hägglund, M.G.A., and Schiöth, H.B. (2008). The solute carrier (SLC) complement of the human genome: phylogenetic classification reveals four major families. FEBS Lett. *582*, 3811–3816. https://doi.org/10.1016/j.febslet.2008.10.016.

Giacomini, K.M., International Transporter Consortium, Huang, S.M., Benet, L.Z., Chu, X., Evers, R., Hillgren, K.M., Dahlin, A., Evers, R., Fischer, V., Hillgren, K.M., et al. (2010). Membrane transporters in drug development. Nat. Rev. Drug Discov. *9*, 215–236. https://doi.org/10.1038/nrd3028.

Gyimesi, G., and Hediger, M.A. (2022). Systematic in silico discovery of novel solute carrier-like proteins from proteomes. PLoS One *17*, e0271062. https://doi.org/10.1371/journal.pone.0271062.

Hediger, M.A., Clémençon, B., Burrier, R.E., and Bruford, E.A. (2013). The ABCs of membrane transporters in health and disease (SLC series): introduction. Mol. Aspects Med. *34*, 95–107. https://doi.org/10.1016/j.mam.2012.12.009.

Hu, N.J., Iwata, S., Cameron, A.D., and Drew, D. (2011). Crystal structure of a bacterial homologue of the bile acid sodium symporter ASBT. Nature *478*, 408–411. https://doi.org/10.1038/nature10450.

Hunte, C., Screpanti, E., Venturi, M., Rimon, A., Padan, E., and Michel, H. (2005). Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. Nature *435*, 1197–1202. https://doi.org/10.1038/nature03692.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., and Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

Lee, S.T., Nicholls, R.D., Jong, M.T., Fukai, K., and Spritz, R.A. (1995). Organization and sequence of the human P gene and identification of a new family of transport proteins. Genomics *26*, 354–363. https://doi.org/10.1016/0888-7543(95)80220-G.

Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259. https://doi.org/10.1093/nar/gkz239.

Levin, E.J., Quick, M., and Zhou, M. (2009). Crystal structure of a bacterial homologue of the kidney urea transporter. Nature *462*, 757–761. https://doi.org/10.1038/nature08558.

Lezon, T.R., and Bahar, I. (2012). Constraints imposed by the membrane selectively guide the alternating access dynamics of the glutamate transporter GltPh. Biophys. J. *102*, 1331–1340. https://doi.org/10.1016/j.bpj.2012.02.028.

Lomize, M.A., Pogozheva, I.D., Joo, H., Mosberg, H.I., and Lomize, A.L. (2012). OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. *40*, D370–D376. https://doi.org/10.1093/nar/gkr703.

Mariani, V., Biasini, M., Barbato, A., and Schwede, T. (2013). lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics *29*, 2722–2728. https://doi.org/10.1093/bioinformatics/btt473.

Meixner, E., Goldmann, U., Sedlyarov, V., Scorzoni, S., Rebsamen, M., Girardi, E., and Superti-Furga, G. (2020). A substrate-based ontology for human solute carriers. Mol. Syst. Biol. *16*, e9652. https://doi.org/10.15252/msb.20209652.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., and Bateman, A. (2021). Pfam: the protein families database in 2021. Nucleic Acids Res. *49*, D412–D419. https://doi.org/10.1093/nar/gkaa913.

Navratna, V., and Gouaux, E. (2019). Insights into the mechanism and pharmacology of neurotransmitter sodium symporters. Curr. Opin. Struct. Biol. *54*, 161–170. https://doi.org/10.1016/j.sbi.2019.03.011.

Ohkuni, A., Ohno, Y., and Kihara, A. (2013). Identification of acyl-CoA synthetases involved in the mammalian sphingosine 1-phosphate metabolic pathway. Biochem. Biophys. Res. Commun. *442*, 195–201. https://doi.org/10.1016/j.bbrc.2013.11.036.

Paradis, E., and Schliep, K. (2019). Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528. https://doi.org/10.1093/bioinformatics/bty633.

Perland, E., and Fredriksson, R. (2017). Classification systems of secondary active transporters. Trends Pharmacol. Sci. *38*, 305–315. https://doi.org/10.1016/j.tips.2016.11.008.

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). Hum. Genet. *109*, 678–680. https://doi.org/10.1007/s00439-001-0615-0.

Prentice, L.M., d'Anglemont de Tassigny, X., McKinney, S., Ruiz de Algara, T., Yap, D., Turashvili, G., et al. (2011). The testosterone-dependent and independent transcriptional networks in the hypothalamus of Gpr54 and Kiss1 knockout male mice are not fully equivalent. BMC Genom. *12*, 209–217. https://doi.org/10.1186/1471-2164-12-209.

Rask-Andersen, M., Masuram, S., Fredriksson, R., and Schiöth, H.B. (2013). Solute carriers as drug targets: current use, clinical trials and prospective. Mol. Aspects Med. *34*, 702–710. https://doi.org/10.1016/j.mam.2012.07.015.

Reece, M., Prawitt, D., Landers, J., Kast, C., Gros, P., Housman, D., et al. (1998). Functional characterization of ORCTL2-an organic cation transporter expressed in the renal proximal tubules. FEBS Lett. *433*, 245–250. https://doi.org/10.1016/S0014-5793(98)00907-7.

Ruprecht, J.J., and Kunji, E.R.S. (2020). The SLC25 mitochondrial carrier family: structure and mechanism. Trends Biochem. Sci. *45*, 244–258. https://doi.org/10.1016/j.tibs.2019.11.001.

Saier, M.H., Reddy, V.S., Moreno-Hagelsieb, G., Hendargo, K.J., Zhang, Y., Iddamsetty, V., Lam, K.J.K., Tian, N., Russum, S., Wang, J., and Medrano-Soto, A. (2021). The transporter classification database (TCDB): 2021 update. Nucleic Acids Res. *49*, D461–D467. https://doi.org/10.1093/nar/gkaa1004.

Shi, Y. (2013). Common folds and transport mechanisms of secondary active transporters. Annu. Rev. Biophys. *42*, 51–72. https://doi.org/10.1146/annurev-biophys-083012-130429.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., and Orengo, C.A. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273. https://doi.org/10.1093/nar/gkaa1079.

Tavoulari, S., Thangaratnarajah, C., Mavridou, V., Harbour, M.E., Martinou, J.C., and Kunji, E.R. (2019). The yeast mitochondrial pyruvate carrier is a hetero-dimer in its functional state. EMBO J. *38*, e100785. https://doi.org/10.15252/embj.2018100785.

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., and Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1.

Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, R.L., Yates, B., and Bruford, E.A. (2021). Genenames. org: the HGNC and VGNC resources in 2021. Nucleic Acids Res. *49*, D939–D946. https://doi.org/10.1093/nar/gkaa980.

Vu, T.M., Ishizu, A.N., Foo, J.C., Toh, X.R., Zhang, F., Whee, D.M., et al. (2017). Mfsd2b is essential for the sphingosine-1-phosphate export in erythrocytes and platelets. Nature *550*, 524–528. https://doi.org/10.1038/nature24053.

Wang, W.W., Gallo, L., Jadhav, A., Hawkins, R., and Parker, C.G. (2019). The druggability of solute carriers. J. Med. Chem. *63*, 3834–3867. https://doi.org/10.1021/acs.jmedchem.9b01237.

Wiederstein, M., and Sippl, M.J. (2020). TopMatch-web: pairwise matching of large assemblies of protein and nucleic acid chains in 3D. Nucleic Acids Res. *48*, W31–W35. https://doi.org/10.1093/nar/gkaa366.

Xie, T., Chi, X., Huang, B., Ye, F., Zhou, Q., and Huang, J. (2021). Rational exploration of fold atlas for human solute carrier proteins. Structure *30*, 1321–1330.e5. https://doi.org/10.1016/j.str.2022.05.015.

Yan, N. (2017). A glimpse of membrane transport through structures—advances in the structural biology of the GLUT glucose transporters. J. Mol. Biol. *429*, 2710–2725. https://doi.org/10.1016/j.jmb.2017.07.009.

Zhang, Y., Zhang, Y., Sun, K., Meng, Z., and Chen, L. (2019). The SLC transporter in nutrient and metabolic sensing, regulation, and drug development. J. Mol. Cell Biol. *11*, 1–13. https://doi.org/10.1093/jmcb/mjy052.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Raw and analyzed data	This paper	Mendeley Data: https://doi.org/10.17632/8bfdy4227s.1
iToL interactive Figure 1	This paper	https://itol.embl.de/tree/9114179215460162903139 1
iToL interactive Figure 2	This paper	https://itol.embl.de/tree/9114169130992416613701 20
Software and algorithms		
HGNC	Povey et al. (2001)	https://www.genenames.org/
TCDB	Saier et al. (2021)	https://www.tcdb.org/
OPM	Lomize et al. (2012)	https://opm.phar.umich.edu/
OMA	Dessimoz et al. (2005)	https://omabrowser.org/
HMMER	Eddy (2011)	http://hmmer.org/
Pfam	Mistry et al. (2021)	http://pfam.xfam.org/
CATH	Sillitoe et al. (2021)	https://www.cathdb.info/
BLAST	Altschul et al. (1990)	https://blast.ncbi.nlm.nih.gov/Blast.cgi
PDB	Burley et al. (2021)	https://www.rcsb.org/
Alpha-Fold	Jumper et al. (2021)	https://alphafold.ebi.ac.uk/
IUpred2	Erdős et al. (2021)	https://iupred2a.elte.hu/
TopMatch	Wiederstein and Sippl (2020)	https://topmatch.services.came.sbg.ac.at/
STAT R package	Bolar et al. (2019)	https://cran.r-project.org/web/packages/STAT
iToL	Letunic and Bork (2019)	https://itol.embl.de/
ape R package	Paridis and Schliep (2019)	https://cran.r-project.org/web/packages/ape
CCTOP	Dobson et al. (2015a)	http://cctop.ttk.hu/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Evandro Ferrada (eferrada@cemm.oeaw.ac.at).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Databases are listed in the Key resources table. Structural similarity data have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the Key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Sequence data annotation

Sequence information for SLC genes was obtained from previous annotations. The list published by Meixner et al. (2020), including 446 SLCs was cross-referenced and complemented with new entries in the HGNC (Povey et al., 2001) and TCDB (Saier et al., 2021) databases, accessed last by September 2021. We excluded 7 reported pseudogenes. The final dataset of SLC human genes included 456 protein-coding genes,

spanning 66 SLC families ranging from 1 to 53 members (Table S1). In addition, we used the OPM classification of 140 alpha-helix polytopic transmembrane proteins as a reference to define a universal set of transmembrane proteins of known structures (Lomize et al., 2012).

Orthology and domain annotation

Groups of orthologous genes were identified using the Orthologous Matrix database (OMA) (Dessimoz et al., 2005). OMA (version Apr, 2021) estimates both orthologous groups and hierarchical orthologs using a set of 2,138 representative and fully-sequenced genomes spanning all kingdoms of life. We used hidden markov models (HMM) and HMMER (Eddy, 2011; v 3.3.2) to annotate protein families and homologous superfamilies as precomputed by the Pfam (Mistry et al., 2021; version 35.0) and CATH (Sillitoe et al., 2021; version 4.3) classifications, respectively. In both cases we used an e-value of 0.001 and kept annotations of at least 80% of sequence coverage over the HMM model.

Structure data and quality control

We used blastp (Altschul et al., 1990) to compare the sequences of 456 human SLC protein-coding genes to sequences of experimentally solved structures deposited in the PDB (Burley et al., 2021) (last accessed by January 2022). We identified 49 experimentally solved structures representative of 35 human SLCs and 15 homologous structures encompassing 38 SLC families and one orphan SLC. Additionally, model structures for all 456 human SLC proteins and the full proteome of 20 species, were obtained from the Alpha-Fold database (Jumper et al., 2021). We studied the quality of Alpha-Fold models using two approaches. First, models with a residue-level pLDDT (local difference test) quality score larger than 70 have accurate protein backbones (Mariani et al., 2013). Thus, for each model we estimated an Alpha-Fold quality score by calculating the fraction of residues with a pLDDT larger than 70 (QScore; Table S6). We only used Alpha-Fold models with a QScore larger than 0.5. In addition, we studied the fraction of disordered segments using IUpred2 (Erdős et al., 2021). This method reports a per residue score that ranges from 0 to 1.0. A protein residue with a score larger than 0.5 is considered disordered at a false positive rate of 5% (Erdős et al., 2021). We calculated an overall per-sequence score of disorder by simply counting the fraction of residues with a score larger than 0.5.

Pairwise, genome-wide structural comparison and definition of transmembrane segments

Protein structure models were aligned in an all-against-all manner using the software TopMatch (Wiederstein and Sippl, 2020). To avoid biases in the comparison of proteins with different lengths, we defined structure similarity (S) as: $S = L_o/\min\{L_q, L_t\}$ Where L_o is the length of the overlapped segments after optimal three-dimensional superposition; and L_q and L_t, the residue lengths of the query and target structures, respectively. Equivalent carbons between superimposed structures were defined at a distance lower than 3.5 Å. S(q,t) ranges between 0 and 1.0. For each structure model, we extracted transmembrane segments using the Positioning of Protein in Membranes (PPM) method (Lomize et al., 2012). The method uses protein structure information and a precomputed energy potential to estimate the most likely length, tilting angle and position of transmembrane segments across an asymmetric model of the membrane. We used the coordinates reported by this method to extract the "transmembrane fold" of each structure model and recompute pairwise structure similarity as described above.

QUANTIFICATION AND STATISTICAL ANALYSIS

In order to analyze the structural diversity of SLC families, we used hierarchical clustering and carried out a perturbation analysis. We constructed a squared matrix (M) of size N, with N the number of structure models under comparison. Entries in the matrix M, are symmetric and record the structural dissimilarity between query and target (i.e., M(i,j) = M(j,i) = 1-S(q,t)). We use M to carry out a hierarchical clustering analysis, using Euclidean distance and the group average clustering method, as implemented in the R function *hclust*, part of the STAT R-package (Bolar et al., 2019). The resulting dendrogram was represented using iToL (Letunic and Bork, 2019). In order to test the invariance of clusters in our analysis, we used a resampling strategy. We obtained a random sample of SLC genes, ensuring that at least 1 member of each SLC family was present in the sample and repeated the clustering analysis. We repeated the resampling 100 times. To analyze differences in the number and composition of clusters we measured pairwise distance between the resulting dendrograms, using the function *dist.topo* and its method 'score' from the R-package ape (Paridis and Schliep, 2019).

Transmembrane segment topology

To study the number of transmembrane segments, as well as their topology (arrangement of N- and C-terminals with respect to the intra and extracellular regions of the membrane), we used CCTOP (Dobson et al., 2015a). This method uses only sequence information to estimate a consensus topology based on 10 different reported methods. We contrasted these predictions with results from the PPM method (see above).