Rational Exploration of Fold Atlas for Human Solute Carrier Proteins

Tengyu Xie[1,2,3#], Ximin Chi[1,2#], Bangdong Huang[1,2], Fangfei Ye[1,2], Qiang Zhou[1,2]*, Jing Huang[1,2,3]*

[1]Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China.

[2]Institute of Biology, Westlake Institute for Advanced Study, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China.

[3]Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, China.

[#]These authors contributed equally to this work.

*To whom correspondence should be addressed: huangjing@westlake.edu.cn (JH), zhouqiang@westlake.edu.cn (QZ).

Abstract

Solute carrier superfamily (SLC) is the largest group responsible for transmembrane transport of substances in human cells. It includes more than 400 members that can be divided into 65 families according to their physiological function and sequence similarity. Different families of solute transporters can adopt the same or different folds that determines the working mechanism and reflects the evolutionary relationship between SLC members. Analysis of structural data in literatures shows that there are 13 different folds in the solute carrier superfamily covering 40 families and 343 members. To further study their working mechanism, we systematically explored the SLC superfamily to find more folds. Our results indicate that there are at least three new folds in the solute transporter superfamily among which the SLC44 family is experimentally verified to adopt a new fold. Our work has laid a foundation and provided important insights for the systematic and comprehensive study on the structure and function of solute carriers.

Main text

Membrane proteins occupy about 30% of the human proteome, playing important roles in signaling transduction, energy generation and utilization, and mass transporting. Solute carrier superfamily (SLC) is the largest group that catalyze the transmembrane transporting of substances in human cells[1,2]. There are more than 400 SLC members that can be divided into 65 families according to their physiological function and sequence homology. The SLC transporters typically contain several transmembrane (TM) helices that can form specific folds to mediate the cross membrane transporting of the substrates. The folds of SLC members usually define their working mechanisms and suggest the evolutionary relationship between them even when the sequence similarity is low. It is well known that distant-homology proteins can share the same fold, with human SLC superfamily as a typical case with high sequence diversity but relatively low structural diversity. Exploring new folds in a rational way can thus accelerate the establishment of the structure landscape, and ultimately facilitate our understanding on the functions of the SLC superfamily. In this work, we have systematically studied the folding of SLC transporters, computationally screened new folds for this superfamily, and established a full structural atlas for the human SLC superfamily.

Protocol to identify new folds

Identification of new protein folds in a genome has been attempted in the early 21st century, where the probability for a sequence adopting a new fold is estimated by statistical models established from sequence-based classification[3-6]. Some secondary structure prediction results are used[7], however, these methods rely mainly on sequence similarity.

The recent surge in the accuracy of protein structure prediction, benefited from the efficient utilization of coevolution information and deep neural networks, provides an opportunity to rethink the task of identifying new protein folds from their sequences. Well-designed neural networks can realize the mapping of multiple sequence alignments to structures by learning from structure database. This is due to the inherent ability of the residue neural network and the attention-based network to infer pairwise relationship. The idea was originally implemented in RaptorX-contact[8] and pushed to a climax in AlphaFold[9,10] that made a great progress recently. RoseTTAFold[11] was developed using a similar end-to-end network architecture and achieved structure prediction accuracy close to that of AlphaFold. The high accuracy of AlphaFold make it possible to offer meaningful predicted structures to biologists, which is applied to human genome[12]. Before the excellent performance of AlphaFold, we had focused on the SLC fold atlas and attempted to provide more knowledge about the function of human SLC superfamily by integrating predicted structures from multiple different methods.

We established an approach to rationally find new folds in the human SLC superfamily (Fig. 1). Starting from all sequences of the SLC superfamily, multiple advanced methods are used to extract accurate structural information at three levels, including 3D structure coordinates, 2D contact maps, and 1D secondary structure contents, respectively. This information is then compared with known protein structures in protein data bank (PDB)[13] using a carefully designed score to measure the fitness of a sequence compared to all known folds (detailed in the Method section). The methods were applied to investigate the structural landscape of the human SLC superfamily.

To ensure that the fitness score can distinguish new folds based on predicted information, analyses were performed using the knowledge of the SLC superfamily. We checked and confirmed that the high-score SLC members have had their own or homologous structures (Extended Data Table 3 and Extended Data Figure 1b). We also validated the calculation protocol by including only SLC

members with LeuT fold in experimentally known structure database and found out that the fitness score correctly predicted that most of SLC members had very low values and were not LeuT fold under such an assumption (Extended Data Figure 1c).

During the process of our research, AlphaFold and RoseTTAFold were released, which can provide better 3D predicted structures for our protocol. Meanwhile, we found a new fold SLC44 by using trRosetta[14] and RaptorX-3DModeling[15] in our protocol. At the next iteration, AlphaFold and RoseTTAFold can be adopted to predict 3D structure coordinates for other new folds of SLC superfamily, which will form a cycle to iteratively complete the structural landscape of SLC superfamily. To better establish the SLC fold atlas, we referred to the prediction results from AlphaFold and RoseTTAFold for a few SLC members, for which the predicted structures generated by previous methods were not good enough.

SLC fold atlas

The literature and database searching indicates that there are 13 folds for the human SLC superfamily or their homologs among the PDB entries before February 2021, which include 40 families and 343 members[16-18], while folds are not determined for 25 families that include 81 members (Extended Data Table 2, Fig. 2b). To obtain a more complete overview of the human SLC family, we analyzed the folds of the human SLC superfamily by combining our prediction results with the experimental structural information. After structure clustering by DaliLitev.5[19], 424 human SLC members are classified into 29 clusters. We survey the common characteristics of each cluster and manually confirm most of clusters as reasonable folds, while some neighboring clusters with the same fold were merged. Among 424 human SLC members, 415 SLC members with more than 5 TMs are classified into 23 folds, while the other remaining 9 SLC members with no more than 5 TMs are classified into 4 classes according to the numbers of TM helices (Fig. 2a and Extended Data Table 2). For each specific fold, the representative experimental structure from PDB database or predicted structure generated by AlphaFold are shown in Suppl. Video1. We note that the members belonging to the same family adopt the same fold, indicating that the folds are consistent with the sequence-based families but offer more general function-related information.

Compared to folds known before February 2021, our calculation identified 10 more folds. Meanwhile, some previously unclassified SLC members are classified into three existing folds

(Fig. 2b). For example, four SLC families (SLC52, SLC59, SLC60, and SLC61) are newly recognized as MFS (Extended Data Table 2). SLC8 and SLC24, phylogenetically belonging to δ family[17], are structurally classified to NCX in the work, which are consistent in the phylogeny and in structure.

Fold distribution follows Zipf's law

We further investigated the fold distribution of the human SLC members. As shown in Fig. 2b, the relationship between the percentages of the folds and corresponding ranking follows the Zipf's law, similar to the power law observed in genome[20] (Fig. 2b). Almost half of SLC proteins (44.81%) belongs to the MFS and LeuT folds. Besides, the released years of the 18 folds are shown in Fig. 2c. New folds were discovered at a rate of about one per year from 2003 to 2012 and at a slowed down rate in recent years, indicating the fold atlas is nearly complete.

Characteristics of SLC folds

Among these human SLC folds, there are 21 folds composed of two or more repeating units that contain 2-7 TM helices in a repeat and are internally correlated with each other via an inverted or rotated pseudo-symmetrical axis, which is a common feature for the SLC folds (Fig. 3). There are 6 folds containing 5 TM helices and 5 folds containing 3 TM helices in their repeating unit, which are the most common number of TM helices in a repeating unit at fold level. Among the 13 folds with inverted repeats, 8 folds are composed of a core and a scaffold domain. The folds with rotated repeats usually have no core and scaffold domains. There are some folds (GltpH, CNT, and PiT) that contain a pair of half-transmembrane hairpin in their core domain. Some SLC members consist of multiple (usually double) duplicates of the fold of the other members in their corresponding SLC family. For example, SLC14A2 contains two duplicates of the AmtB fold linked by a long linker, while AmtB fold is adopted by the other members in SLC14 (Fig. 4a). In addition, some SLC members exert their function in homo-oligomers, the folds of which consist of multiple chains such as SLC31, a trimer, forming a fold with 9 TM helices (Fig. 4b). The combination of different folds in a SLC member is also observed. The AlphaFold predicted structure shows that the fold of SLC30A5 is composed of the YiiP and the YddG folds that are adopted by SLC30 and SLC35, respectively (Fig. 4c).

SLC57 and SLC64 are new folds that have not been experimentally verified. The prediction structures by AlphaFold for members of SLC57 and SLC64 family are similar with those predicted by trRosetta[14] and RaptorX-3DModeling[15], which is convincing for further analysis. According to the ranking results, all members of SLC57 family were not in top positions and were predicted to be structurally most similar with sodium-pumping pyrophosphatase (PDB ID: 4AV6), but their sequential TM helical arrangements of the two proteins are different. SLC57 forms "5+5" inverted repeats with TM 3 inserted in the between of TM 1 and TM 2. As for SLC64 with only one member SLC64A1, which is ranked 11th in ranking table (Extended Data Table 1), there is no similar structures in database found. The TMs 1-3 and TMs 4-6 of SLC64A1 are folded as "3+3" inverted repeats linked by a long TM helix, which is parallel to the membrane plane. Experimental results are in demand for validation for these computational results. SLC44 is a new fold validated by experiments, which is introduced in the next section.

SLC44 is a new fold in SLC protein superfamily

We applied our prediction protocol to 424 human SLC proteins and hypothesized that SLC members with low fitness scores might have new folds. The fitness score distribution overall is approximate to a normal distribution (Extended Data Figure 1). The first version of calculation results indicated that the members of the SLC20 family (SLC20A1 and SLC20A2) probably adopted new folds. The later published PDB structure of a homology of SLC20 from *Thermotoga maritima* adopts a new fold, named as PiT fold[21] (Fig. 2a). The ranking results was updated, which still indicated the members of SLC44 and SLC64 family probably adopted a new fold. We further chose SLC44A1 for structural validation using single-particle cryogenic electron microscopy (cryo-EM). SLC44A1 sequence was cloned and purified. The cryo-EM samples of SLC44A1 were prepared for data acquisition and processing (Fig. 5a-b, Extended Data Figure 2-3). Finally, a cryo-EM map of SLC44A1 was obtained at a resolution of 4.4 Å (Fig. 5c and Extended Data Figure 3).

Molecular dynamics flexible fitting[22] (MDFF) was carried out to obtain the atomic models of SLC44A1 (Fig. 5e) starting from the predicted structures generated by RaptorX-3DModeling and guided using the cryo-EM map as a bias potential (details in the Method section). The cross-correlation coefficient (CCC) between the structure model and the cryo-EM map increases from

0.295 to 0.610 with MDFF, which is accurate enough to reflect the exact fold of SLC44A1 (Fig. 5f and Extended Data Figure 4).

SLC44A1 has 10 TMs arranging as a corolla shape that contains 4 pairs of TM hairpins formed by TM3/4, TM5/6, TM7/8, and TM9/10, respectively. The TM3/4 hairpin has different topology than the other three hairpins that assemble a rotated repeating pattern. The TM1 and TM2, which are in proximity to TM5/6 and TM9/10 hairpins in peripheral region of TM domain, respectively, are linked by a long loop that folds as a domain. To further confirm that human SLC44A1 is a new fold, we aligned the cryo-EM structure with all structures in the PDB database with TM-align[23]. The highest TM-score is only 0.35, and there are no similar structures for SLC44A1, indicating SLC44A1 corresponds to a new fold.

We also compared the cryo-EM map and built model with the AlphaFold predicted structure. The TM helices of AlphaFold2 can be aligned with the cryo-EM density map well except TMs 3-4 (CCC=0.436). With the cryo-EM density map we also managed to further refine the AlphaFold predicted structure, and the TM arrangements of the two built models are almost the same (Fig. 5f). In general, the folding pattern is the same between the cryo-EM structure and the AlphaFold predicted structure (TM-score=0.807). Furthermore, the model of SLC44A1 dimer was built based on the cryo-EM map of the SLC44A1 dimer (Fig. 5d). These results demonstrate that the developed computational approach can identify proteins with new folding pattern and thus broaden our knowledge on the sequence-structure-function relationship of proteins.

Discussion

Proteins are the major component in cells that fulfills the life activities. Classification of the SLC superfamily is usually performed based on the sequence homology, while the sequence-structure-function relationship is central to understand the working mechanism of proteins. We present a full description on the structural atlas of human SLC proteins in this work, which will enhance our understanding in membrane transport. Our results indicate that it is possible to rationally discover new protein folds by combining computational and experimental information, allowing to systematically study the sequence-structure-function relationship for a whole protein superfamily. The similar methods can be applied to other membrane superfamilies and other species.

The SLC members usually work in single chain that folds as single domain in TM regions. There are several exceptional cases. SLC54 and SLC55 transport substrates by forming multiple chains and SLC30A5 has two TM domains. Prediction of single domain or chain of SLC55 and SLC30A5 cannot give complete insights of their functions. The information of the numbers of chains in those complexes is necessary and robust 3D structure prediction methods for multiple domains and multiple chains are in demand.

In addition, structure clustering reveals relationship among all folds of the human SLC superfamily (Fig. 6). The major features for each fold, such as containing the inverted or rotated repeats and whether including the half-transmembrane hairpin or not, are also labeled. It is worthy to note that the folds with similar features cluster closely. The 3TMs repeat unit in MFS also exists in its neighboring ZIP, YddG, and SLC57 folds. All of these four folds are different in the number of TM helices and the assembly manner of the repeating units. The SLC members of the AaC, YiiP, SLC44, and SLC56 folds with rotated repeats are located adjacently. So do the LeuT, NhaA, GltpH, NCX, UraA, and SLC57 folds that are composed of core and scaffold domains. The relationship among the SLC folds revealed based on the structural similarity may facilitate the mechanism study of the human SLC superfamily.

Acknowledgements

We thank the cryo-EM facility and the high-performance computing center of Westlake University for providing supports. This work was supported by the National Key R&D Program of China (2020YFA0509300) from Ministry of Science and Technology of China, the National Natural Science Foundation of China (projects 31971123, 32022037, 21803057), and Zhejiang Provincial Natural Science Foundation of China (LR19B030001).

Author contributions

J.H. and Q.Z. conceived this work. T.X., X.C., B.H., and F.Y. performed the experiments. T.X. performed the modelling and computational work. X.C. and B.H. performed biochemical work. X.C. and F.Y. determined the cryo-EM structure of SLC44A1. All authors analyzed the data. J.H., Q.Z., T.X., and X.C. wrote the manuscript with critical input from all authors.

Competing interests

The authors declare no competing interests.

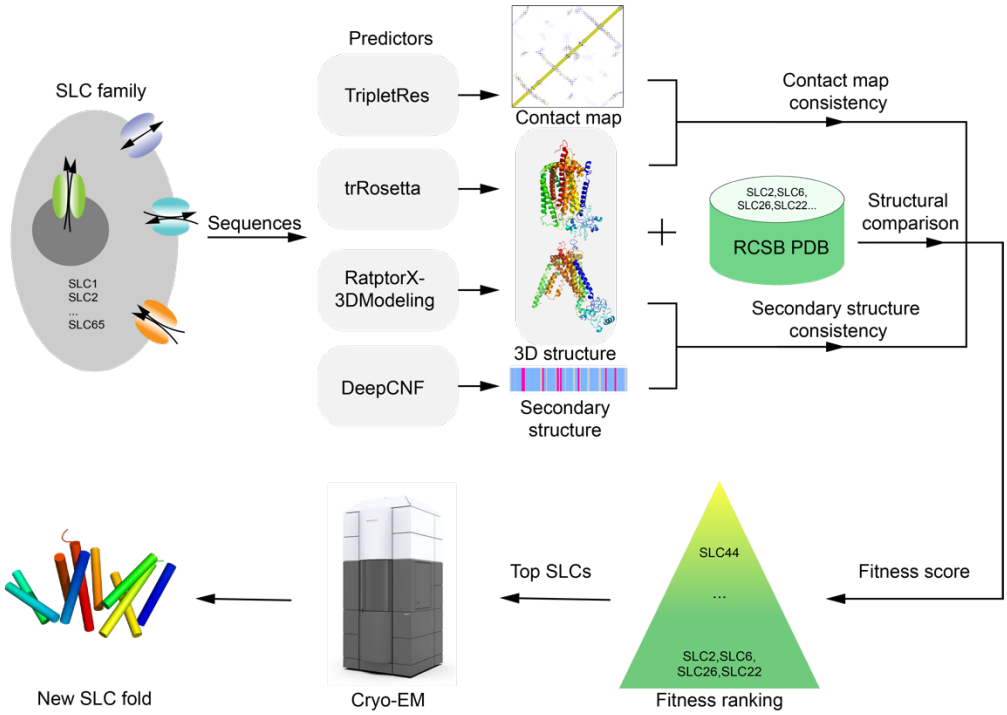


Fig. 1 | The discovery strategy of new fold of human SLC superfamily.

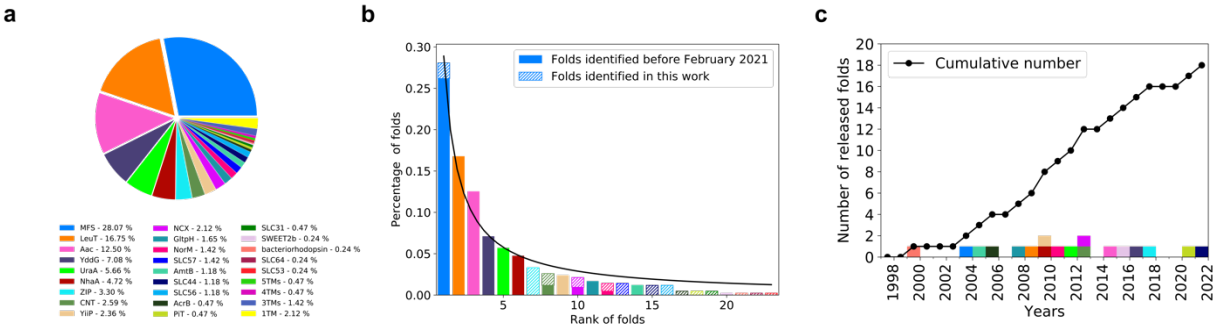


Fig. 2 | The overview of SLC folds. a, The percentages of confirmed folds in SLC superfamily. b, The percentages of folds before February 2021 (solid color block) and after our calculation (shaded color block). The percentages follow the Zipf's law. c, The number of released folds in each year (color blocks) and the cumulative number of released folds along years (black dot line). The released year for each fold is the earliest year of all PDBs belonging to the fold. The color of bars in b or blocks in c corresponds to the color legend of a.

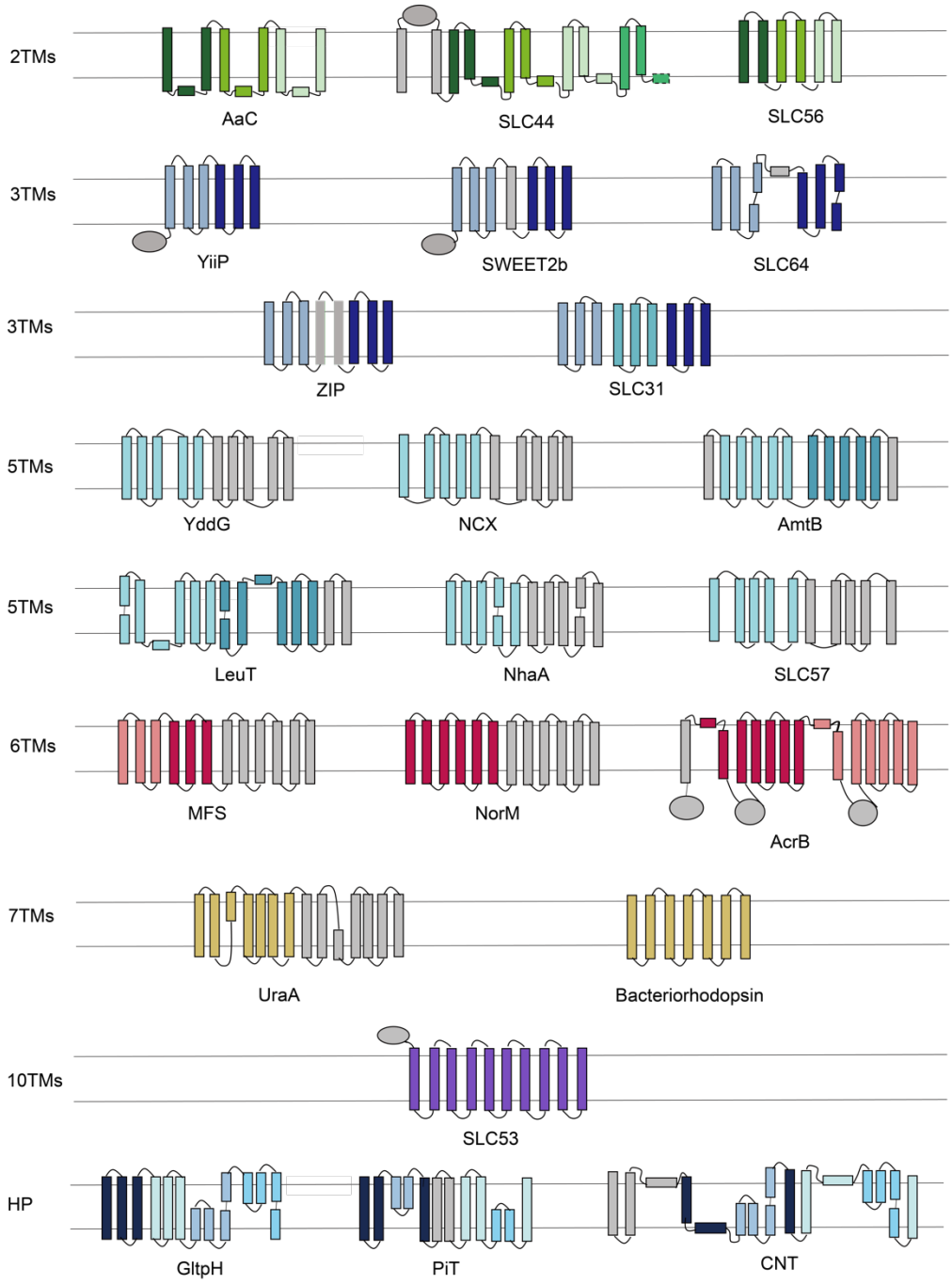


Fig. 3 | Cartoon representation of SLC folds. These folds are listed according to their number of TM helices in a structural repeat unit.

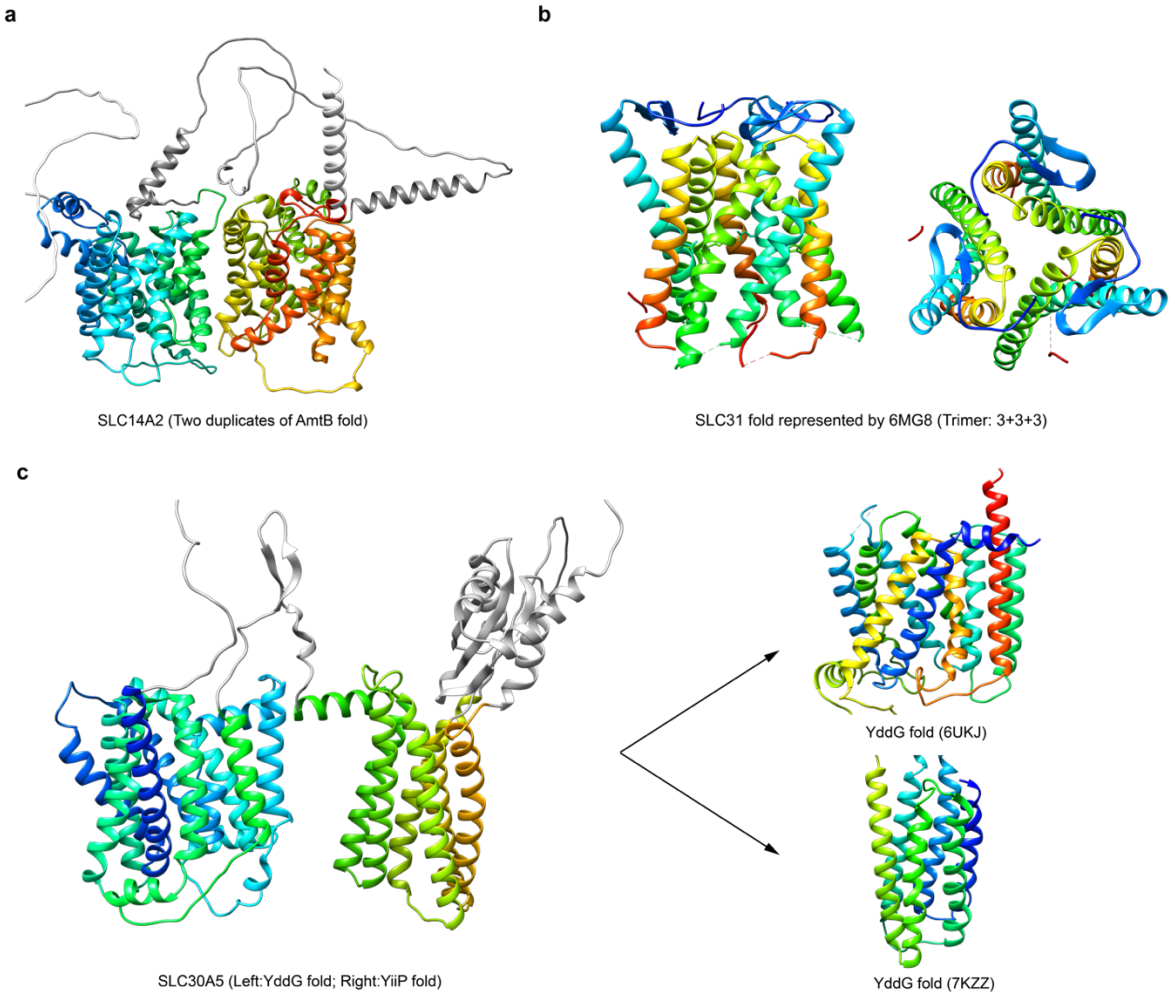


Fig. 4 | Folds of SLC14A2, SLC31, and SLC30A5. a, SLC14A2 adopts the two duplicates of the AmtB fold. b, The fold of SLC31 is composed of three chains, each of which has 3 TM helices. c, SLC30A5 is composed of the YiiP and the YddG folds. The predicted structures of SLC14A2 and SLC30A5 are generated by AlphaFold.

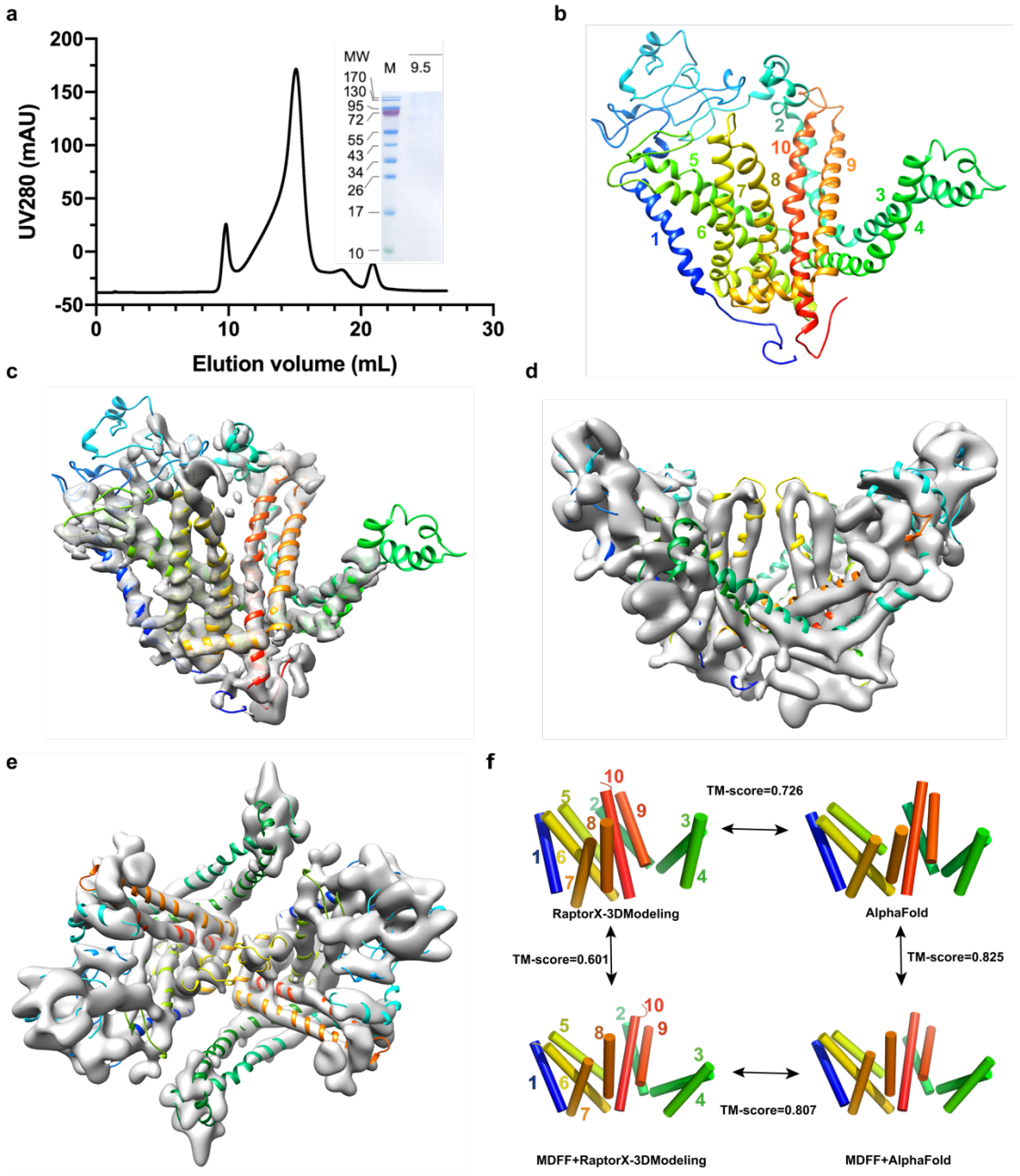


Fig. 5 | SLC44A1 adopts a new fold. a, Size exclusive chromatography (SEC) results and the corresponding SDS-PAGE analysis of the SEC. UV280 is applied to detect proteins. b, The SLC44A1 model. c, The alignments of SLC44A1 model and its density map. d, e, The side view, and top view of the alignment of SLC44A1 dimer and its density map. f, Transmembrane helix arrangements of SLC44A1 model and the predicted structure of RaptorX-3DModeling and AlphaFold.

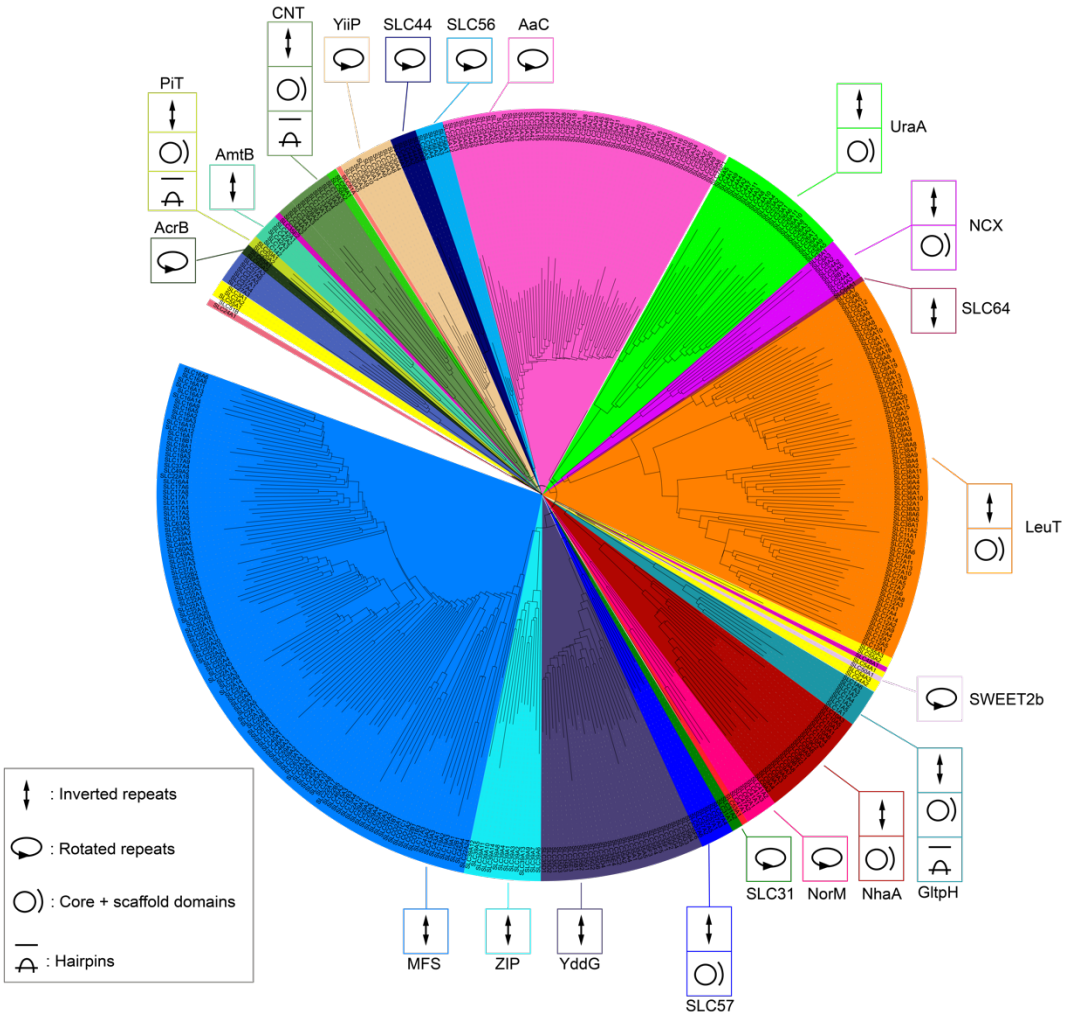


Fig. 6 | Classification result of SLC structures. The SLC members are colored according to the corresponding folds.

Methods

Datasets of SLC sequences and experimental structures

There are 65 SLC gene families with 458 different human transporter genes, including repeated genes with different isoforms (such as SLC14A1_UT-B1 and SLC14A1_UT-B2). As 27 SLC genes have no available amino acid sequences and 6 SLC members are isoforms, these 33 SLC members were excluded in our calculation. In total, 425 SLC genes was included in structure prediction phase. *SLC7A5P1*, which is a pseudogene, is not considered in fold recognition. Thus, 424 SLC genes left. As the purpose is to recognize the overall fold pattern, only TM regions are subjected to structure prediction. Sequences that are 20 residues away from TM-regions are deleted during prediction. The TM regions are recognized by TOPCONS[24].

To identify new folds, structural comparison between SLC members to all known folds is achieved by comparing SLC members to all available structures in the PDB database. In PDB database, many proteins have multiple experimental structures, which are redundant for the structural comparison. Thus, we made use of sequence clustering data (https://cdn.rcsb.org/resources/sequence/clusters/bc-100.out) to get only one structure for a unique sequence. This reduces the 560,635 protein chains in the RCSB PDB database as of February 2021 to 97,033 sequences that are retained after clustering. We further reduce computational cost by filtering out protein sequences with no-TM regions, which results in 4,812 structures to be used in the computational protocol for new fold identification.

Fitness score

To find SLC members with new fold, structural information is predicted at three levels that are 3D structures, contact maps, and secondary structures. 3D coordinate structure is the most informative and directly used to compared with known protein structures. Contact map and secondary structure prediction are less informative but more accurate to compute, so they are used to account for possible errors in 3D structure prediction. The information at each level can be predicted by customized and multiple methods. The three levels of information are integrated to a fitness score S, which reflects the dissimilarity of a target sequence to all experimentally determined structures and is defined as follow:

$$S = \frac{1}{2} w * S_{3D} * (1 + (C_{max}^{all} - C_{3D}) * (1 - \frac{1}{2}(\overline{C_{2D}} + \overline{C_{1D}}))) \quad (1)$$

, where w is a weight factor, S_{3D} is the similarity of the predicted 3D structures to all known structures, C_{3D} is the maximum similarity of M 3D prediction methods for one SLC (Eq. 3), C_{max}^{all} is the maximum similarity value of C_{3D} among all SLC members, $\overline{C_{2D}}$ is the averaged consistency of predicted 2D contact maps to all predicted 3D structures (Eq. 4), $\overline{C_{1D}}$ is the averaged consistency of predicted 1D secondary structures to all predicted 3D structures (Eq. 6). The fitness score S is a number ranged from 0 to 1, with higher value means a protein sequence is more like known structures.

One of advantages of the fitness score is that it can integrate different prediction methods. Once advanced methods are newly released, the fitness score can directly make use of that and keep pace with the development of predicted methods. In this work, we use two structure prediction methods: trRosetta[14] and RaptorX-3DModeling[25]. TripletRes[26] and DeepCNF[27] are adopted to predict contact map and secondary structure, respectively. Predicted 3D structures are compared with all known protein structures in RCSB PDB database through TM-align[23]. TM-align rely on structural alignment instead of sequence alignment and a TM-score value of greater than 0.5 indicates that the two compared structures have similar folds[23]. Due to the inevitable inaccuracy in predicted 3D structures, two strategies are adopted to compensate the errors. One is that more than one 3D protein structure prediction methods are adopted, resulting multiple predicted 3D structures for one SLC. These predicted structures are compared to all known structures, and the corresponding maximum TM-align value S_{3D} reflects the similarity of the predicted 3D structures to all known structures. The similarity S_{3D} is defined as follow:

$$S_{3D} = \max_{\substack{i \in \{1,2,\ldots,M\} \\ n \in \{1,2,\ldots,N\}}} TMscore(P_i, E_n) \quad (2)$$

, where M is the number of 3D structure prediction methods adopted, N is the number of experimental structures in the structure database. Assume M 3D structure prediction methods are adopted, the predicted structure P_i of ith ($i \in \{1,2,\ldots,M\}$) method for one target sequence is compared to in N experimental structures ($E_n, n \in \{1,2,\ldots,N\}$) in the database by TM-align.

In addition, the consistency of these predicted structures is also key information. If these predicted structures are similar with each other, the structures will be more reliable, and the comparison result S_{3D} based on predicted 3D structures is more convincing. Thus, the maximum similarity of all predicted 3D structures comparison (C_{3D}), representing the consistency of M prediction methods, is integrated to S, where the similarity C_{ij} of each pair of predicted structures P_i and P_j ($i, j \in \{1,2, \dots, M\}$) are calculated by TM-score. The definition of C_{3D} is as follow:

$$C_{3D} = \max_{i,j \in \{1,2,\dots,M\}} C_{ij} = \max_{i,j \in \{1,2,\dots,M\}} TMscore(P_i, P_j) \tag{3}$$

As secondary structures and contact maps can be predicted in general with higher precision, these two types of predicted information are also included in the inference of new folds. A good, predicted 3D structure should be consistent with the predicted secondary structure and contact map. Thus, $\overline{C_{2D}}$ reflects the consistency of M 3D predicted structures to predicted contact maps and is integrated to the fitness score. Its definition is as follows:

$$\overline{C_{2D}} = \frac{1}{M} \sum_{i=1}^{M} C_{2D}(P_i) \tag{4}$$

$$C_{2D}(P_i) = \frac{1}{L} \sum_{n=m+12}^{L} \sum_{m=1}^{L-13} I_{\left[d_{mn}^{P_i} \leq 8.0\right]} I_{[Rank(c_{mn}) \leq L]} \tag{5}$$

, where L is the sequence length of a target protein, m and n are residue indices, c_{mn} is the contact probability predicted by an advanced method for the residue pair (m and n), $d_{mn}^{PS_i}$ is the Euclidean distances of the residue pair in the predicted structure P_i, the function $I_{[condition]}$ is equal to 1 when $condition$ is satisfied otherwise 0. For one predicted structure P_i, only top L medium- and long-contacts ($|m - n| \geq 12$) are checked. If the Euclidean distances $d_{mn}^{P_i}$ of their C_β atoms are no more than 8.0 angstrom and the predicted contact probabilities c_{mn} are ranked in top L ($Rank(v_{mn}) \leq L$), the contacts of the residue pairs are consistent. So, a larger $\overline{C_{2D}}$ indicates higher the consistency of contact map. c_{mn} is obtained from TripletRes in this work.

Similarly, the consistency of M predicted structures to predicted secondary structures is integrated to the fitness score. The hit rate of secondary structures of predicted 3D structures to predicted secondary structures is defined as

$$\overline{C_{1D}} = \frac{1}{M}\sum_{i=1}^{M} C_{1D}(P_i) \tag{6}$$

$$C_{1D}(P_i) = \frac{1}{L}\sum_{t\in\{\mathrm{E,H,C}\}}\sum_{m=1}^{L} p_m(t)\, I_{\left[T_m^{P_i}=t\right]} \tag{7}$$

, where E, H, C are types of secondary structures, *t* is one of the three types, *m* is the residue index, $p_m(t)$ is the possibility of the *m*th residue to be the secondary type *t*, $T_m^{P_i}$ is the secondary structure type of *m*th residue of the predicted structure P_i. DSSP[28] is adopted to obtain the secondary structure of 3D structure P_i. Predicted secondary structure is obtained by DeepCNF. The consistency of secondary structures is higher if $\overline{C_{1D}}$ is close to 1.

Finally, the last term w will reweight the fitness score when the 3D structural consistency is good ($C_{3D} \geq 0.5$). Its weighting down or up is determined by the similarity of predicted structures to the structural database S_{3D}. Its definition is as follow:

$$w = \begin{cases} \frac{S_{3D}}{0.5}, & C_{3D} > 0.5 \\ 1, & C_{3D} \leq 0.5 \end{cases} \tag{8}$$

Analysis on the fitness score

Several analyses were performed to evaluate the rationale of the fitness score *S*. All 425 members in SLC family are evaluated using the fitness score. The 15 SLC members with the highest fitness scores (Extended Data Table 3) have their own or homologous structures, which are recognized by finding the most similar structures to the predicted 3D structures from all known structures. The SLC members with experimental structures are logically known folds and should have high fitness scores. Among all SLC members, 28 SLC genes have experimental structures containing transmembrane regions. Compared with all SLC members, the scores of these 28 SLC members are all higher than 0.4 (Extended Data Figure 1a-b). That means these proteins are less likely to have new folds and illustrates the fitness score can recognize known folds.

Next, we assume that all known structures are SLC members with the LeuT[29] fold, rather than all structures in PDB database and recalculate the scores for all SLC members. In principle, all other folds like MFS are different from LeuT and should be recognized as new folds with low fitness score values. As shown in Extended Data Figure 1c, most of scores of SLC members are less than

0.2 while only small percentage of SLC members have high score values, which confirms that the score can distinguish new fold and known fold. Finally, an interesting observation is that SLC20 was listed in the top position when the ranking result was firstly obtained in July 2020. Later, the experimental structure 6L85 of SLC20 was released in September 2020[21], which confirmed it to be a new fold. All these validate the ranking protocol in this work.

SLC members cluster determination

We use 3D structure alignment-based DaliLite.v5[19] to identify SLC fold patterns. The structural similarity of each pair of SLC members is calculated by the intramolecular $C\alpha - C\alpha$ differences in a set of optimized structurally equivalent residue pairs from predicted structures.

Expression and purification of human SLC44A1

The full length sequence of human *SLC44A1* was amplified by PCR from human cDNA library, and subcloned into pCAG vector with C-terminal FLAG-tag. The plasmids were purified devoid of endotoxin before transfecting HEK293F cell lines for protein overexpression. Approximately 1.5 mg plasmids were incubated with 3 mg polyethylenimines (PEIs, Polysciences) in 50 mL SMM 293T-II medium (Sino Biological Inc.) for 30 min before transfection. Cells were cultured in Multitron-Pro shaker (Infors, 130 rpm) under 37°C for 48hrs before collection and resuspension in Lysis Buffer (20 mM Hepes, pH 7.4 and 150 mM NaCl). Then the crude extraction was applied with 1.5% DDM and 0.3% CHS (Anatrace) in 4°C for 2 hours after addition of 1.3 mg/ml aprotinin, 5 mg/ml leupeptin, 1 mg/ml pepstatin (Amresco), and 0.2 mM PMSF (Sigma). After centrifugation at ~25000 g for 1.5 hours, the supernatant containing solubilized SLC44A1 protein was load to anti-FLAG M2 affinity gel (Sigma), then washed by Lysis Buffer with 0.015% TWEEN-20. The target protein was eluted by 0.2 mg/ml FLAG peptide and further purified through size exclusion chromatography (SEC, Superose 6, 10/300, GE Healthcare). The peak fractions were checked by SDS-PAGE analysis and concentrated to ~10 mg/ml for cryo-EM sample preparation.

Cryo-EM sample preparation, data acquisition and procession

To prepare cryo-EM samples, aliquots (3 μL) of the concentrated protein were placed on glow-discharged holey carbon grids (Quantifoil Au R1.2/1.3) and blotted for 5 s, then flash-frozen in liquid ethane cooled by liquid nitrogen with Vitrobot (Mark IV, Thermo Fisher Scientific). Data was collected in 300kV Titan Krios equipped with Gatan K3 detector and GIF Quantum energy

filter. Movie stacks were automatically collected using AutoEMation[30] at a defocus range from -2.2 μm to -1.2 μm in super-resolution mode at a nominal magnification of 81,000×. Each stack was exposed for 2.56 s with an exposure time of 0.08 s per frame, resulting in a total of 32 frames per stack. The total dose rate was approximately 50 $e^-/Å^2$ for each stack. The stacks were motion corrected with MotionCor2[31] and binned 2-fold, resulting in a pixel size of 1.087 Å/pixel. Meanwhile, dose weighting[32] was performed. The defocus values were estimated with Gctf[33].

Particles were automatically picked using Relion 3 [34]. After 2D classification, particles were subjected to global angular searching 3D classification against an initial model generated from similar protein by Relion 3. Then, these particles were selected by multi-reference 3D classification with C2 symmetry. After 3D auto-refinement with reference of noise-reduced map[35], the dataset was C2 symmetry-expanded and then subjected to multi-reference 3D classification with adapted and re-weighted mask applied on single protomer with reduced weight on micelle region. Higher regularization parameter was applied to contain more high-resolution information in 3D classification. The selected particles were further optimized by 3D auto-refinement. The resolution was estimated with the gold-standard Fourier shell correlation 0.143 criterion[36,37] with high-resolution noise substitution[38]. Refer to Extended Data Figure 2-3 for details of the data collection and processing.

Model building of human SLC44A1

The model building of SLC44A1 was performed using the predicted structure from RaptorX-3DModeling and AlphaFold through MDFF, which was guided by the 4.4Å density map. As the final quality of the model building heavily relies on the initial pose of the structure relative to the density map, the predicted structures were manually docked to the density map through Chimera[39]. The simulation process adopted 1 fs as time step and CHARMM36m[40] as the force field and the temperature was set to 300 K. The system was first minimized for 5 ps and simulated for 1.5 ns under the bias potential with a scaling factor of 0.3. Next, simulations were performed in three phases with scaling factors of 1, 5, and 10, respectively. The increasing bias potential drove the structure model gradually close to the density map. The cross-correlation coefficient of the structure from RaptorX-3DModeling and density map increased from 0.295 to 0.610, while that from AlphaFold increased from 0.436 to 0.601.

Code availability

Source code for the computation part of the protocol and the complete SLC ranking results are available under an open-source license at https://github.com/JingHuangLab/SLCFold.

Data availability

All input data are available from public sources. Sequences of the human SLC superfamily were downloaded from UniProt database, IDs of which are available in SLC TABLES (http://slc.bioparadigms.org/). For prediction methods, the input databases include UniRef90 v2019_10 (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2019_10/uniref/), UniRef30_2020_03 (http://wwwuser.gwdg.de/~compbiol/uniclust/2020_03/), Uniclust30 v2018_08 (https://wwwuser.gwdg.de/~compbiol/uniclust/2018_08/), metagenome data metaclust_50 in Nov. 2020 (https://metaclust.mmseqs.org/current_release/). Atomic coordinates and EM density maps of SLC44A1 have been deposited in the Protein Data Bank (http://www.rcsb.org) and the Electron Microscopy Data Bank (https://www.ebi.ac.uk/pdbe/emdb/). Correspondence and requests for materials should be addressed to huangjing@westlake.edu.cn (J.H.) and zhouqiang@westlake.edu.cn (Q.Z.).

Reference

1 Hediger, M. A. *et al.* The ABCs of solute carriers: physiological, pathological and therapeutic implications of human membrane transport proteinsIntroduction. *Pflugers Arch* **447**, 465-468, doi:10.1007/s00424-003-1192-y (2004).

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

3 Portugaly, E. & Linial, M. Estimating the probability for a protein to have a new fold: A statistical computational model. *P Natl Acad Sci USA* **97**, 5161-5166, doi:10.1073/pnas.090559497 (2000).

4 Brenner, S. E. Target selection for structural genomics. *Nat Struct Biol* **7**, 967-969, doi:10.1038/80747 (2000).

5 Portugaly, E., Kifer, I. & Linial, M. Selecting targets for structural determination by navigating in a graph of protein families. *Bioinformatics* **18**, 899-907, doi:10.1093/bioinformatics/18.7.899 (2002).

6 Sasson, O. & Linial, M. ProTarget: automatic prediction of protein structure novelty. *Nucleic Acids Res* **33**, W81-W84, doi:10.1093/nar/gki389 (2005).

7 McGuffin, L. J. & Jones, D. T. Targeting novel folds for structural genomics. *Proteins* **48**, 44-52, doi:10.1002/prot.10129 (2002).

8 Wang, S., Sun, S. Q., Li, Z., Zhang, R. Y. & Xu, J. B. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *Plos Comput Biol* **13**, doi:10.1371/journal.pcbi.1005324 (2017).

9 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature*, doi:10.1038/s41586-021-03819-2 (2021).

10 Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706-710, doi:10.1038/s41586-019-1923-7 (2020).

11 Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, eabj8754, doi:10.1126/science.abj8754 (2021).

12 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature*, doi:10.1038/s41586-021-03828-1 (2021).

13 Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res* **47**, D520-D528, doi:10.1093/nar/gky949 (2019).

14 Yang, J. Y. *et al.* Improved protein structure prediction using predicted interresidue orientations. *P Natl Acad Sci USA* **117**, 1496-1503, doi:10.1073/pnas.1914677117 (2020).

15 Xu, J., Mcpartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *bioRxiv*, 2020.2010.2012.336859, doi:10.1101/2020.10.12.336859 (2020).

16 Bai, X. Y., Moraes, T. F. & Reithmeier, R. A. F. Structural biology of solute carrier (SLC) membrane transport proteins (2018). *Mol Membr Biol* **34**, 65-65, doi:10.1080/09687688.2018.1503851 (2017).

17 Perland, E. & Fredriksson, R. Classification Systems of Secondary Active Transporters. *Trends Pharmacol Sci* **38**, 305-315, doi:10.1016/j.tips.2016.11.008 (2017).

18 Bai, X. Progress in Structural Biology of Solute Carriers. *Current Molecular Biology Reports* **7**, 9-19, doi:10.1007/s40610-021-00144-5 (2021).

19 Holm, L. Benchmarking fold detection by DaliLite v.5. *Bioinformatics* **35**, 5326-5327, doi:10.1093/bioinformatics/btz536 (2019).

20 Luscombe, N. M., Qian, J. A., Zhang, Z. L., Johnson, T. & Gerstein, M. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol* **3**, doi:10.1186/gb-2002-3-8-research0040 (2002).

21 Tsai, J. Y. *et al.* Structure of the sodium-dependent phosphate transporter reveals insights into human solute carrier SLC20. *Sci Adv* **6**, doi:10.1126/sciadv.abb4024 (2020).

22 Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673-683, doi:10.1016/j.str.2008.03.005 (2008).

23 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**, 2302-2309, doi:10.1093/nar/gki524 (2005).

24 Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res* **37**, W465-W468, doi:10.1093/nar/gkp363 (2009).

25 Xu, J., McPartlon, M. & Li, J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell* **3**, 601-609, doi:10.1038/s42256-021-00348-5 (2021).

26 Li, Y. *et al.* Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *Plos Comput Biol* **17**, e1008865, doi:10.1371/journal.pcbi.1008865 (2021).

27 Wang, S., Peng, J., Ma, J. Z. & Xu, J. B. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci Rep-Uk* **6**, doi:10.1038/srep18962 (2016).

28 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577-2637, doi:10.1002/bip.360221211 (1983).

29 Yamashita, A., Singh, S. K., Kawate, T., Jin, Y. & Gouaux, E. Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature* **437**, 215-223, doi:10.1038/nature03978 (2005).

30 Lei, J. & Frank, J. Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. *J Struct Biol* **150**, 69-80, doi:10.1016/j.jsb.2005.01.002 (2005).

31 Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).

32 Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 A reconstruction of rotavirus VP6. *Elife* **4**, e06980, doi:10.7554/eLife.06980 (2015).

33 Zhang, K. Gctf: Real-time CTF determination and correction. *J Struct Biol* **193**, 1-12, doi:10.1016/j.jsb.2015.11.003 (2016).

34 Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, doi:10.7554/eLife.42166 (2018).

35 Wang, N. *et al.* Structural basis of human monocarboxylate transporter 1 inhibition by anti-cancer drug candidates. *Cell* **184**, 370-383 e313, doi:10.1016/j.cell.2020.11.043 (2021).

36 Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J Mol Biol* **333**, 721-745 (2003).

37 Scheres, S. H. & Chen, S. in *Nature methods* Vol. 9 853-854 (2012).

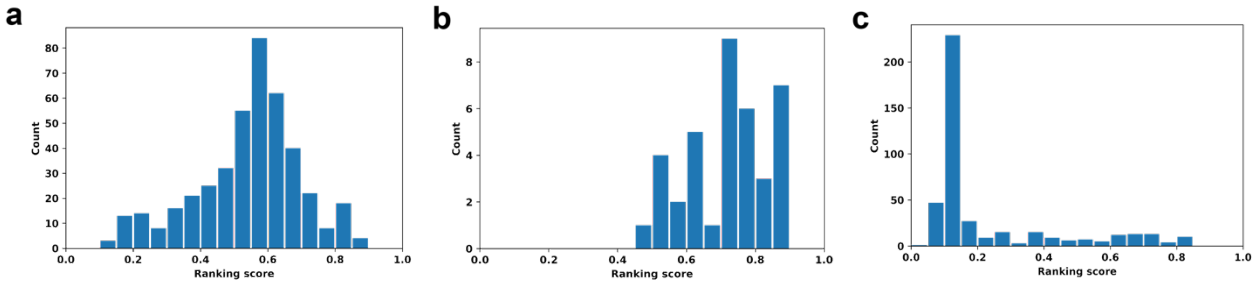
38 Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135**, 24-35, doi:10.1016/j.ultramic.2013.06.004 (2013).

39 Pettersen, E. F. *et al.* UCSF chimera - A visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).

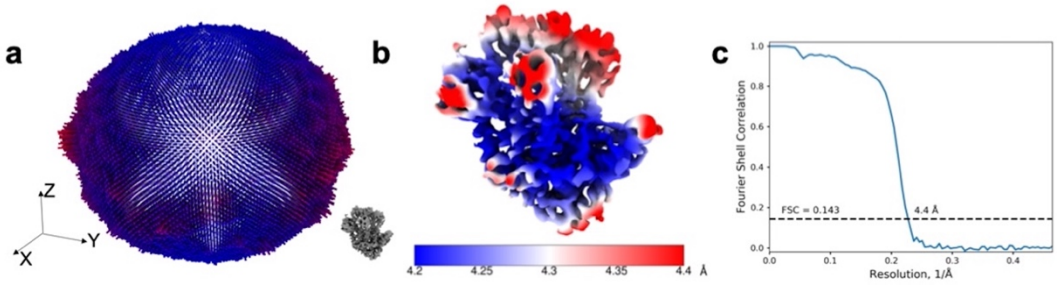
40 Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71-73, doi:10.1038/Nmeth.4067 (2017).

41 Baek, M., Park, T., Heo, L., Park, C. & Seok, C. GalaxyHomomer: a web server for protein homo-oligomer structure prediction from a monomer sequence or structure. *Nucleic Acids Res* **45**, W320-W324, doi:10.1093/nar/gkx246 (2017).

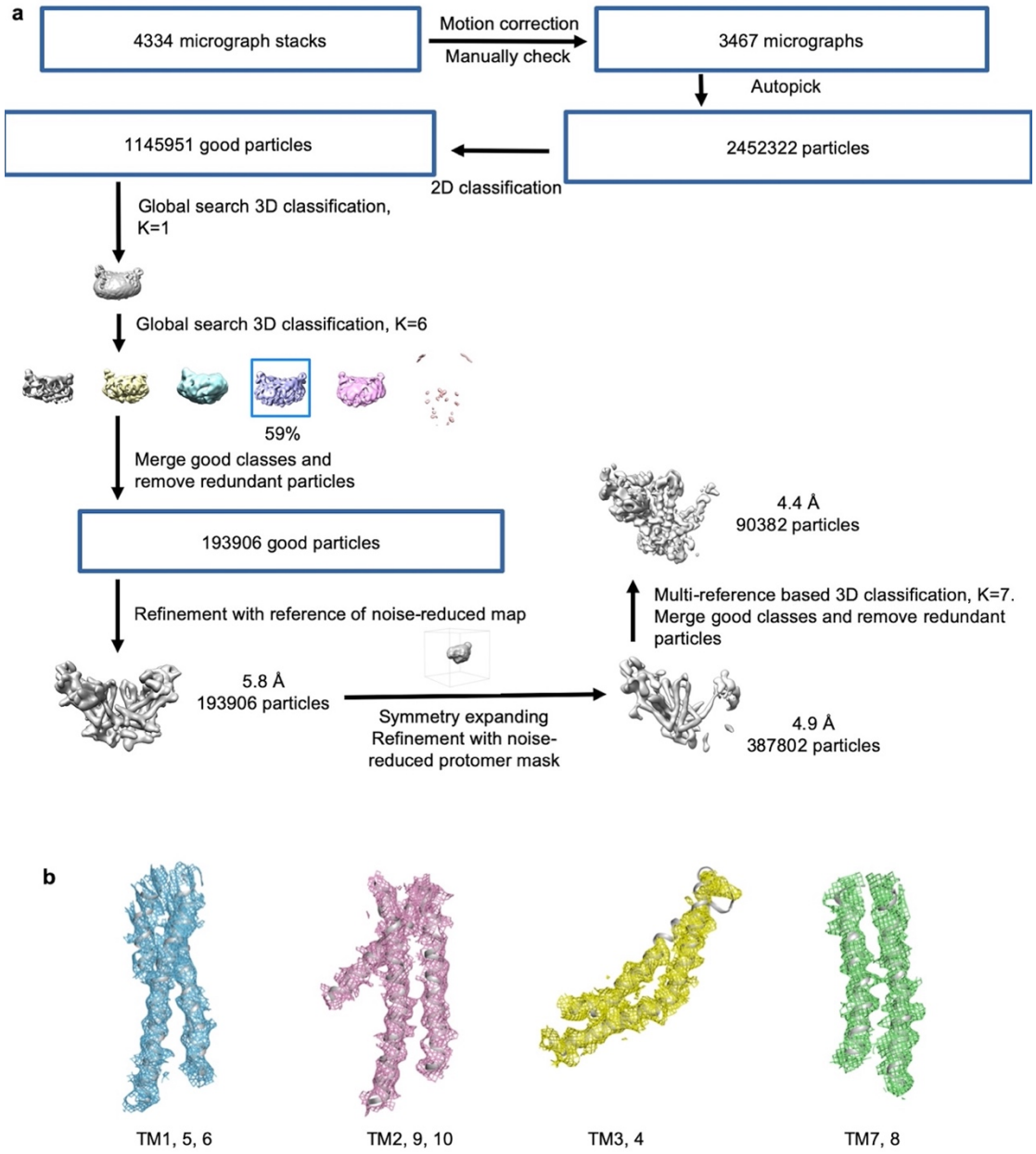
42 Ren, F. *et al.* X-ray structures of the high-affinity copper transporter Ctr1. *Nat Commun* **10**, doi:10.1038/s41467-019-09376-7 (2019).



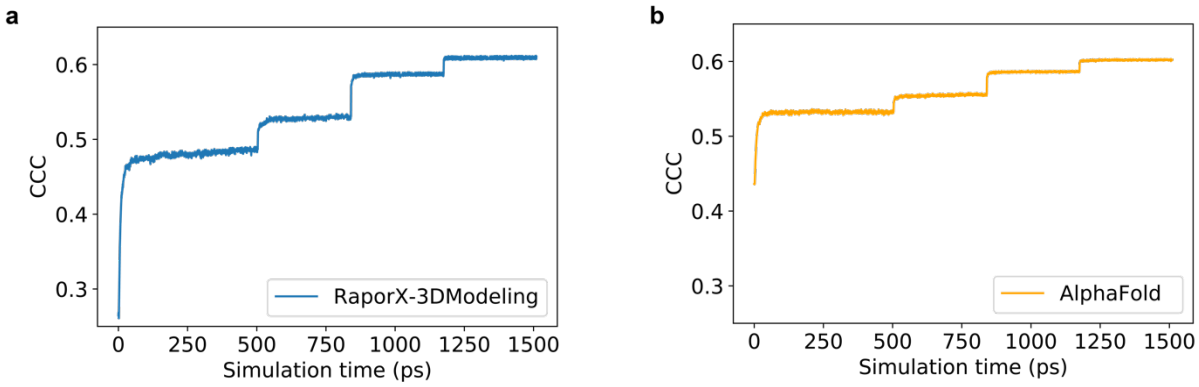
Extended Data Figure 1 | Validation of the fitness score. Histogram of the fitness ranking scores for all of 425 SLC members (a) and structural-known SLC members (b) relative to all known membrane proteins. c, Distribution of the fitness ranking scores for 425 target proteins assuming only LeuT fold structures are known.



Extended Data Figure 2 | Cryo-EM analysis of SLC44A1. a, Euler angle distribution for the 3D reconstruction. b, Local resolution map for the 3D reconstruction. c, Gold standard FSC curves for the 3D refinements applied with protomer mask.



Extended Data Figure 3 | Flowchart for cryo-EM data processing and representative regional maps. a, Flowchart for cryo-EM data processing, please refer to the 'Cryo-EM sample preparation, data acquisition and procession' section in Methods for details. b, electron-density map of the indicated transmembrane (TM) regions. The quality of the map is good enough for identification of all the transmembrane helices.



Extended Data Figure 4 | Cross-correlation coefficient along MDFF simulation. a, RaptorX-3DModeling b, AlphaFold.

Extended Data Table 1 | Ranking score of the top 35 SLC members that probably have a new fold

No.	SLC	S_{3D}	C_{3D}	$\overline{C_{2D}}$	$\overline{C_{1D}}$	S
1	SLC44A5	0.3738	0.6284	0.5653	0.7737	0.1577
2	SLC44A4	0.3844	0.6435	0.5632	0.8017	0.1651
3	SLC39A4	0.3907	0.5317	0.4374	0.8860	0.1808
4	SLC44A2	0.404	0.6251	0.5852	0.7822	0.1835
5	SLC44A1	0.4169	0.6463	0.6207	0.8305	0.1923
6	SLC44A3	0.4167	0.6168	0.6229	0.7981	0.1945
7	SLC55A2	0.3992	0.5414	0.2322	0.8311	0.1965
8	SLC34A3	0.4369	0.7675	0.5709	0.7909	0.2048
9	SLC34A2	0.4295	0.6570	0.5379	0.7588	0.2071
10	SLC8A3	0.3572	0.4359	0.6181	0.7434	0.2135
11	SLC64A1	0.4428	0.6548	0.5864	0.8490	0.2171
12	SLC58A2	0.4349	0.5030	0.6207	0.7976	0.2181
13	SLC55A3	0.4396	0.7086	0.3417	0.8131	0.2182
14	SLC39A10	0.3208	0.2777	0.2395	0.8929	0.2197
15	SLC34A1	0.4471	0.6861	0.5803	0.7755	0.2206
16	SLC8A1	0.3673	0.3988	0.5601	0.7586	0.2219
17	SLC24A4	0.4457	0.6166	0.5852	0.8378	0.2221
18	SLC39A7	0.4413	0.5943	0.5032	0.8729	0.2230
19	SLC24A3	0.4388	0.5388	0.5497	0.8008	0.2232
20	SLC24A2	0.4419	0.5551	0.5882	0.8523	0.2241
21	SLC25A13	0.3641	0.4278	0.5141	0.8069	0.2247
22	SLC8A2	0.3848	0.4500	0.6533	0.7671	0.2256
23	SLC39A6	0.3435	0.3397	0.2185	0.8771	0.2298
24	SLC25A12	0.3882	0.4886	0.5133	0.7969	0.2355
25	SLC53A1	0.4586	0.6671	0.4855	0.8073	0.2359
26	SLC14A2_UT-A1	0.4052	0.4464	0.5495	0.8095	0.2425
27	SLC39A12	0.3872	0.4305	0.4110	0.8434	0.2430
28	SLC24A1	0.3149	0.1043	0.2157	0.8110	0.2447
29	SLC25A24	0.4695	0.6115	0.5912	0.8277	0.2473
30	SLC58A1	0.4417	0.4925	0.6209	0.7952	0.2557

Extended Data Table 2 | SLC Fold information in this work. The IDs of the newly recognized SLC members are shown in orange and bold.

Fold names	No. TMs	Fold features	PDB ID of Rep.	No. families	No. SLCs	Year	PDB ID/chain ID of the earliest structure	SLC family IDs	Fold name previously
MFS	12	TMs 1-6 and TMs 7-12 are "6+6" inverted repeats; TMs 1-3 and TMs 4-6 are "3+3" inverted repeats; TM 4 is surround by TMs 1-3; TMs 1, 2, 4, 5 form the repeat interface while TMs 3, 6 locates in the distance. TMs 1-3 are arranged in a left-handed spiral mode.	1PV6	21	119	2003	1PV6/A	2, 15-19, 22, 29, 33, 37, 40, 43, 45, 46, 49, **52**, **59**, **60**, **61**, 63, O	MFS
LeuT	10	TMs 1-5 and TMs 6-10 are "5+5" inverted repeats; there are unwound regions in the middle of TM 1 and TM 6; between TMs 2-3 and TMs 7-8 locate soluble short helices. Other TMs are in the peripheral area.	6TL2	8	71	2008	3DH4/A	5-7, 11, 12, 32, 36, 38	LeuT
Aac	6	TMs (1, 6), TMs (2,3), and TMs (4, 5) are "2+2+2" rotated repeats.	4C9J	1	53	2014	4C9G/A	25	NA

YddG	10	TMs 1-5 and TMs 6-10 are "5+5" inverted repeats; TMs 1-4 and TMs 6-9 form central cavity, beside which locate TM 5 and TM 10.	6UKJ	1	30	2016	5I20/A	35	NA
UraA	14	TMs 1-7 and TMs 8-10 are "7+7" inverted repeats; there are unwound regions in the middle of TMs 3 and 10; TMs 1-4 and TMs 8-11 form core domain; TMs 5-7 and TMs 12-14 are the scaffold domain.	5I6C	3	24	2011	3QE7/A	4, 23, 26	NA
NhaA	10	TMs 1-5 and TMs 6-10 are inverted repeats; TM4 and TM8 are partially unwound in the middle. TMs 3-5 and TMs 8-10 are the core domain; TMs 1, 2, 6, and 7 are the scaffold domain; the interface of the two domains is composed of TMs 3, 4, 8, and 9.	7CYK	2	20	2009	3FI1/A	9, 10	CPA/AT
ZIP	8	TMs 1-3 and TMs 6-8 are inverted repeats. TMs 4, 5 are sandwiched by the two 3-TM repeats.	6PGI	1	14	2017	5TSB/A	**39**	NA

CNT	10	TMs 1-5 and TMs 6-10 are "5+5" inverted repeats; the first three TMs from N terminal are marked as -2, -1, 0; TMs 1 and 6 are two long, slanted, and bent helices; TMs 2-3 and TMs 7-8 form hairpin; there are unwound regions in the middle of TMs 4 and 9; TMs 3, 4, the end of TM 2, and the start of TM 5 form right-handed spirals.	3TIJ	2	11	2012	3TIJ/A	13, **28**, **34**	NA
YiiP	6	It functions as a dimer. TMs 1-3 and TMs 4-6 are "3+3" rotated repeats and face with each other.	7KZZ	1	10	2009	3H90/A	30	NA
NCX	10	TMs 1-5 and TMs 6-10 are inverted "5+5" repeats; TMs 3-5 and TMs 8-10 are core domain; TMs 1, 2, 6, and 7 are the scaffold domain.	3V5U	2	9	2012	3V5U/A	8, **24**	CaCA
GltpH	12	Two "3+3" inverted repeats: TMs 1-3 and TMs 4-6; TMs 7-9 and TMs 10-12; TMs 1, 2, 4, 5 are core domain; TMs 1-5 are the scaffolding; TMs 7-8 and TMs 10-11 forms two hairpins; unwound or twisted regions exist in the middle of TMs 9 and 12.	6WYL	1	7	2007	2NWL/A	1	Glt

NorM	12	TMs 1-6 and TMs 7-12 are "6+6" rotated repeats. Each repeat forms a bundle domain. TMs 1, 2, 7, and 8 are the domain interface. TMs 3, 4, 5, and 6 locates along a line that surrounds TM1-2. TM 6 is inserted in the between of TMs 4 and 5.	5Y50	3	6	2010	3MKU/A	41, 47, 62	MviN, MATE
AmtB	10	The first and last TMs are marked as 0 and 11. TMs 1-5 and TMs 6-10 are "5+5" inverted repeats. Each repeat form left-handed spirals. TMs 1, 3, 5, 6, 8, and 10 locates at the inter-domain interface. TMs 1 and 11 are locate at the beginning and the end of the fold.	3HD6	2	5	2004	1U77/A	14, 42	Channel like
SLC44	10	TMs 1 and 2 are on the outside of the protein. TMs 3-10 are core domain with 4 groups of two joint TMs. Both TMs (6, 7) and TMs (8, 9) are linked by a helix.	NA	1	5	2021	NA	44	NA

AcrB	12	The first TM is marked as 0. TMs 1-6 and TMs 7-12 are "6+6" rotated repeats. TMs 1 and 7 both have a right-angle bend, half of which is along with the membrane surface. TMs 2, 4, and 5 are the interface. TMs 2-6 forms resistance nodulation division fold.	6V3H	1	2	2005	1T9W/A	**65**	NA
PiT	12	TMs 1-5 and TMs 8-12 are "5+5" inverted repeats. TMs 6 and 7 are inserted in the two repeats. TMs 3-4 and TM 8-9 are two hairpins, where substrate bound.	6L85	1	2	2020	6L85/A	**20**	NA
SWEET 2b	7	TMs 1-3 and TMs 5-7 are rotated repeats, linked by TM 4.	5XPD	1	1	2015	5TCT/A	50	MtN3-like
bacteriorhodopsin	7	TMs 1-4 and TM5-7 are located in different planes, and both behave zigzag shape.	5W0P	1	1	1999	1C3W/A	**51(A)**	NA
SLC64	6	TMs 1-3 and TMs 4-6 are inverted repeats linked by a long helix, which is parallel to the membrane plane.	NA	1	1	NA	NA	**64**	NA
SLC57	10	TMs 1-5 and TMs 6-10 are inverted repeats. TM 3 is inserted in the between of TM 1 and TM 2.	NA	1	6	NA	NA	**57**	NA

SLC53	10	10 TMs without structural repeats.	Part of 6QQ6	1	1	NA	NA	**53**	NA
SLC31	9	Trimer of SLC31 by GalaxyHomomer server[41] is similar with Ctr1(PDB ID: 6M97)[42].	6M97	1	2	NA	NA	**31**	NA
SLC56	6	TMs 1-2, TMs 2-3, and TMs 4-5 are "2+2+2" repeats. It is a channel-like fold.	6C70	1	5	NA	Part of 1T9W	**56**	NA
5TMs	5	SLC58 members have a partial experimental structure (PDB ID: 6S7T) and functions as heteromer.	NA	1	2	NA	NA	58	NA
4TMs	4	4 TM helices	NA	2	2	NA	NA	35(E2), 48	NA
3TMs	3	3 TM helices	NA	1	6	NA	NA	27	NA
1TM	1	SLC51B, SLC54, SLC55 function in complex. SLC3 members are chaperons.	NA	4	9	NA	NA	3, 51(B), 54, 55	NA

"Fold names" are the gene names of the protein belonging to the fold, which was first experimentally resolved, or the number of TMs when the number of TM helices is less than 6, or SLC family ID. "PDB ID of Rep." refers to the PDB ID of the representative structure for a SLC member. "Year" refers to the year that a fold was resolved at the first time.

Extended Data Table 3 | The fitness score and most similar structures of 15 SLC proteins ranked with highest *S* scores.

SLC gene	*S*	PDB ID of the most similar structure	Gene name	Homologs/SLC family/Gene
SLC6A8	0.8038	6VRL	SLC6A4	SLC6
SLC26A9	0.8063	6RTF	Slc26A9	Homologs
SLC6A6	0.8161	6DZY	SLC6A4	SLC6
SLC6A7	0.8186	6VRL	SLC6A4	SLC6
SLC2A5	0.8204	7CRZ	SLC2A3	SLC2
SLC2A11	0.8247	7CRZ	SLC2A3	SLC2
SLC6A4	0.8281	5I75	SLC6A4	SLC6A4
SLC6A1	0.8327	6VRL	SLC6A4	SLC6
SLC6A3	0.8391	5I75	SLC6A4	SLC6
SLC2A1	0.8409	7CRZ	SLC2A3	SLC2
SLC6A11	0.8415	6DZY	SLC6A4	SLC6
SLC6A13	0.8460	6VRL	SLC6A4	SLC6
SLC14A1_UT-B1	0.8515	6QD5	SLC14A1	SLC14A1
SLC2A3	0.8525	7CRZ	SLC2A3	SLC2A3
SLC6A12	0.8547	6DZY	SLC6A4	SLC6

"PDB ID of the most similar structure" is the ID of the structure most similar with the 3D structure predicted by trRosetta among all known structures in RCSB PDB. "Gene name" refers to gene name of those PDB structures. "Homologs/SLC family/Gene" refers to the common ground between a SLC member and the most similar structure, where "Homologs" means the two proteins are homologous, "SLC family" means the two proteins belong to a SLC family but are different genes, and "Gene" means the two structures correspond to the same gene.