S.T. Yau High School Science Award

Research Report

The Team

Name of team member: Chenhang Christopher Zhang
School: Concord Academy
City, Country: Concord, Massachusetts, USA

Name of supervising teacher: Xianjun Dong

Job Title: Associate Professor of Neurology
School/Institution: Yale School of Medicine
City, Country: New Haven, Connecticut, USA

Title of Research Report

SLC Atlas: an integrative multi-scale resource for exploring the human solute carrier superfamily

Date

September 15, 2026

SLC Atlas: an integrative multi-scale resource for exploring the human solute carrier superfamily

Chenhang Christopher Zhang, Xianjun Dong

Abstract

More than 400 solute carrier (SLC) membrane proteins transport ions, nutrients, metabolites, neurotransmitters, and drugs across the membranes that separate biological compartments. Although SLCs have been extensively studied, their sequence relationships, evolutionary conservation, tissue expression, regulatory context, and structural annotations remain distributed across heterogeneous resources, limiting systematic cross-scale analysis. We developed SLC Atlas (https://slc.yalepages.org), a one-stop interface integrating gene annotation, sequence relationships, cross-species conservation, tissue expression, multi-omic tracks, protein structures, and binding domains for 462 human SLC genes assigned to 76 families. Leveraging these data, we examined molecular organization and tissue specialization across the superfamily. Protein-sequence clustering recapitulated substantial family structure (adjusted Rand index 0.637), and ortholog-record breadth across 25 nonhuman species was inversely associated with tissue specificity across 52 anatomical tissue categories (Spearman $\rho = -0.209$, permutation $P = 0.0001$). In contrast, protein identity showed no detectable global association with similarity of tissue-expression profiles ($\rho = -0.0056$, permutation $P = 0.5941$). Seventy-nine genes met a broad brain-enrichment criterion, although no individual family was significantly enriched after correction for multiple testing. The structural inventory linked 163 genes (35.3%) to experimental structure records, while 299 (64.7%) had AlphaFold records only. Family members also shared more annotated binding ligands and similar transmembrane-helix counts than genes from different families. These analyses identify complementary patterns of sequence organization, expression specialization, and structural coverage, while distinguishing recorded data availability from biological absence.

Keywords: solute carrier; membrane transporter; sequence conservation; tissue expression; protein structure; bioinformatics resource

Acknowledgement

Chenhang Christopher Zhang served as the lead developer and student researcher. He designed and implemented SLC Atlas, assembled and processed the relevant datasets, conducted the analyses, prepared the visualizations, interpreted the results, and drafted the research report. Xianjun Dong supervised the project, provided scientific guidance, advised on data collection and analysis, assisted with interpretation of the findings, and reviewed the research report. Members of the Dong Lab provided comments and feedback on the development of SLC Atlas. No other external assistance was received.

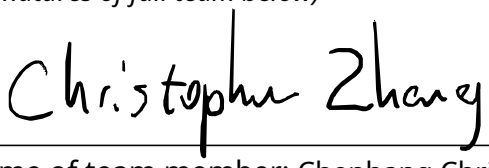
Commitments on Academic Honesty and Integrity

We hereby declare that we

1. are fully committed to the principle of honesty, integrity and fair play throughout the competition.
2. actually perform the research work ourselves and thus truly understand the content of the work.
3. observe the common standard of academic integrity adopted by most journals and degree theses.
4. have declared all the assistance and contribution we have received from any personnel, agency, institution, etc. for the research work.
5. undertake to avoid getting in touch with assessment panel members in a way that may lead to direct or indirect conflict of interest.
6. undertake to avoid any interaction with assessment panel members that would undermine the neutrality of the panel member and fairness of the assessment process.
7. observe the safety regulations of the laboratory(ies) where we conduct the experiment(s), if applicable.
8. observe all rules and regulations of the competition.
9. agree that the decision of YHSA is final in all matters related to the competition.

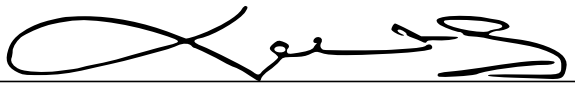
We understand and agree that failure to honour the above commitments may lead to disqualification from the competition and/or removal of reward, if applicable; that any unethical deeds, if found, will be disclosed to the school principal of team member(s) and relevant parties if deemed necessary; and that the decision of YHSA is final and no appeal will be accepted.

(Signatures of full team below)



X _____

Name of team member: Chenhang Christopher Zhang



X _____

Name of supervising teacher: Xianjun Dong

Table of Contents

INTRODUCTION **1**
MATERIALS AND METHODS **2**
Datasets behind the SLC Atlas 2
Sequence acquisition and similarity analysis 3
Cross-species conservation 4
Tissue-level expression 5
Brain enrichment 6
Protein structure and domain annotation 6
Family similarity in ligand annotations and membrane topology 7
Web implementation 8
Statistical analysis 8
RESULTS **9**
SLC Atlas provides a multi-scale view of the human solute carrier superfamily 9
Sequence relationships recapitulate SLC family organization but reveal heterogeneous family cohesion 10
Broader evolutionary conservation accompanies broader tissue expression 11
Protein similarity is largely unrelated to tissue-expression similarity 12
Brain-enriched expression is distributed across SLC families 13
Family members share binding-ligand annotations and membrane topology 13
Structural coverage across the SLC superfamily remains heterogeneous 14
Cross-scale exploration of neurotransmitter transport illustrates the utility of SLC Atlas 14
DISCUSSION **15**
CONCLUSION **17**
DATA AVAILABILITY **17**
FIGURE LEGENDS **18**
REFERENCES **28**

INTRODUCTION

The SLC designation covers a diverse set of membrane proteins that carry ions, metabolites, nutrients, neurotransmitters, and other solutes across cellular and organelle boundaries. In doing so, they contribute to metabolism, ion balance, signaling, and communication between cells. More than 400 human genes are grouped into dozens of SLC families according to sequence-based and functional criteria(1). Changes in SLC activity have been linked to numerous human diseases. Because these proteins are accessible membrane components with defined transport activities, they are attractive but still underused therapeutic targets(2–5). Detailed functional and structural information is still unavailable for many members of the superfamily.

Several SLC transporters are central to nervous-system physiology. Indeed, multiple SLC genes serve as routine markers of neuronal and glial populations. Two markers of major glutamatergic populations are SLC17A7 (VGLUT1) and SLC17A6 (VGLUT2); dopaminergic neurons are marked by the dopamine transporter gene SLC6A3 (DAT); and monoamine loading into vesicles is mediated by the SLC18A2 product VMAT2. Additional SLCs contribute to neurotransmitter uptake, amino-acid transport, ion balance, and metabolite exchange in neurons, glia, and the brain vasculature(5). These features make the superfamily useful for relating molecular evolution to cellular identity and specialized physiology in neural and nonneural tissues.

The breadth of SLC biology has motivated several systematic efforts to classify and characterize the superfamily. Previous studies have examined phylogenetic relationships among SLCs, their genetic landscape, tissue distribution, protein-protein and genetic interactions, structural properties, disease associations, and therapeutic potential(1–14). Dedicated resources, including SLC Tables(1), iMusta4SLC(15), RESOLUTE, and others(16, 17), have substantially advanced the organization of SLC-related knowledge. However, information needed to investigate an individual transporter or compare transporters across families remains distributed across specialized genomic, transcriptomic, evolutionary, and structural resources. Moving from an SLC gene to its family relationships, evolutionary conservation, tissue or cell-

type specificity, local regulatory landscape, and protein structure therefore commonly requires querying multiple independent platforms.

To address this need, we developed **SLC Atlas**, an interactive resource that integrates gene and transcript annotation, nucleotide and amino-acid sequence relationships, cross-species conservation, tissue expression, brain transcriptomic and chromatin-accessibility tracks, and protein structural information. Its family-level and gene-level views connect these data around the same transporter. We used the integrated data to quantify agreement between sequence relationships and family nomenclature, relate conservation to tissue specificity, compare sequence and expression similarity, assess brain enrichment, and identify gaps in experimental structural coverage.

MATERIALS AND METHODS

Datasets behind the SLC Atlas

The supplied data inventory behind SLC Atlas contained 462 human SLC genes assigned to 76 families, including 14 families with a single gene member. Ensembl gene identifiers were used to harmonize the supplied annotation, sequence, ortholog, expression, and structure exports. Gene membership and family nomenclature were obtained from the HGNC “Solute carrier families” group (group 752). Gene and transcript coordinates were taken from Ensembl release 116 (GRCh38.p14) via BioMart, and gene function summaries from NCBI Gene. The transcript table retains every annotated Ensembl transcript, and the Ensembl canonical transcript of each gene provided the representative protein and coding sequence used in the sequence analyses. From the GTEx Portal, we obtained the gene-level TPM matrix GTEx_Analysis_2026-05-19_v11_RNASeQCv2.4.3_gene_tpm.gct.gz together with the sample-attribute file GTEx_Analysis_v11_Annotations_SampleAttributesDS.txt(18). This matrix comprised 19,616 samples distributed among 54 tissue categories. Ortholog records came from the Ensembl REST homology API for the vertebrates division of Ensembl Compara, release 116.

SLC Atlas supplies genomic context through an interactive browser that places gene and transcript structures beside functional genomic tracks. For the current release, we

converted single-nucleus RNA-seq and ATAC-seq results from 82 midbrain samples in our ASAP PD5D project (Lu et al. manuscript in preparation) into normalized density tracks. We also obtained the Parkinson's disease GWAS dataset GCST009325, based on Nalls et al. (19), from the GWAS Catalog and display it with the coverage tracks. Every track uses the GRCh38 human reference assembly.

Protein structure data were retrieved from PDBe/3D-Beacons, which reports experimental (PDBe) and predicted (AlphaFold DB, AlphaFill, SWISS-MODEL) structures per UniProt accession in a single format. AlphaFold predicted models and per-residue pLDDT confidence were downloaded from AlphaFold DB (v6). Membrane-topology and ligand-binding features were retrieved from UniProt.

All analyzed gene identifiers matched across the relevant input layers. The annotation layer connects the sequence, conservation, expression, genomic, and structural records. Gene symbols are displayed alongside stable identifiers in the analysis tables so that source-specific naming can be traced.

Sequence acquisition and similarity analysis

Protein and coding DNA sequences came from the single canonical Ensembl transcript selected for each gene through the Ensembl REST API (release 116; GRCh38.p14). Protein multiple-sequence alignment used MAFFT v7.505 in FFT-NS-2 mode (`mafft --retree 2 --maxiterate 0 --quiet --thread -1`) (20); the Atlas displays the supplied protein hierarchy. For any pair, identity was the fraction of aligned, nongap positions containing the same amino acid or nucleotide. The DNA alignment was projected codon by codon from the protein alignment, so it cannot serve as an independent alignment validation. We computed every pairwise identity from these alignments rather than using a precomputed identity or similarity matrix.

Pairwise distances were computed as 1 − identity before average-linkage agglomeration. Each resulting hierarchy was partitioned into 76 clusters; agreement with named families was quantified by adjusted Rand index (ARI), normalized mutual information (NMI), and adjusted mutual information. We also summarized identities for every unordered pair

within the same family and for every pair from different families. To generate null distributions without changing family sizes, 9,999 family-label permutations tested the within-minus-between mean difference, ARI, and NMI separately for protein and DNA. To test correspondence between the DNA and protein matrices, one matrix underwent synchronized permutations of gene labels along both dimensions. Benjamini–Hochberg adjustment covered the full set of seven global comparisons. Family-specific cohesion was evaluated against random, size-matched gene sets and corrected across the 62 families containing more than one gene.

Alignment sensitivity was evaluated using an identity measure weighted by the longer protein length, removal of the supplied PTCHD3 record containing an internal stop, and alternative global pairwise alignments for selected cross-family candidates. The unrooted daylight display preserves the supplied protein hierarchy but uses unit display edges; it is a visualization of sequence clustering, not a time-calibrated phylogeny. Tip color and shape combinations identify families, and tip areas are proportional to the number of members in the corresponding family.

Cross-species conservation

The supplied export contained 10,273 ortholog records across 25 nonhuman species: 14 mammals and 11 other chordates, including Ciona. Ortholog-record breadth was the fraction of these 25 species with an exported record for each human gene; mean identity was calculated across available species records. Missing entries were retained as missing, rather than assigned zero identity or interpreted as confirmed evolutionary losses. Where a gene had multiple orthologs in a species, the representative retained was the ortholog with the highest percent identity to the human gene.

Each conservation-heatmap cell gives the identity recorded for one human gene and one nonhuman species. The 76 families define the row groups, and the protein tree determines order within each family. For SLC17 alone, row order follows the numeric gene suffix; VGLUT2, VGLUT1, and VGLUT3 correspond to SLC17A6, SLC17A7, and SLC17A8, respectively. Columns keep taxonomic clades together and are arranged broadly by relatedness to human. The accompanying species tree follows the same columns, but its branch-length units are unknown and were not treated as divergence

times. We evaluated conservation–specificity relationships with Spearman coefficients and 9,999 permutations of gene labels. Uncertainty intervals came from 1,999 bootstrap samples drawn in family blocks. The within-family analysis first centered global ranks within each family, and its null distribution was generated by permuting specificity residuals within families. Multiplicity correction was applied jointly to the two primary conservation tests and separately to sensitivity analyses.

Tissue-level expression

The primary analysis retained 18,637 samples in 52 anatomical categories after excluding cultured fibroblasts and EBV-transformed lymphocytes. Sample-level TPM values were summarized by the median within each category; tissues were then weighted equally. Source float32 values were promoted to float64 before aggregation.

Tissue specificity was quantified using τ on untransformed tissue-median TPM values(21). For each gene, expression in each tissue was divided by its maximum across the 52 categories; τ is the sum of one minus these normalized values, divided by 51. Larger values indicate greater specificity. No analyzed gene had an all-zero or constant primary tissue profile. Sensitivity analyses used all 54 categories, broader parent-tissue aggregation, and log-transformed expression.

For the sequence–expression comparison, tissue-median TPM values were converted to $\log_2(\text{TPM} + 0.1)$, then standardized across the 52 tissues separately for each gene. For a given gene pair, the Pearson coefficient between standardized tissue profiles served as the expression-similarity measure. Centering and positive rescaling do not alter Pearson correlation, so standardization changes the plotted scale but not the correlation obtained from the log-transformed profiles. Across 106,491 unordered pairs, we related protein identity to expression similarity with a matrix Spearman coefficient and 9,999 permutations that moved each gene's row and column together. Additional analyses used family-restricted permutations or only within-family pairs to assess sensitivity to family structure. Because many pairs share a gene, the permutation framework retained this dependence instead of assuming independent pairs. A Spearman analysis of untransformed TPM profiles served as a sensitivity check. Illustrative pairs met one of three definitions: identity ≥ 0.4 with expression correlation ≥ 0.6, identity ≥ 0.4 with

correlation ≤0, or identity ≤0.15 with correlation ≥0.8. Each member also required maximum tissue-median expression ≥19.3418 TPM, the cohort's 25th percentile. We ranked candidates deterministically by identity × correlation, identity × (1 − correlation)/2, or (1 − identity) × correlation, respectively, and selected pairs without reusing genes. No pair-specific significance tests were assigned.

Brain enrichment

Broad brain enrichment was defined using the median of 12 brain-region TPM medians (B) and the median of 39 nonbrain anatomical-category medians (N). The enrichment score was log2((B + 0.1)/(N + 0.1)). A gene was classified as brain-enriched when B was at least 1 TPM and the score was at least 1. Spinal cord was excluded from this primary brain-versus-nonbrain comparison and included in a separate central nervous system sensitivity analysis. Peripheral nerve and pituitary were retained in the nonbrain group according to the source metadata.

For each family, a one-sided Fisher exact test compared the proportion of brain-enriched genes with that among all remaining SLC genes; Benjamini–Hochberg correction covered 76 families. Median enrichment scores were tested using 9,999 family-size-preserving gene-label permutations and corrected across 62 nonsingleton families. Sensitivity analyses varied the expression floor, fold-enrichment threshold, pseudocount, regional summary, and inclusion of spinal cord. These analyses quantify tissue-level enrichment and do not estimate neuronal versus glial expression; no cell-type expression matrix was available.

Protein structure and domain annotation

Structural coverage was assessed from the supplied PDB and AlphaFold inventory. Experimental records were deduplicated to unique gene–PDB pairs. Genes were classified as having an experimental record, an AlphaFold record only, or neither; an experimental association did not require full-length coverage. Analyses requiring at least 50% or at least 80% reported sequence coverage assessed this definition. AlphaFold version 6 records were present for all genes, including genes with experimental records. Mean pLDDT was summarized descriptively; residue-level confidence arrays were available for 461 genes.

We evaluated experimental undercoverage in 181 supplied protein-hierarchy clades ranging from 5 to 457 genes. For each clade, a one-tailed Fisher exact test contrasted its genes with all genes outside the clade. Because the clades overlap, Bonferroni adjustment defined the primary multiplicity control, with Benjamini–Hochberg values reported secondarily. We kept PDB chemical-component records distinct from experimentally validated ligand interactions and did not treat an AlphaFill link by itself as evidence for a predicted binding site.

Structural information is displayed using Mol* Viewer(22), an interactive three-dimensional viewer. Users can rotate, zoom, and pan each model, switch between the AlphaFold predicted model (colored by pLDDT confidence) and any experimental PDB structures, and examine individual residues, transmembrane, and intramembrane segments, binding sites, and bound ligands. The viewer is coupled with a membrane-topology diagram so that hovering a residue, topology region, or binding site in either panel highlights the corresponding residues in the other, so features identified in the two-dimensional topology map can be inspected directly in the three-dimensional structure.

Family similarity in ligand annotations and membrane topology

Binding-site information from UniProt covered 94 genes. We reduced each gene's records to a set of ligand identities, prioritizing ChEBI identifiers and otherwise using ligand names. For each gene pair, ligand overlap was measured as intersection size divided by union size (the Jaccard index), and we separately recorded whether the pair shared any ligand. Across all 462 genes, we counted unique annotated transmembrane helices but excluded intramembrane segments. A second topology measure divided the total number of residues covered by transmembrane intervals by UniProt sequence length. We verified annotation coordinates against the supplied UniProt sequences.

The two primary endpoints were ligand-set Jaccard similarity and absolute transmembrane-helix count difference. One-sided tests compared within-family and between-family means using 9,999 permutations of whole gene profiles among fixed family slots, preserving family sizes and pairwise dependence. Benjamini–Hochberg correction covered the two primary tests. Sensitivity analyses equally weighted eligible

families, stratified ligand-profile permutations by annotation richness (one, two, or at least three distinct ligands), excluded common ions, or used exact helix-count matches and transmembrane residue fractions. Exploratory family–ligand enrichment used one-sided Fisher tests across families and ligands represented by at least two annotated genes each, with correction across all tested combinations. Sodium-binding annotations in six SLC6 proteins were projected onto SLC6A4 by global pairwise alignment using BLOSUM62 and gap-open/extension scores of −10/−0.5. This descriptive projection tests sequence-position correspondence, not three-dimensional binding geometry.

Web implementation

SLC Atlas was implemented using AtlasForge, an open-source framework we developed for automated construction of interactive gene-family atlases. AtlasForge is designed to generate web atlases for arbitrary gene families through automated data retrieval and analysis.

The SLC Atlas frontend is a single-page application written in React 19 and TypeScript and built with Vite, using MUI for the interface, TanStack Query for server state, Zustand for client state, and React Router for navigation. Interactive views use Mol* Viewer for three-dimensional protein structures and anatomograms from the EBI Expression Atlas (© EMBL-EBI, CC BY 4.0) for the tissue diagrams. Every data endpoint maps to a file path, so the atlas is served as a fully static site, serving the precomputed data as files with no backend or database. The same routes can alternatively be served dynamically by a FastAPI service that reads the Parquet files with Polars. The resource is hosted at https://slc.yalepages.org and requires no registration or login.

Statistical analysis

All analyses ran in Python 3.12.14 with NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, Matplotlib 3.11.1, and Biopython 1.88. We used seed 20260907 for sequence–family comparisons and seeds 20260909, 20260910, and 20260911 for conservation, sequence–expression, and brain-enrichment analyses, respectively. Under the add-one calculation, the reporting floor is $P = 1/(9{,}999 + 1) = 0.0001$. Correlation tests were two-sided; family-restricted matrix tests were referenced to the center of their permutation

null. As defined above, enrichment and undercoverage tests were one-sided. Adjusted P values below 0.05 were considered significant.

RESULTS

SLC Atlas provides a multi-scale view of the human solute carrier superfamily

The current release comprises 462 human SLC genes assigned to 76 families, using a common set of Ensembl gene identifiers to connect gene and transcript annotation, nucleotide and protein sequences, cross-species conservation, tissue expression, genomic and epigenomic tracks, and protein structural information (**Figure 1A**). This harmonized organization enables both family-level comparisons across the SLC superfamily and detailed investigation of individual transporters without requiring users to reconcile identifiers or navigate independently among multiple source databases.

SLC Atlas organizes these data into interconnected family- and gene-centered views. Sequence relationships can be explored through DNA- and protein-based clustering, including rectangular and radial tree representations. The conservation view places individual SLC genes in a cross-species context, whereas the expression view summarizes their abundance in multiple tissues across human body and different regions across human brain. An integrated genome browser connects gene and transcript models with human brain RNA-seq and chromatin-accessibility tracks and GWAS signals. Structural views combine AlphaFold and experimental PDB models with topology, residue, and feature annotations, making it possible to compare the same features in two- and three-dimensional representations.

Because these data layers are organized around the same set of SLC genes, SLC Atlas also provides a platform for examining relationships that are difficult to evaluate from individual resources alone. We therefore used the integrated dataset to ask how sequence relationships correspond to established SLC family organization, whether evolutionary conservation is associated with tissue-expression breadth, whether closely related proteins retain similar tissue-expression programs, and how experimental structural coverage is distributed across the SLC superfamily. These analyses use the Atlas not only as a reference resource but also as a demonstration for examining how

molecular evolution, tissue specialization, and structural characterization relate across a major human transporter superfamily.

Sequence relationships recapitulate SLC family organization but reveal heterogeneous family cohesion

Protein relationships largely recovered the established SLC family organization (Figure 2A). After the 462 proteins were partitioned into 76 clusters, the clustering agreed substantially with annotated families (adjusted Rand index [ARI] = 0.6370; normalized mutual information [NMI] = 0.8243). The corresponding DNA hierarchy produced ARI = 0.6129 and NMI = 0.8125. Every global agreement metric for both sequence types exceeded its family-label permutation distribution (P = 0.0001), showing that primary sequence contains a strong signature of the current family classification.

Pairwise comparisons further supported this organization (Figure 2B). Among 106,491 unordered gene pairs, mean protein identity was 0.2437 for the 3,459 within-family pairs compared with 0.1200 for the 103,032 between-family pairs; corresponding mean DNA identities were 0.3993 and 0.3153, respectively. The distributions nevertheless overlapped substantially, indicating considerable sequence divergence within some families and sequence similarity across family boundaries.

The degree of sequence cohesion varied markedly among SLC families (Figure S1B). Fifty-seven of 62 nonsingleton families showed greater mean within-family protein identity than random gene groups of matched size at FDR < 0.05. Among families with at least three members, SLC36 and SLC41 were highly sequence-cohesive, with mean within-family protein identities of 0.5872 and 0.5581, respectively, whereas SLC35 and SLC66 were substantially more heterogeneous (0.1389 and 0.1436). Thus, although sequence relationships support the overall SLC classification, individual families encompass markedly different degrees of molecular divergence.

DNA and protein identities were moderately correlated across gene pairs (Spearman ρ = 0.5937, permutation P = 0.0001; Figure 2C). Interpretation of highly divergent pairs was limited by alignment coverage: 82,487 protein pairs had non-gap overlap covering less than half of the shorter sequence, and weighting identity by the longer protein length reduced the protein-clustering ARI from 0.6370 to 0.5738. These analyses therefore

support the broad sequence organization of the SLC superfamily without implying reclassification of individual genes.

The marked heterogeneity in sequence divergence across SLC genes and families raised the question of how this molecular organization relates to evolutionary and functional constraint. We therefore next examined conservation across 25 nonhuman chordates and its relationship to the breadth of SLC expression across human tissues.

Broader evolutionary conservation accompanies broader tissue expression

Across 462 human SLC genes and 25 nonhuman chordate species, we identified 10,273 ortholog records among 11,550 possible gene–species combinations (Figure 3A). Conservation breadth varied substantially across genes and families, providing a complementary measure of evolutionary constraint beyond pairwise sequence similarity among human SLCs. Missing ortholog records were treated as missing data rather than evidence of gene absence.

Genes recorded in more of the 25 nonhuman species generally had broader expression across the 52 human tissue categories (Figure 3B). This pattern appeared as a negative Spearman association between ortholog breadth and τ ($\rho = -0.2089$; permutation $P = 0.0001$). The family-block bootstrap interval remained below zero (95% CI, −0.3228 to −0.0967), showing that a small set of individual genes did not account for the relationship.

A complementary conservation measure based on mean protein identity to available nonhuman orthologs showed a similar relationship with tissue specificity (Spearman $\rho = -0.2731$, permutation $P = 0.0001$; Figure 3C). Importantly, the association persisted after centering conservation and tissue-specificity values within SLC families ($\rho = -0.2532$), indicating that it was not explained solely by differences among families. Thus, even among members of the same SLC family, more evolutionarily conserved genes tended to be more broadly expressed across human tissues.

Together, these analyses link evolutionary conservation to the breadth of tissue deployment across the SLC superfamily. The association was modest, however,

suggesting that evolutionary constraint explains only part of the diversity in SLC expression programs.

Protein similarity is largely unrelated to tissue-expression similarity

We next tested whether proteins that are alike in sequence are also deployed similarly across human tissues. They were not: among all 106,491 gene pairs, protein identity was essentially unrelated to tissue-expression similarity (Spearman ρ = −0.0056, permutation P = 0.5941; Figure 4A). Sequence relationships therefore align strongly with SLC family organization but do not predict how similar two genes' tissue-expression profiles will be.

In the raw-TPM Spearman sensitivity analysis, no association was detected when comparisons were restricted to genes within the same SLC family (ρ = −0.0959, permutation P = 0.7398), indicating that the global result was not simply attributable to comparisons between distantly related families. Instead, closely related SLCs could exhibit either similar or markedly divergent tissue-expression programs.

Representative gene pairs illustrate this decoupling between protein sequence and tissue expression (Figure 4B). *SLCO1B3* and *SLCO1B1* combined high protein identity (approximately 0.80) with strongly correlated tissue expression (Pearson r = 0.86). In contrast, *SLC12A7* and *SLC12A5* had similarly high protein identity (approximately 0.75) but divergent tissue-expression profiles (Pearson r = −0.66). Conversely, *SLC25A24* and *SLC66A3* showed low protein identity (approximately 0.04) despite highly similar tissue-expression profiles (Pearson r = 0.92). These correlations describe standardized log2(TPM + 0.1) profiles; values are rounded as displayed in Figure 4B. Additional selected pairs and sensitivity results appear in Figure S3.

Together, these results distinguish two aspects of SLC evolutionary and functional organization. Evolutionarily conserved SLC genes tend to be more broadly expressed across tissues, whereas pairwise protein similarity does not determine the similarity of their tissue-expression programs. Tissue deployment therefore appears to represent a partially independent dimension of SLC diversification beyond primary-sequence relationships.

Brain-enriched expression is distributed across SLC families

SLC family classification primarily reflects sequence homology rather than tissue-specific function. Given the weak relationship between sequence similarity and tissue-expression similarity, we therefore asked whether nervous-system specialization is distributed across individual members of multiple SLC families rather than concentrated within particular families. Comparing 12 brain regions with 39 nonbrain tissues identified 79 of 462 SLC genes (17.1%) as brain enriched, revealing nervous-system specialization among many individual transporters. *SLC1A2*, *SLC12A5*, and *SLC17A7* were among the strongest examples, with enrichment scores of 9.270, 7.845, and 6.945, respectively. Family-level score summaries appear in Figure 5A.

Brain-enriched genes appeared in many SLC families instead of clustering in a few sequence-defined groups (Figure 5B). Nominal enrichment appeared in several families, although every signal fell short after multiplicity adjustment. SLC1 illustrates this distinction: five of its seven members met the brain-enrichment definition, giving P = 0.0021 before adjustment and P = 0.1569 afterward.

These results indicate that prominent brain expression is a property of individual transporters distributed across the SLC superfamily rather than a feature confined to particular SLC families. Together with the lack of association between protein sequence identity and tissue-expression similarity, this pattern further illustrates that tissue deployment represents a distinct dimension of SLC functional specialization beyond sequence-defined family organization.

Family members share binding-ligand annotations and membrane topology

We next asked whether another major dimension of this membrane transporter gene family—structural characterization—is evident in the family organization. Among 94 genes from 32 families with binding-site records in UniProt, mean ligand-set Jaccard similarity was 0.495 within families versus 0.035 between families (202 and 4,169 pairs, respectively; Figure 6A). At least one ligand was shared by 63.4% of within-family pairs versus 6.4% of between-family pairs. All six binding-annotated SLC6 genes had Na^+ annotations, compared with 15 of 88 other annotated genes (Fisher adjusted P = 0.00255). Sodium-binding annotations mapped to seven corresponding sequence

positions in all six proteins. These comparisons concern recorded annotations: a missing ligand annotation does not demonstrate absence of binding, and SLC17 could not be assessed because it lacked binding-site records.

Within-family pairs differed by an average of 0.53 annotated transmembrane helices, compared with 3.34 for between-family pairs (Figure 6B). The two genes in a pair had identical counts in 71.1% of within-family comparisons and 16.1% of between-family comparisons. Every one of the 19 SLC6 proteins had 12 annotated helices, while all 53 SLC25 proteins had six (Figure 6C). For both primary endpoints, permutation P and adjusted P were 0.0001, the smallest attainable values, and sensitivity analyses gave consistent results. Family membership therefore retains information about ligand annotations and coarse membrane topology even when tissue-expression profiles vary. The complete gene–ligand matrix shows both family-associated patterns and gaps in annotation coverage (Figure S4).

Structural coverage across the SLC superfamily remains heterogeneous

SLC Atlas also linked every analyzed gene to an AlphaFold record and a subset to experimental PDB records on their Structure page. Of 462 genes, 163 (35.3%) had at least one experimental structure record and 299 (64.7%) had AlphaFold records only; no gene lacked both record types. Deduplication of 1,848 experimental source rows produced 1,298 unique gene–PDB pairs spanning 1,214 distinct PDB identifiers. Twenty-one families had no experimental structure record, including SLC57 (six genes), SLC24 (five), and SLC56 (five). Six of 181 tested protein-hierarchy clades showed significant experimental undercoverage after Bonferroni correction. AlphaFold mean pLDDT was descriptively similar in the AlphaFold-only and experimental-record groups (medians 80.75 and 82.19), but confidence scores do not replace experimental validation. These results locate gaps in the supplied structural inventory; they are not an exhaustive assessment of the current PDB or of transporter druggability.

Cross-scale exploration of neurotransmitter transport illustrates the utility of SLC Atlas

The tissue-level results provide a concrete entry point for cross-scale exploration. The dopamine transporter gene, *SLC6A3* (DAT), illustrates how SLC Atlas connects tissue

specialization with molecular features. Its highest tissue-median expression occurred in the substantia nigra (23.561 TPM), whereas its median abundance across the 12 brain regions was only 0.012 TPM, highlighting regional specialization that is obscured by a broad brain summary. Users can locate DAT within the SLC6 family, compare its sequence relationships with other neurotransmitter transporters, and inspect ortholog records across 18 of the 25 surveyed nonhuman species (Figure 3A). The structural view links DAT to 17 experimental PDB records and an AlphaFold model, alongside UniProt annotations identifying 12 transmembrane helices and six binding residues for dopamine in four transmembrane helices (TM1, TM3, TM6, and TM8). Its genomic view provides complementary access to transcript models and RNA-seq and ATAC-seq density tracks of cell types in midbrain (Figure 1B). Together, these views connect regional expression with sequence conservation, membrane topology, and ligand annotations, while distinguishing tissue-level observations and annotated features from direct evidence of cell-type specificity or binding mechanisms.

DISCUSSION

SLC Atlas organizes sequence relationships, cross-species conservation, tissue expression, regulatory genomic tracks, and structural records around a common set of 462 human SLC genes in 76 families. This common inventory enables comparisons across molecular and tissue-level data while retaining access to individual gene records.

SLC Atlas complements existing SLC-focused resources that organize nomenclature, phylogeny, interactions, structural biology, functional characterization, and pharmacology(1, 15–17). Its integrated views connect these dimensions with tissue expression and brain-focused genomic tracks.

Feature	SLC Tables	iMusta4SLC	RESOLUTE	HPA	SLC Atlas
Family annotation	✓	✓	✓	partial	✓
DNA similarity					✓
Protein similarity		✓	✓		✓
Cross-species conservation					✓
Tissue expression	✓		✓	✓	✓
Brain cell-type expression			partial	✓	partial

Feature	SLC Tables	iMusta4SLC	RESOLUTE	HPA	SLC Atlas
Brain RNA-seq tracks				✓	✓
Brain ATAC-seq tracks					✓
Genome-browser view					✓
Protein structures		✓		✓	✓
Unified cross-view navigation			partial		✓

Different data layers yielded different relationships. Sequence clusters aligned well with family labels, and genes recorded across more species tended to have broader tissue expression. In contrast, protein identity between a pair of genes was not globally related to the similarity of their tissue profiles. Brain enrichment was also evident for individual genes without surviving family-level significance after multiplicity adjustment. Although predicted structures were broadly available, experimental records covered only 35.3% of genes. These results support evaluating sequence, expression, and structural evidence independently when forming hypotheses about a transporter.

The nervous system provides a particularly compelling setting for such cross-scale investigation. Several of the most widely used markers of neuronal identity are SLC genes, reflecting the close relationship between transporter repertoire and neuronal physiology. Vesicular glutamate transporters, monoamine transporters, and neurotransmitter reuptake systems define distinct aspects of neurotransmitter identity, while SLC transporters expressed by astrocytes, vascular cells, and other non-neuronal populations regulate extracellular neurotransmitter concentrations, ion homeostasis, and metabolite exchange. By placing these familiar marker genes within the broader context of >400 related transporters, SLC Atlas may facilitate identification of less well-characterized family members with analogous or complementary cellular specializations.

These findings have several limitations. The ortholog export spans 25 chordates, and an unrecorded ortholog could reflect loss, divergence, or the database's annotation and selection process; the lack of a record therefore cannot establish true gene absence. The species represented in the export also determine the calculated mean

identity. For distant SLC pairs, nongap identity may be inflated when only a short region aligns, and the codon-level DNA alignment was projected from the protein alignment. Medians from bulk tissue cannot separate cell-type composition, while the available midbrain bigWig summaries are not a cell-type expression matrix. The brain-enrichment definition is a descriptive threshold rather than a donor-level differential-expression analysis, and RNA abundance is not a direct measure of transporter activity. Experimental structure records can cover only part of a protein or a related sequence; neither AlphaFold confidence nor a chemical-component annotation validates a binding site. These constraints make regular updates and source-level provenance necessary.

Future development may incorporate expression and chromatin-accessibility QTLs, transporter substrate annotations, and per-cell-type expression. The underlying atlas-generation workflow has also been generalized as AtlasForge, an open-source framework for constructing comparable resources for other gene families. This separation between the SLC-specific biological resource and its generalizable software infrastructure should facilitate continued development of SLC Atlas while enabling analogous resources for other gene families.

CONCLUSION

SLC Atlas provides an integrated framework for exploring the human solute carrier superfamily through sequence, conservation, tissue expression, genomic context, and protein structure. The accompanying analyses support family organization and a modest conservation–expression breadth relationship, while identifying limits to sequence-based prediction of tissue deployment and gaps in experimental structural coverage.

DATA AVAILABILITY

SLC Atlas is freely accessible at https://slc.yalepages.org. No registration or login is required. Data incorporated into SLC Atlas were obtained from publicly available resources as described in Materials and Methods. The software framework underlying SLC Atlas, AtlasForge, is available at https://github.com/TheDongLab/AtlasForge.

FIGURE LEGENDS

Figure 1. SLC Atlas architecture and cross-scale exploration of an individual SLC gene. (A) Schematic overview of SLC Atlas, which integrates complementary data layers for the human solute carrier superfamily, including gene annotation, family and evolutionary context, tissue expression, sequence and expression-based clustering, regulatory genomics, and structural information. Regulatory genomic tracks include cell-type-resolved RNA-seq and ATAC-seq data together with genetic association signals. (B) Representative views of SLC6A3 (DAT) illustrating cross-scale navigation across the Atlas. Gene-centered information can be explored through annotation, family, and clustering relationships, tissue-expression patterns, genomic and regulatory tracks, and two- and three-dimensional protein structural views. All views are interconnected through a common gene-centered interface, enabling information from sequence, evolution, expression, regulation, and protein structure to be examined within the same resource. SLC Atlas is available at slc.yalepages.org.

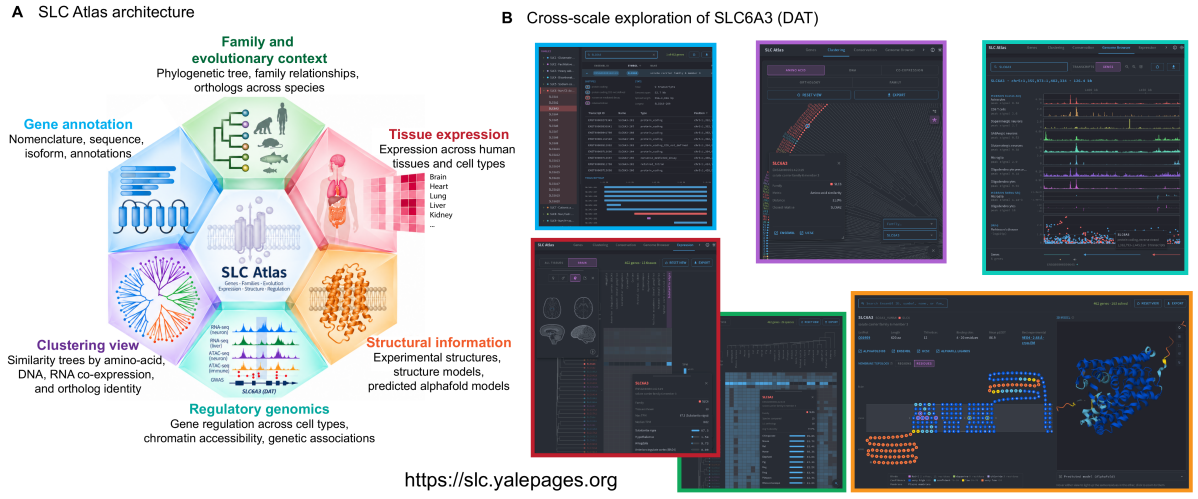


Figure 2. Sequence organization and family concordance. (A) Unrooted daylight display of the protein average-linkage hierarchy for 462 genes. Each tip represents one gene; color and shape combinations denote the 76 families, and marker area is proportional to family size. Unit display edges preserve topology but do not encode evolutionary time or sequence distance. (B) Empirical cumulative distributions of protein and DNA identity for 3,459 within-family and 103,032 between-family pairs. (C) DNA versus protein identity for all 106,491 pairs, with hexagonal-bin counts on a logarithmic scale. DNA identities in panel C come from codons projected onto the protein alignment rather than from a separate DNA alignment. Global tests used 9,999 label permutations.

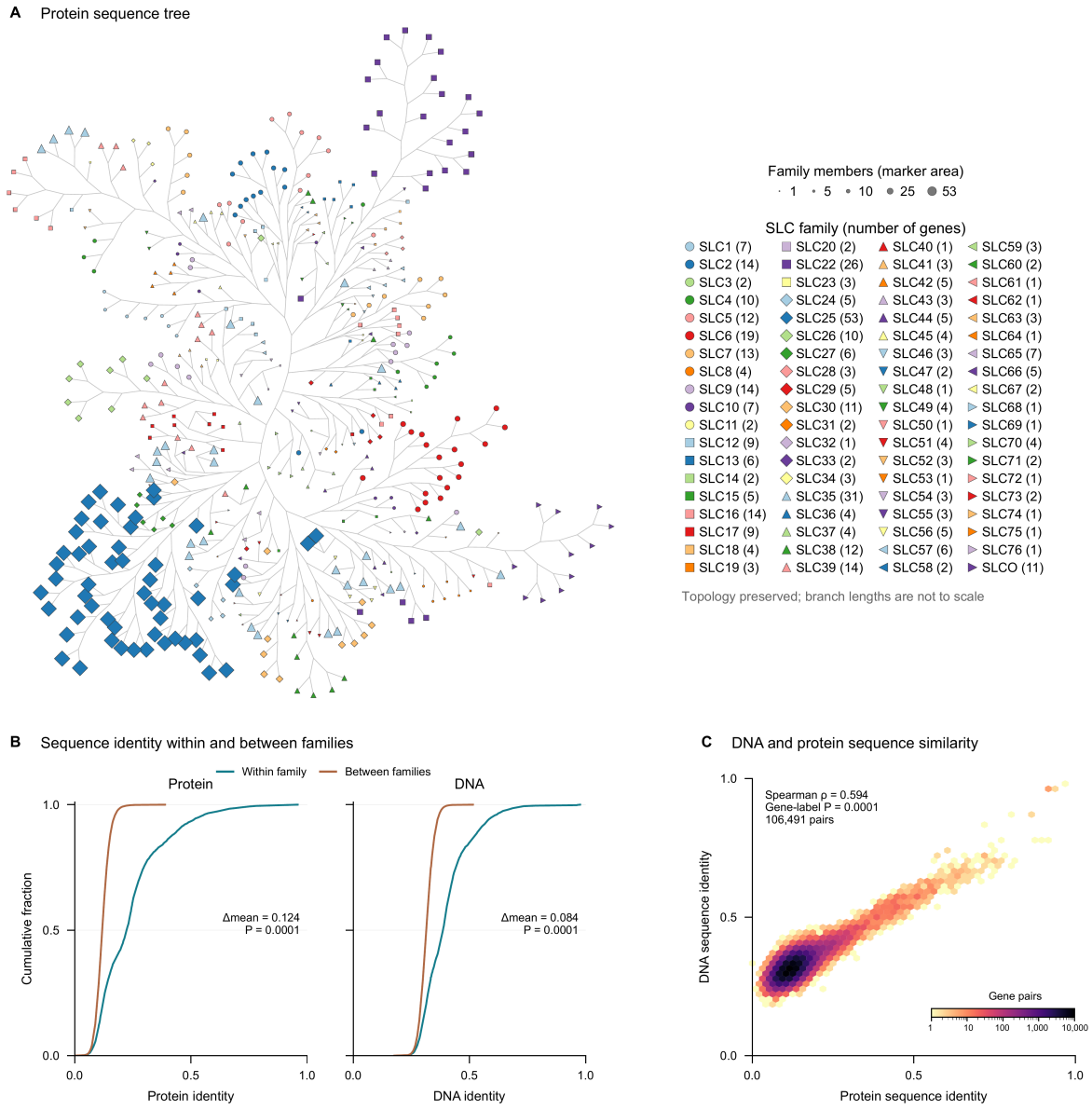


Figure S1. (A) Mean protein identity between families; diagonal cells summarize within-family pairs excluding self-comparisons, with singleton diagonals shown as missing. (B) Sequence cohesion of non-singleton SLC families.

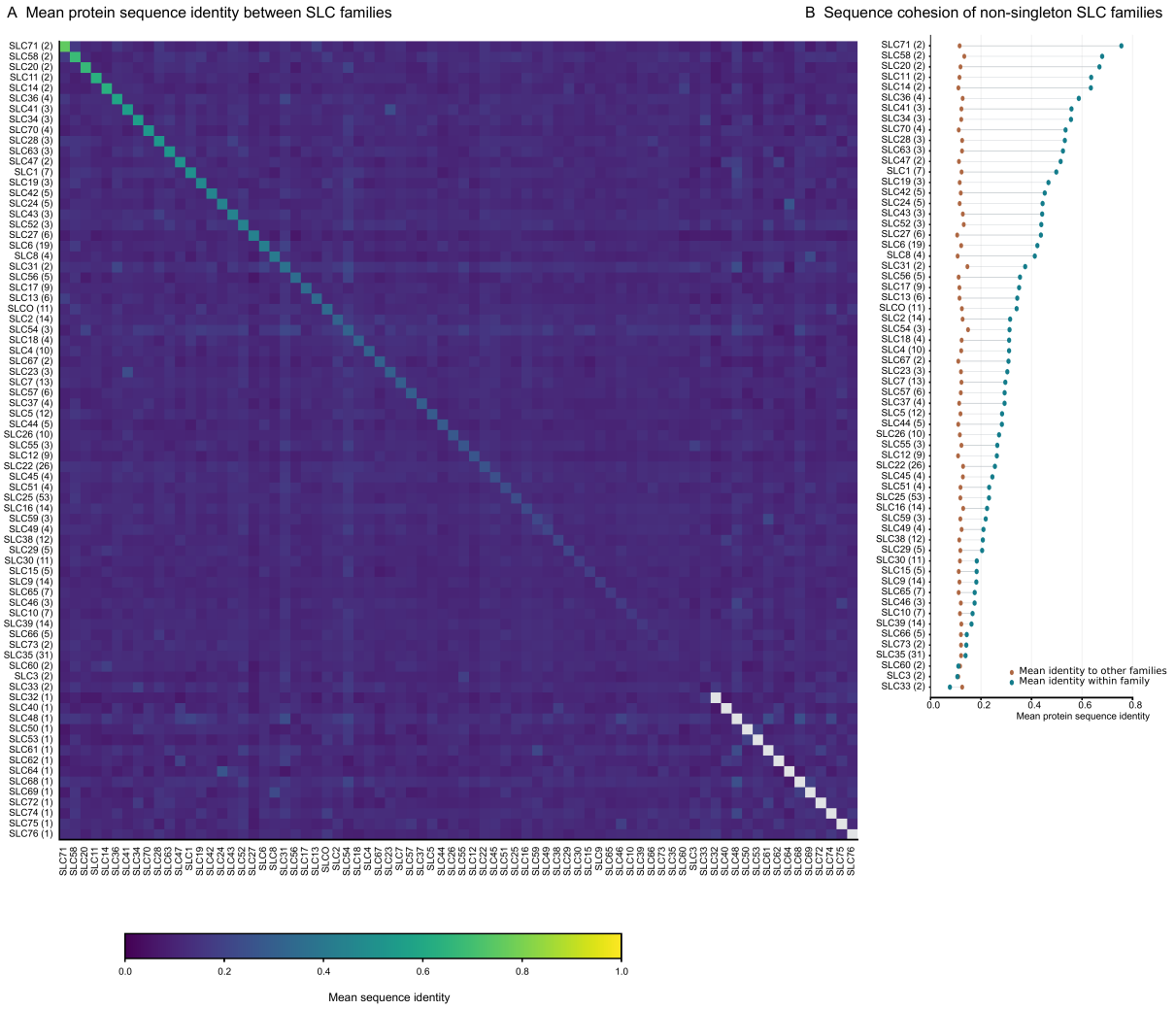


Figure 3. Recorded evolutionary conservation and tissue specificity. (A) Protein identity for 462 human genes across 25 nonhuman species. Rows are grouped into 76 SLC families, with VGLUT1–3 highlighted within SLC17. Columns follow the supplied species tree with clades kept together and ordered by relatedness to human. Gray cells denote missing records, not confirmed absence; tree branch units are undocumented. (B) Ortholog-record breadth versus τ across 52 anatomical tissue categories (Spearman ρ = −0.2089). (C) Tissue specificity by quartile of mean identity across available ortholog records; the continuous association is ρ = −0.2731. Both primary associations have permutation P = 0.0001 and adjusted P = 0.0001. Higher τ indicates narrower expression. For panel C, box midlines indicate medians, box limits indicate the first and third quartiles, whiskers stop at the most extreme observation no farther than 1.5 interquartile ranges from the box, and horizontally displaced dots are individual genes; no test compared the quartiles as discrete groups.

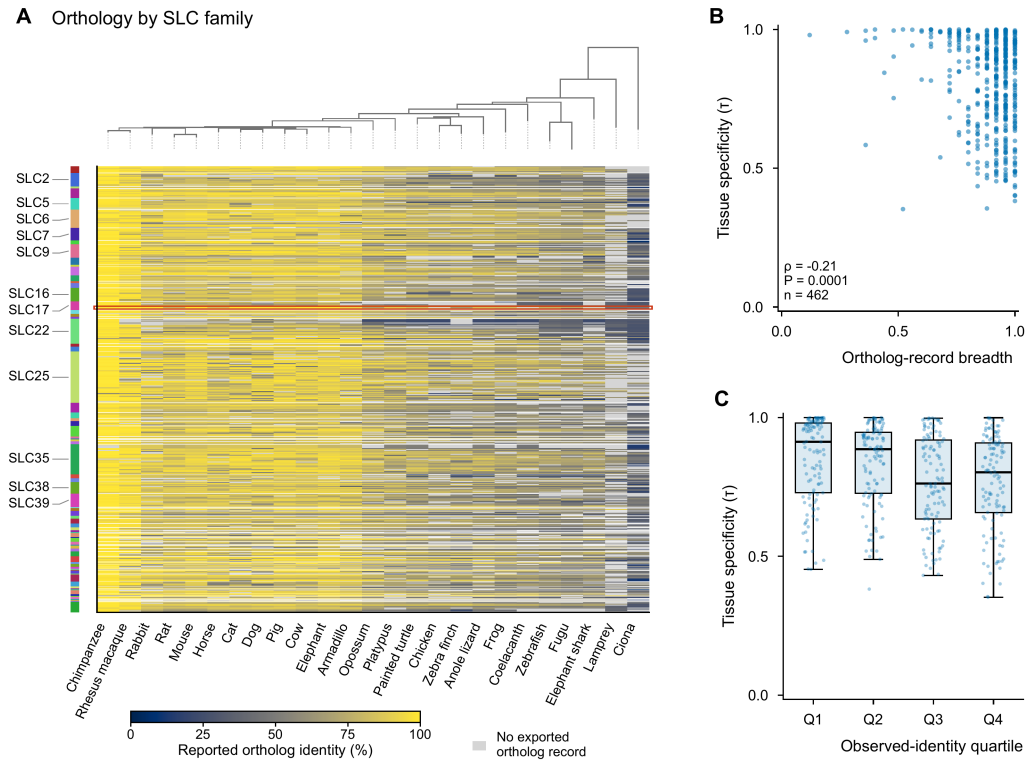


Figure S2. Conservation sensitivity and SLC17/VGLUT detail. (A) Spearman correlations between tissue-expression specificity (τ) and alternative conservation or expression-analysis definitions across 462 SLC genes. The primary analysis uses ortholog-record breadth across 25 nonhuman species and median expression in 52 anatomical GTEx tissue categories. Sensitivity analyses use mean identity among observed orthologs, reciprocal identity, mammalian or nonmammalian record breadth, clade-balanced breadth, all 54 source categories, 30 parent-tissue categories, log-transformed expression, and within-family rank associations. Correlation coefficients appear as points; where calculated, their 95% family-block bootstrap intervals are horizontal, and a vertical reference line identifies zero. **(B)** Percent amino-acid identity of reported orthologs for SLC17A1–SLC17A9 across the 25 nonhuman species, ordered by evolutionary relatedness to human with the supplied species tree aligned above the columns. Red styling identifies the three vesicular glutamate transporters: SLC17A6 (VGLUT2), SLC17A7 (VGLUT1), and SLC17A8 (VGLUT3). An open circle indicates that the corresponding exported orthology record has a classification other than one-to-one; an empty gray cell means that no record was exported and does not establish absence. All three VGLUT genes have Ciona records classified as many-to-many relationships.

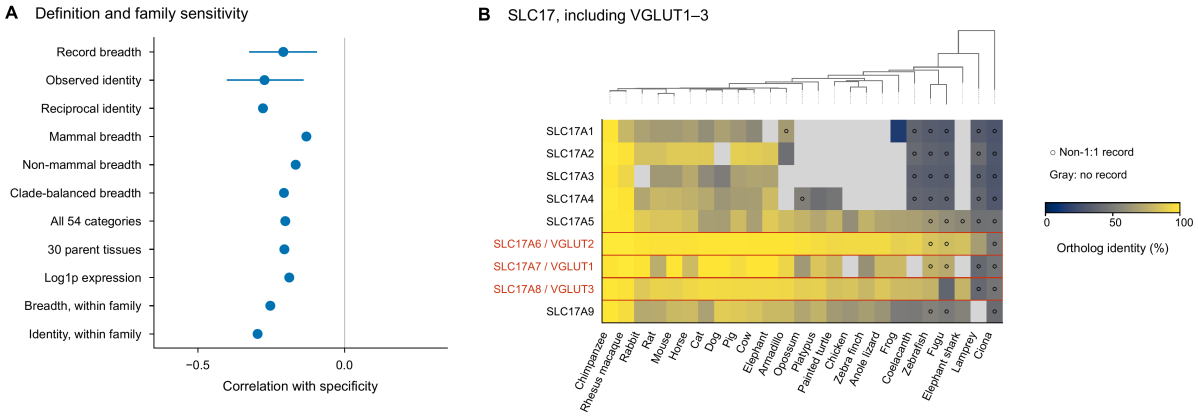


Figure 4. Protein identity and tissue-expression similarity. (A) Pearson similarity of gene-wise standardized log2(TPM + 0.1) profiles across 52 tissues. Hexagonal-bin counts are shown on a logarithmic scale. The association between protein identity and expression similarity is summarized by a matrix Spearman coefficient of −0.0056 (P = 0.5941; 9,999 joint gene-label permutations). (B) Illustrative pairs with high sequence/high expression similarity, high sequence/low expression similarity, or low sequence/high expression similarity. Colors show gene-wise standardized log2(TPM + 0.1) values, truncated from −3 to 3 for display; Pearson correlations use the untruncated standardized profiles. Pairs are descriptive examples selected using prespecified thresholds, without pair-specific significance tests.

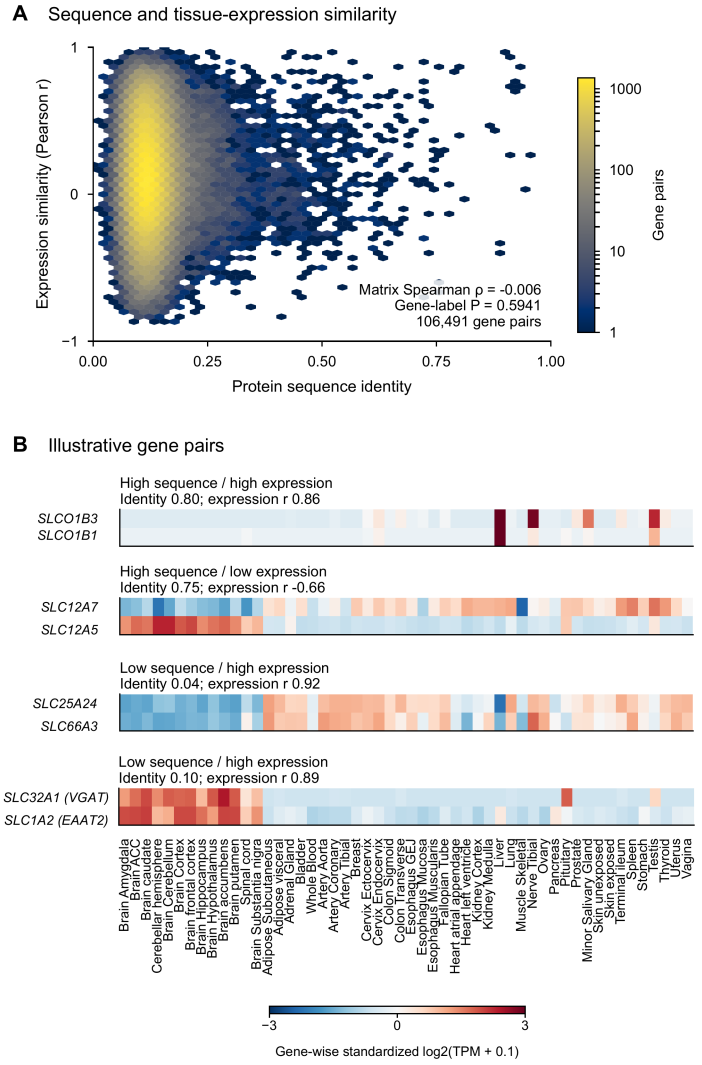


Figure S3. Additional sequence–expression examples and sensitivity analysis. (A) Tissue-expression profiles for six additional gene pairs selected from the predefined high-sequence/high-expression, high-sequence/low-expression, and low-sequence/high-expression categories. **(B)** Sensitivity of the global sequence–expression association to expression transformation, tissue grouping, and tissue inclusion. Points show Spearman correlations between pairwise protein identity and expression similarity for all 106,491 gene pairs (blue), 3,459 within-family pairs (orange), and 103,032 between-family pairs (green). The primary definition uses Pearson correlation of gene-wise standardized log2(TPM + 0.1) profiles across 52 tissues; alternatives use raw-TPM Spearman correlation, all 54 source categories, 30 parent-tissue categories, tissue means, or tissues represented by at least 100 samples. These subgroup estimates are descriptive and do not treat gene pairs as independent observations.

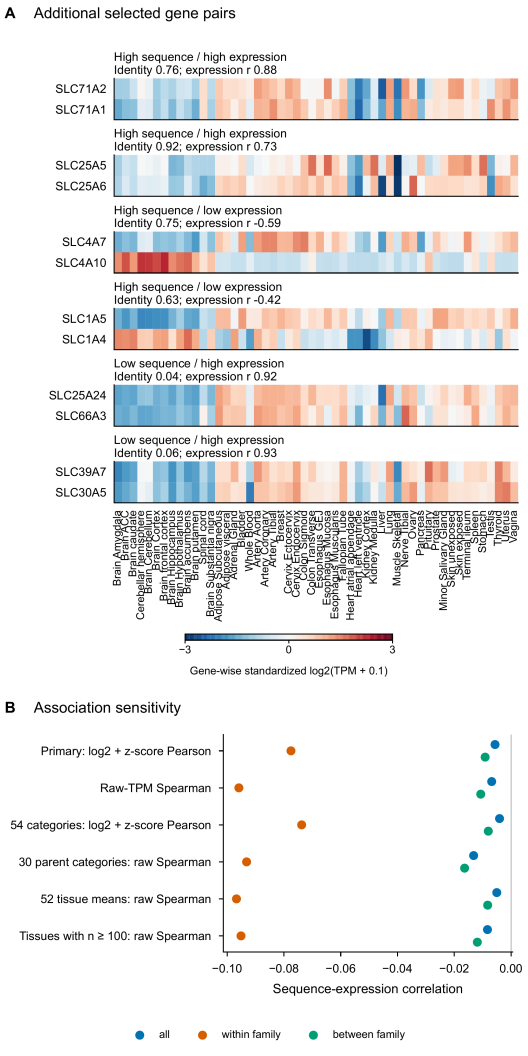


Figure 5. Brain enrichment across SLC families. (A) The 20 families with the highest median brain-enrichment score among families with at least three genes. Scores are log2((B + 0.1)/(N + 0.1)), where B and N are the medians of 12 brain-region and 39 nonbrain tissue medians. Across-gene spread is described using the median and interquartile range; these intervals are descriptive summaries and do not represent confidence intervals. (B) Proportions meeting B ≥1 TPM and score ≥1 in the same families, with enriched/total counts. Seventy-nine of 462 genes meet this criterion. No family is significant at FDR <0.05 in either the median-score permutation test or the one-sided Fisher test of enriched proportions.

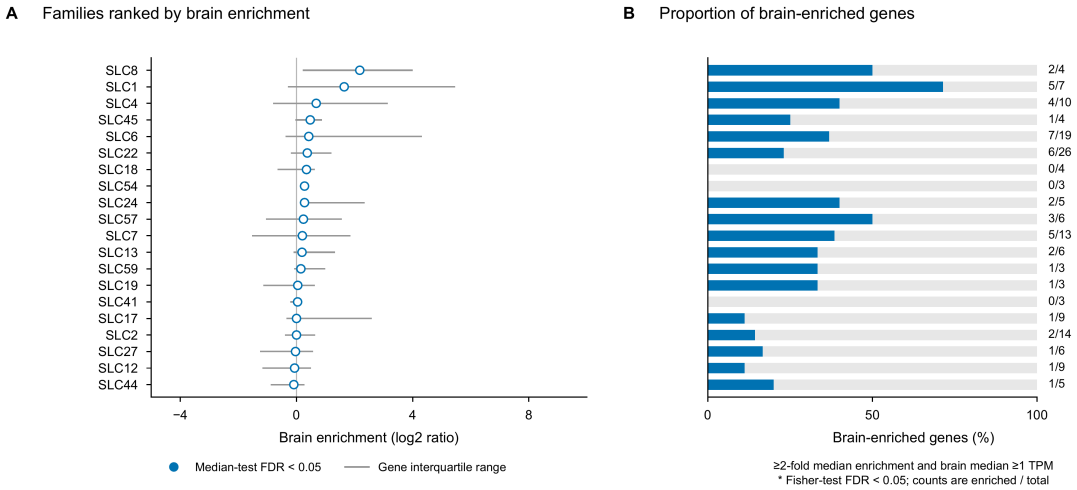


Figure 6. Family similarity in binding-ligand annotations and membrane topology. (A) Mean ligand-set Jaccard similarity for 202 within-family and 4,169 between-family pairs among 94 binding-annotated genes. (B) Mean absolute transmembrane-helix count difference for 3,459 within-family and 103,032 between-family pairs among 462 genes. Both primary comparisons used 9,999 whole-gene profile permutations preserving family sizes (P = 0.0001 and Benjamini–Hochberg adjusted P = 0.0001 for each). Bars report the observed means; the analysis retained the dependence created when pairs shared a gene. (C) Annotated helix counts across 76 families; in panel C, each box covers the first-to-third-quartile range with a median line, whiskers are limited to values no more than 1.5 interquartile ranges beyond the box, and observations outside the whiskers appear as separate points. Helix counts describe annotated topology, not measured three-dimensional structural similarity.

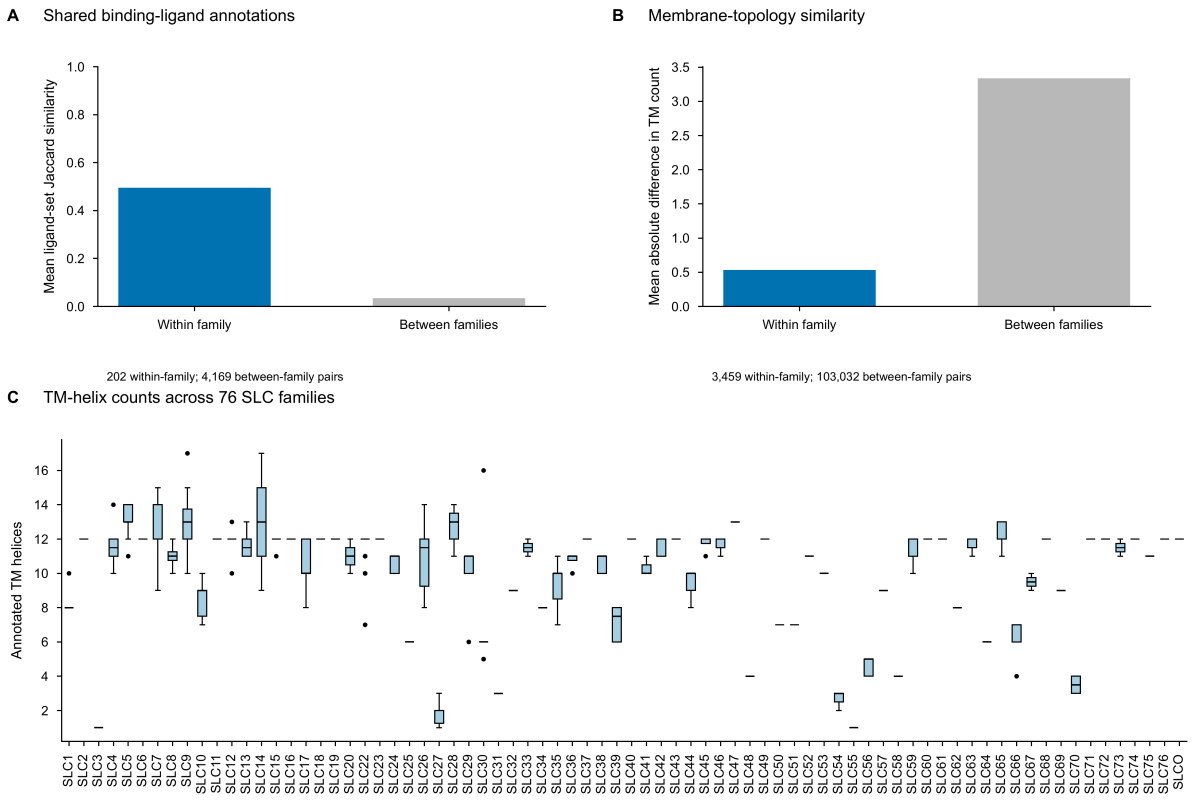
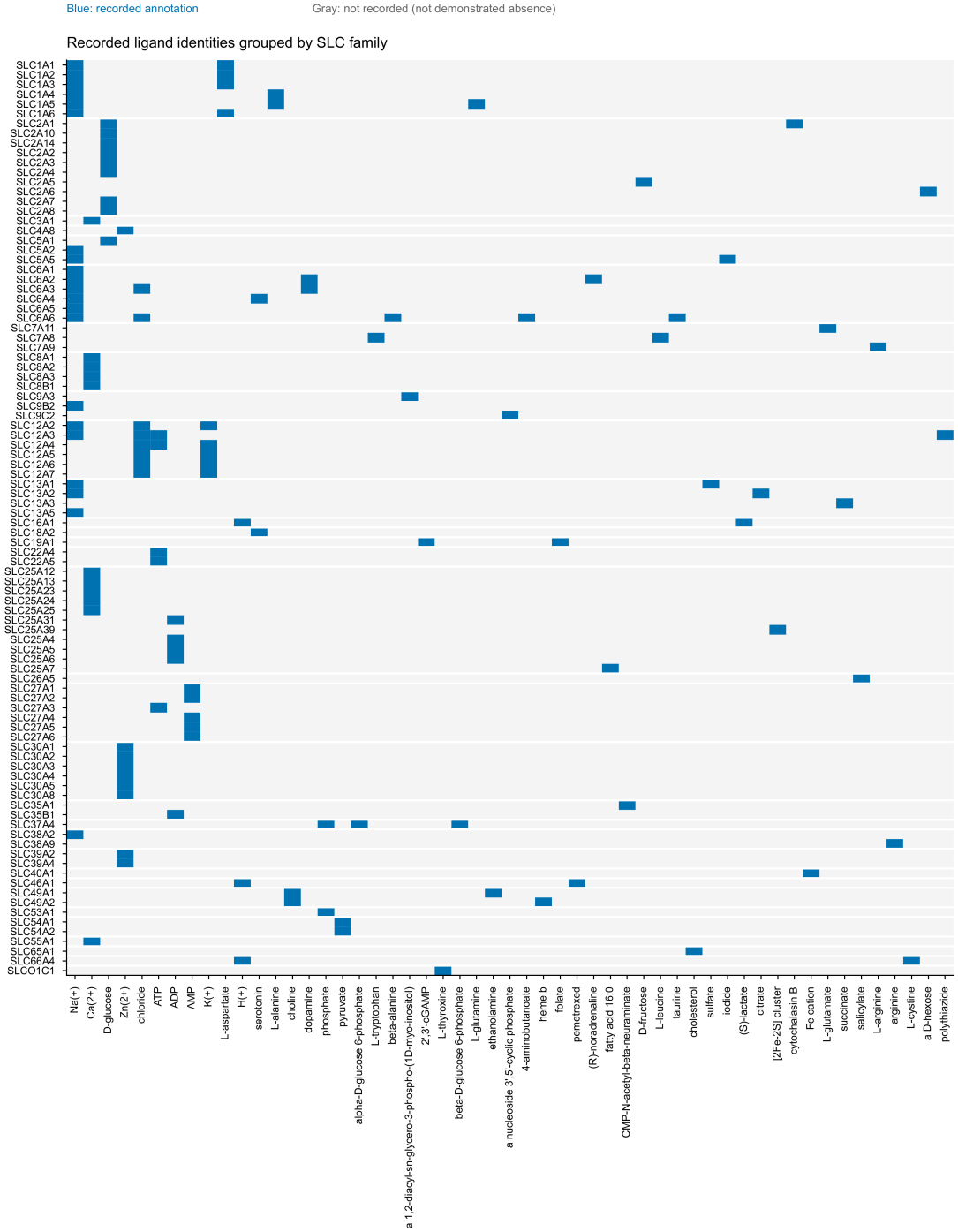


Figure S4. Recorded ligand identities for all 94 binding-annotated genes, grouped by family. Blue denotes a recorded annotation and gray denotes not recorded, rather than demonstrated absence.



REFERENCES

1. Hediger,M.A., Clémençon,B., Burrier,R.E., *et al.* (2013) The ABCs of membrane transporters in health and disease (SLC series): introduction. *Mol Aspects Med*, **34**, 95–107.

2. Cho,S.Y. and Kang,N.S. (2026) The Solute Carrier (SLC) Transporter Superfamily as Therapeutic Targets for the Treatment of Ovarian Serous Cystadenocarcinoma. *Reprod. Sci.*, **33**, 399–410.

3. Chen,R. and Chen,L. (2021) Solute carrier transporters: emerging central players in tumour immunotherapy. *Trends Cell Biol*, **32**, 186–201.

4. Dvorak,V. and Superti-Furga,G. (2023) Structural and functional annotation of solute carrier transporters: implication for drug discovery. *Expert Opin Drug Discov*, **18**, 1099–1115.

5. Hu,C., Tao,L., Cao,X., *et al.* (2019) The solute carrier transporters and the brain: Physiological and pharmacological implications. *Asian J Pharm Sci*, **15**, 131–144.

6. Fredriksson,R., Nordström,K.J.V., Stephansson,O., *et al.* (2008) The solute carrier (SLC) complement of the human genome: phylogenetic classification reveals four major families. *FEBS Lett*, **582**, 3811–3816.

7. Dahlin,A., Royall,J., Hohmann,J.G., *et al.* (2009) Expression profiling of the solute carrier gene family in the mouse brain. *J Pharmacol Exp Ther*, **329**, 558–570.

8. Ho,H.T.B., Dahlin,A. and Wang,J. (2012) Expression Profiling of Solute Carrier Gene Families at the Blood-CSF Barrier. *Front Pharmacol*, **3**, 154.

9. Schaller,L. and Lauschke,V.M. (2019) The genetic landscape of the human solute carrier (SLC) transporter superfamily. *Hum Genet*, **138**, 1359–1377.

10. Denecke,S.M., Driva,O., Luong,H.N.B., *et al.* (2020) The Identification and Evolutionary Trends of the Solute Carrier Superfamily in Arthropods. *Genome Biol Evol*, **12**, 1429–1439.

11. Frommelt,F., Ladurner,R., Goldmann,U., *et al.* (2025) The solute carrier superfamily interactome. *Mol Syst Biol*, **21**, 632–675.

12. Wiedmer,T., Teoh,S.T., Christodoulaki,E., *et al.* (2025) Metabolic mapping of the human solute carrier superfamily. *Mol Syst Biol*, **21**, 560–598.

13. Wolf,G., Leippe,P., Onstein,S., *et al.* (2025) The genetic interaction map of the human solute carrier superfamily. *Mol Syst Biol*, **21**, 531–559.

14. Newstead,S. (2024) Future opportunities in solute carrier structural biology. *Nat Struct Mol Biol*, **31**, 587–590.

15. Higuchi,A., Nonaka,N. and Yura,K. (2018) iMusta4SLC: Database for the structural property and variations of solute carrier transporters. *Biophys Physicobiol*, **15**, 94–103.

16. Superti-Furga,G., Lackner,D., Wiedmer,T., *et al.* (2020) The RESOLUTE consortium: unlocking SLC transporters for drug discovery. *Nat Rev Drug Discov*, **19**, 429–430.

17. Uhlén,M., Fagerberg,L., Hallström,B.M., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

18. Adult GTEx bulk tissue expression data, release v11 *GTEx Portal*.

19. Nalls,M.A., Blauwendraat,C., Vallerga,C.L., *et al.* (2019) Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-genome wide association study. *Lancet Neurol*, **18**, 1091–1102.

20. Katoh,K. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059–3066.

21. Kryuchkova-Mostacci,N. and Robinson-Rechavi,M. (2016) A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 10.1093/bib/bbw008.

22. Sehnal,D., Bittrich,S., Deshpande,M., *et al.* (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res*, **49**, W431–W437.