Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Inferring transporter specificity through language model–guided graph learning

Tong Pan
University of Adelaide
Jie Liu
University of Adelaide
Jing Xu
University of Adelaide
Rong Han
Tsinghua University
Zixu Ran
Northwest A&F University
Xudong Guo
Northwest A&F University
Yi Hao
Northwest A&F University
Yue Bi
Monash University
Fuyi Li
University of Adelaide
Jiangning Song
jiangning.song@monash.edu

Monash University

Research Article

Keywords: Membrane transporters, Substrate specificity, Transporter deorphanization, Protein–small molecule interactions, Deep learning, Protein Language model

Posted Date: March 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8963245/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Genome Biology on October 15th, 2026.
See the published version at https://doi.org/10.1186/s13059-026-04309-2.

Inferring transporter specificity through language model–guided graph learning

Tong Pan[1,2], Jie Liu[1], Jing Xu[1], Rong Han[3], Zixu Ran[4], Xudong Guo[4], Yi Hao[4], Yue Bi[2], Fuyi Li[1,*], Jiangning Song[2, 5,*]

[1]South Australian immunoGENomics Cancer Institute (SAiGENCI), The University of Adelaide, Adelaide, 5005, Australia;

[2]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia;

[3]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;

[4]College of Information Engineering, Northwest A&F University, Yangling 712100, China.

[5]Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Melbourne, Parkville 3010, Australia

*Corresponding author(s). Jiangning Song, Email: jiangning.song@monash.edu; Fuyi Li, Email: fuyi.li@adelaide.edu.au.

Abstract

Membrane transporters orchestrate chemical exchange across cells, playing a central role in cellular metabolism and serving as prime targets for disease intervention. Yet, the specific substrates of most transport proteins remain unknown, largely because experimental characterization is hindered by the instability of membrane proteins in experimental assays. Existing computational tools are often limited to narrow substrate classes or fail to generalize to distant homologs. Here, we present TraSPIN, a geometric deep learning framework that jointly models paired representations of transporter proteins and small-molecule substrates. TraSPIN consistently outperforms existing predictors, demonstrating excellent generalization to both novel substrates and sequence-dissimilar transporters. The model moves beyond mere prediction to reveal the underlying structural mechanisms and motifs governing transporter-substrate recognition. Through large-scale inference across diverse species and successful de-orphanization of uncharacterized solute carrier proteins, we highlight TraSPIN's broad applicability. This powerful framework establishes a robust foundation for deciphering membrane transport, with immediate implications for drug discovery and systems medicine.

Keywords: Membrane transporters; Substrate specificity; Transporter deorphanization; Protein–small molecule interactions; Deep learning; Protein Language model

Background

Membrane transport proteins mediate the selective exchange of molecules across biological membranes and are central to cellular metabolism, signaling and homeostasis [1]. Accurate annotation of transporter substrate specificity is essential for systems-level models of cellular physiology and underpins diverse applications ranging from metabolic engineering to pharmacology [2]. Despite their importance, high-confidence functional annotations for transport proteins remain sparse [3]. Experimental characterization is challenging because membrane proteins are unstable outside lipid environments, and their substrates span a chemically diverse space [4], limiting the scalability of in vitro assays. Consequently, a substantial fraction of transporter genes lacks reliable substrate information, even in well-studied organisms.

Computational methods have traditionally inferred transporter function via homology-based approaches [5, 6], using sequence similarity searches or profile hidden Markov models [7]. These methods perform well when closely related, annotated homologs are available, but their accuracy rapidly declines for divergent transporters [5], and they provide limited resolution at the level of individual substrates. Machine learning approaches partially address these limitations by learning sequence-derived representations to predict substrate classes. Models such as TooT-SC [8] and TranCEP [9] employ classifiers including support vector machines or shallow neural networks and achieve moderate accuracy within predefined substrate categories. Nevertheless, these models remain constrained by limited training data and coarse output spaces. Most models are trained on relatively small datasets despite the availability of far larger protein databases and are designed to predict broad substrate classes rather than specific molecules across diverse transporter families [10].

Recent work SPOT represents a major advance as the first model capable of predicting specific substrates for arbitrary transporters [11]. SPOT jointly encodes transporter sequences and small-molecule representations without relying on experimentally determined protein structures. However, its reliance on sequence-only features limits its ability to capture structural determinants of specificity, hindering generalization to unseen substrates and structurally divergent transporters [12, 13]. These limitations underscore the need for computational frameworks that integrate expressive protein and substrate representations with structural and geometric inductive biases, enabling accurate, generalizable inference of transporter specificity across sequence, structure, and chemical space.

In practical applications, transporter specificity prediction is most critical when experimental annotations are unavailable for closely related transporters [14]. Predictive models must therefore generalize robustly to unseen substrates and to transporters with low sequence similarity to the training set. Existing approaches also offer limited interpretability and only coarse insights into transporter–substrate interactions.

Here, we present TraSPIN, a generalizable transporter specificity inference framework based on geometric deep learning. To capture complex interactions, TraSPIN represents chemical substrates via molecular fingerprints detailing features at both atomic and molecular levels, while transporter properties are encoded using embeddings from state-of-the-art, fine-tuned protein language models. By integrating these representations with predicted protein structures and geometric inductive biases, the framework identifies spatially coherent interaction patterns, pinpointing the specific residues directly involved in substrate binding and transport activity. TraSPIN consistently outperforms similarity-based baselines and existing machine learning methods. Notably, it generalizes effectively to unseen substrates and to transporters sharing less than 40% sequence identity with training data. When applied to the solute carrier (SLC) family [15]—where roughly one-third of members lack annotated

substrates—TraSPIN provides systematic substrate predictions, offering a scalable strategy for transporter deorphanization.

Results

Model overview

The TraSPIN framework is illustrated in Fig. 1. To integrate complementary information from transporter sequences, inferred structural organization, and chemical substrates, we designed TraSPIN as a multi-channel geometric deep learning architecture for transporter substrate specificity inference. By incorporating sequence-derived representations, structure-aware residue partitioning, and molecular-level substrate encodings, TraSPIN enables accurate and generalizable prediction of transporter–substrate relationships across diverse protein families and chemical space. TraSPIN comprises three core architectural components: a transporter–substrate feature encoding module, a structure-aware residue partitioning channel that decomposes transporters into spatially coherent substructures, and an interaction and fusion module that integrates information across channels to infer substrate specificity.

The Transporter Encoding Channel extracts contextualized residue-level representations from the primary amino acid sequence using pretrained protein language models, including ProtT5 [16] and ESM2 [17]. These models encode evolutionary and contextual information learned from large-scale sequence corpora, generating high-dimensional embeddings that capture residue identities, long-range dependencies, and sequence-level constraints. The resulting embeddings provide a sequence-informed representation of transporter properties independent of experimentally resolved structures. In parallel, the substrate encoding channel represents chemical substrates using dual-level molecular encodings. Atomic-level features capture local physicochemical properties of individual atoms, while a pretrained

SMILES-based transformer [18] generates a global molecular embedding that encodes higher-order chemical context. Together, these complementary representations capture substrate information across both atomic and molecular scales, enabling effective modeling of transporter–substrate interactions.

To incorporate structural inductive biases, TraSPIN includes a Residue Partition Channel that models transporter substructures as graphs. Residues are represented as nodes connected by edges defined by spatial proximity inferred from ESM2-predicted residue contact maps based on spatial proximity [17]. A graph convolutional network propagates information among neighboring residues [19], capturing local geometric context. The resulting residue graph is partitioned into strongly connected subgraphs corresponding to spatially coherent residue regions, which are pooled to generate partition-level embeddings that represent putative functional substructures within the transporter.

During the feature fusion and interaction stage, embeddings from the transporter sequence channel, residue partition channel, and substrate channel are integrated using a multilayer perceptron. An attention mechanism explicitly models interactions between substrates and transporter substructures, enabling the network to weight residue partitions based on their relevance to substrate recognition. Through joint optimization, TraSPIN learns a unified representation that captures spatially localized transporter–substrate interaction patterns. The fused embeddings are finally projected to scalar probability scores corresponding to predicted transporter substrate specificity.

By unifying sequence-derived, structure-aware, and chemical representations within a coherent multi-channel architecture, TraSPIN provides an expressive and interpretable framework for transporter specificity prediction, enabling robust generalization beyond sequence similarity and facilitating mechanistic insights into transporter–substrate interactions.

TraSPIN enables generalizable inference of transporter–substrate specificity

To evaluate the predictive power of TraSPIN, we first evaluated its performance on the standardized SPOT dataset [11], which comprises experimentally validated transporter–substrate pairs (Methods section). This benchmark enabled a systematic and fair comparison of TraSPIN with existing computational approaches within a unified evaluation framework. As a foundational baseline, we compared TraSPIN against similarity-based k-nearest neighbor (k-NN) models [20] that infer transporter specificity via homology transfer (Fig. 2a). For each query transporter, the k (k = 1, 2, 3) most similar proteins in the training set with experimentally characterized substrates were identified, and their associated substrates were assigned as candidate substrates for the target protein. This strategy reflects the conventional assumption that substrate specificity can be propagated through sequence similarity [21]. To further contextualize TraSPIN's performance beyond homology-based inference, we benchmarked it against a diverse panel of machine learning models, including multilayer perceptrons (MLP) [22], logistic regression [23], decision trees [24], random forests [25], and recurrent neural networks (RNNs) [26]. Notably, we compared TraSPIN with SPOT [11], a recently proposed gradient boosting framework for general transporter specificity prediction that represents transporter–molecule pairs using ESM-1b [17] and ChemBERTa [27] embeddings. Predictions were evaluated by comparing inferred transporter–substrate associations against experimentally validated annotations. Model performance was quantified using multiple complementary metrics, including accuracy, maximum F1-score (F_{max}), area under the receiver operating characteristic curve (AUROC), and area under the precision–recall curve (AUPRC), providing a comprehensive assessment of predictive robustness across imbalanced datasets [28].

As shown in Fig. 2a, we first compared TraSPIN's overall accuracy with that of baseline approaches, including similarity-based homology transfer. For the k-nearest neighbor (k-NN) strategy,

prediction accuracy progressively declined with increasing k, highlighting the limited robustness of sequence similarity alone for substrate inference. Even at k = 1, the similarity-based approach achieved an accuracy of only 64.2%, substantially lower than that of TraSPIN (Acc = 92.7%). These results indicate that TraSPIN enables markedly more precise substrate identification for a given transporter, outperforming conventional homology-driven strategies.

We also benchmarked TraSPIN against a panel of machine learning baselines that estimate the probability of a molecule being a non-substrate for a given transporter (Fig. 2a, Fig. S1). All models were trained using identical transporter–molecule pair representations derived from ESM-1b [17] and ChemBERTa [27] embeddings to ensure a fair comparison. Simpler classifiers, including multilayer perceptrons (MLP) and logistic regression, exhibited limited predictive performance (MLP: Acc = 81.3%; logistic regression, Acc = 88.8%). In contrast, ensemble methods and sequential deep learning models, such as random forests and RNNs, achieved only modest improvements (random forests: Acc = 90.3%; RNN: Acc = 90.7%). Consistent trends were observed across additional evaluation metrics, including AUROC, AUPRC and F_{max} . Notably, SPOT and TraSPIN outperformed all other methods across evaluation metrics, achieving the strongest overall predictive performance. Although SPOT and TraSPIN achieved broadly comparable performance in standardized evaluations, TraSPIN did not show a statistically significant global improvement, suggesting that both frameworks possess similar predictive capacity across the test dataset.

We next evaluated model generalization under a more stringent setting in which all substrates in the test set were absent from training (Fig. 2b). Under this unseen substrate split, SPOT exhibited a pronounced decline in predictive performance (F_{max} = 0.04, AUPRC = 0.34), indicating limited robustness to distributional shifts. In contrast, TraSPIN maintained strong performance, achieving substantially higher F_{max} and AUPRC values (F_{max} = 0.44, AUPRC = 0.59), reflecting a markedly

improved precision–recall trade-off in challenging scenarios involving novel substrates. Consistent improvements were also observed for additional metrics. TraSPIN attained an accuracy of 78.3% and an AUROC of 0.843, surpassing SPOT (Acc = 70.5%, AUROC = 0.712) by 11% and 18%, respectively. As shown in Fig. 2c, for the substrate L-alanyl-gamma-D-glutamyl-meso-diaminopimelate(1-) [29], there is no overlapped substrate in the training set. For this unseen peptide anion, TraSPIN successfully identified all five potential transporter proteins within the Solute Carrier (SLC) family across five distinct organisms. In contrast, the SPOT baseline failed to identify any of these transporters, highlighting TraSPIN's superior ability to predict interactions for novel substrates. Collectively, these results demonstrate that TraSPIN generalizes more effectively than SPOT in transporter–substrate specificity prediction under stringent evaluations involving previously unseen substrates.

TraSPIN can accurately predict transporter specificity across diverse substrate classes

Transporters in our dataset span multiple mechanistic categories, including channels/pores, electrochemical potential–driven transporters, primary active transporters, and group translocators (Fig. 2d). The class distribution is markedly imbalanced, with group translocators substantially underrepresented relative to electrochemical potential–driven systems. Here we stratified transporter–substrate pairs in the test set into five chemically defined substrate classes: anions, cations, sugars, amino acids, and a residual "other" category for molecules not belonging to the preceding groups (Fig. 2d, Fig. S2). Across these chemically diverse groups, TraSPIN consistently achieved high precision and recall values (Fig. 2e, Fig. S3). Performance was particularly strong for anionic substrates (mean precision = 0.895; recall = 0.924), while cations and other molecules also yielded precision and recall values exceeding 0.89. Predictions for sugars and amino acids were slightly lower but remained robust (precision = 0.858 and 0.839; recall = 0.883 and 0.844, respectively). Collectively, these results

demonstrate that TraSPIN accurately captures transporter–substrate specificity across chemically heterogeneous substrate classes.

We next examined whether predictive performance depended on prior exposure to specific substrates during training (Fig. 2f, Table S1). For each substrate in the test set, we quantified its occurrence frequency among positive training examples and partitioned the test data accordingly. TraSPIN retained strong performance regardless of substrate frequency, including substrates rarely or never observed during training. This analysis indicates that, despite being trained on a limited repertoire of substrates, TraSPIN learns transferable representations that generalize effectively to a broad chemical space.

To dissect the contributions of individual architectural components, we conducted ablation analyses (Fig. 2g, Table S2). Removing the cross-attention module, which explicitly models transporter–substrate atomic interactions, reduced AUROC from 0.956 to 0.898. Similarly, excluding the substructure-aware module decreased AUROC to 0.880, underscoring the importance of incorporating structural inductive biases to represent transporter substructures. Together, these findings highlight the critical role of explicit interaction modeling and structural features in driving TraSPIN's superior performance across diverse substrate and transporter classes.

TraSPIN captures transporter-specific signals beyond homologous similarity

We evaluated TraSPIN's ability to predict substrate specificity across diverse transporter families by stratifying the dataset according to Transporter Classification Database (TCDB) classes [30] (Fig. 3a). Across all transporter categories, TraSPIN achieved consistently strong performance when compared with experimental annotations. Notably, robust accuracy was maintained even for group translocators, despite their limited representation in the dataset.

We further evaluated model performance across the principal TCDB categories [30], including Channels/Pores, Electrochemical Potential-Driven Transporters (EPDTs), and Primary Active Transporters (PATs), as well as their associated substrates (Fig. 3b). TraSPIN demonstrated particularly strong performance for anion substrates across all transporter classes, achieving an AUC of 0.98 and an AUPRC of 0.93 for EPDTs. For amino acid transport via Channels/Pores, the model reached perfect discrimination (AUC = 1.00; AUPRC = 1.00). High reliability was also observed for sugar transport, with EPDTs attaining an AUC of 0.94.

A critical challenge in protein function prediction is generalizing to sequences that are distantly related to those seen during training. This issue is particularly pronounced for transporter specificity prediction, where close homologs with known function may be unavailable [31]. To investigate TraSPIN's robustness under such conditions, we quantified, for each transporter–substrate pair in the test set, the maximum pairwise sequence identity to any protein in the training set (Fig. 3c, Fig. S4-5). Model performance was then evaluated across bins of decreasing sequence similarity using accuracy and the F_{max}, and compared against the baseline method SPOT. For test instances with high sequence identity to the training data (sequence identity >60%), SPOT and TraSPIN exhibited comparable performance. However, as sequence identity decreased, TraSPIN consistently outperformed SPOT. For proteins sharing less than 40% sequence identity with any training sequence, TraSPIN retained markedly superior predictive accuracy, whereas SPOT performance degraded substantially. As expected, predictive performance improved with increasing sequence identity. Accuracy was approximately 0.80 for proteins with the lowest identity (0–30%) and increased steadily, reaching a plateau of ~0.97–0.98 for sequences with identity >60%. Other metrics followed a similar trend (Fig. 3c, Table S3). These results indicate that TraSPIN delivers reliable predictions even for transporters that are only distantly related to those in the training set.

To further illustrate this capability, we selected three representative fatty acid–transporting proteins from the test set (Fig. 3d), spanning a wide range of sequence identities while sharing similar substrate profiles, including palmitate and linoleate. The first example, Q08857 (Fatty Acid-Binding Protein 1, FABP1) [32], predominantly expressed in the liver, exhibits high sequence identity (0.94) with the training proteins; under these conditions, both SPOT and TraSPIN perform accurately. The second case, Q60714 (Fatty Acid-Binding Protein 3, FABP3) [33], which is highly expressed in cardiac and skeletal muscle, represents a moderate level of sequence identity scenario (0.39). At this level, SPOT fails to recover the correct specificity, whereas TraSPIN remains accurate. The most challenging example, Q91VE0 (Fatty Acid-Binding Protein 12, FABP12) [34], displays similarly low sequence identity (0.38) and is expressed in specialised tissues such as the epididymis and retina. Despite the substantial divergence from known templates, TraSPIN successfully predicts its substrate specificity, while SPOT loses predictive power. Collectively, these results highlight TraSPIN's capacity to capture the intrinsic biochemical and functional properties of transport proteins, enabling accurate specificity inference across diverse protein families and varying level of sequence similarity.

TraSPIN uncovers functional substructures essential for transport activity

To investigate why TraSPIN remains accurate for transporters with low sequence identity to the training set, such as Q60714 (FABP3; identity = 0.39) [33] and Q91VE0 (FABP12; identity = 0.38) [34], where sequence-based methods fail, we examined the structural basis of its predictions (Fig. 4). Although these proteins are highly diverged at the sequence level, their substrate recognition relies on conserved three-dimensional arrangements of functionally critical residues.

Transporter specificity is determined by the spatial organization of residues that form substrate-binding pockets rather than by sequence similarity alone [35]. TraSPIN captures this principle by

learning long-range dependencies and evolutionary constraints, enabling the identification of structurally coherent functional sites even in low-homology regimes. Visualization of residue-level activation scores for Q60714 and Q91VE0 (Fig. 4a), mapped onto their three-dimensional structures in complex with palmitate/linoleate [36] (Fig. 4a), reveals spatially contiguous clusters that define putative fatty acid–binding pockets. Residues assigned high importance localize within these pockets and directly interact with the lipid substrates, despite limited sequence conservation.

This structural interpretability extends beyond lipid transporters. For Neutral amino acid transporter A (P43007) [37], which shares high sequence similarity with the training set (92.1%), sequence-based approaches nonetheless fail to recover correct specificity. In contrast, TraSPIN accurately identifies residues coordinating both Na^+ and proline (Fig. 4b). The highlighted residues cluster near the substrates in three-dimensional space and correspond to functional hotspots essential for ion coupling and conformational transitions during transport. Together, these results demonstrate that TraSPIN captures structurally grounded determinants of transporter specificity, enabling accurate prediction and mechanistic insight even when sequence similarity is low.

Phylogenetic and subcellular determinants of TraSPIN predictive performance

To evaluate TraSPIN's generalisation across evolutionary diversity, we assessed transport specificity prediction performance across species spanning the three domains of life [38]: Archaea, Bacteria, and Eukaryota, using a combined phylogenetic tree (Fig. 5a, Fig. S6-7). Despite being underrepresented in the training data, archaeal transporters exhibited uniformly high performance across all evaluation metrics. This minimal interspecies variability indicates strong robustness to training data imbalance. As shown in Fig. 5b, bacterial transporters similarly showed high accuracy and discrimination across diverse lineages. In contrast, eukaryotic transporters exhibited moderately reduced performance and

greater variability. The predominance of high-performance scores across the phylogeny demonstrates that TraSPIN maintains strong predictive power across most species regardless of evolutionary distance from the training set. The relatively lower performance observed for eukaryotic transporters may arise from increased amino acid sequence diversity and structural complexity characteristic of eukaryotic transport proteins.

We next examined whether subcellular localization contributes to the reduced performance observed for eukaryotic transporters (Fig. 5c). Whereas prokaryotic transporters are confined to one or two membranes separating the cytosol from the external environment, eukaryotic transporters operate across multiple intracellular compartments, potentially increasing topological and functional complexity. Subcellular annotations retrieved from UniProt were used to stratify the eukaryotic test set into four categories [39]: cell membrane, nuclear membrane, organelle membrane, and ambiguous localization. Prediction accuracy was modestly lower for transporters residing in organelle membranes compared with those in the cell membrane. Notably, performance for eukaryotic cell membrane transporters, despite being the strongest among eukaryotic categories, remained inferior to that achieved for prokaryotic transporters. These results indicate that subcellular localization partially contributes to the reduced performance in eukaryotes but does not fully account for the disparity, implying that broader evolutionary and structural differences underlie the remaining gap.

TraSPIN enables systematic deorphanization of human SLC transporters

Solute carriers (SLCs) constitute a large and diverse family of integral membrane proteins that mediate the transmembrane transport of metabolites, signalling molecules and therapeutic compounds [40]. Although they play central roles in cellular metabolism, intercellular communication and pharmacology, approximately one-third of human SLCs remain “orphan” transporters with no experimentally defined

substrates [41]. To assess the potential of TraSPIN for systematic de-orphanisation, we applied the model to a curated compendium of SLC annotations assembled from prior reports [41], comprising 148 transport activities with experimentally supported substrates (Fig. 5e). Notably, 54 of the corresponding transporters were absent from the training dataset, thereby providing an opportunity to evaluate predictive generalization beyond previously seen proteins. TraSPIN generated substrate probability profiles for each SLC, and in every case the top-ranked predictions were fully consistent with the corresponding experimental annotations across the entire benchmark set (Table S4).

Illustrative examples highlight this capability. The protein Q7Z3Q1 [42], sharing a maximal sequence identity of only 0.457 with any training sequence, has been characterized as a lysosomal proton-coupled steroid conjugate and bile acid symporter. TraSPIN assigned a high probability to estrone-3-sulfate transport [43], consistent with its experimentally validated substrate profile. Similarly, Q5SVS4 [44], with a maximal sequence similarity of 0.454 to the training set and identified as a kidney mitochondrial carrier protein, was predicted to transport dicarboxylates such as malonate, in agreement with functional studies. Together, these results demonstrate that TraSPIN can recover known substrate–transporter relationships even for evolutionarily distant proteins, supporting its utility as a scalable framework for annotating orphan SLCs in the human genome.

SPOT improves substrate classes prediction for transport proteins

To further assess model versatility, we extended TraSPIN to a related yet less fine-grained task: direct prediction of transporter substrate classes from primary protein features. A dedicated classifier was trained to classify transporter proteins into seven substrate categories [11]: cations, anions, sugars, amino acids/oligopeptides, proteins/mRNA, electrons, and other substrates, using proteins-derived

representations alone. Following standardized annotation and mapping procedures, 11,664 transporters were categorized and subsequently divided into training (80%) and independent test (20%) sets.

The final model achieved an accuracy of 98.8% on an independent test set comprising 2,333 samples (Fig. S8), outperforming the previous state-of-the-art method SPOT. Both models were trained and assessed on the same dataset, ensuring an unbiased head-to-head comparison. Performance remained consistently high across all seven substrate classes (Fig. 5e, Fig. S9), although the cation and "other substrates" categories exhibited comparatively higher misclassification rates. Visualization of the learned protein embeddings [45] further substantiates this performance (Fig. 5f). Proteins belonging to the same substrate class form compact, well-separated clusters in representation space, indicating that the learned sequence embeddings capture substrate-specific biochemical and functional signatures with high discriminative power.

Conclusions

In this study, we introduce TraSPIN, a geometric deep learning framework for predicting the specific substrates of membrane transport proteins. By jointly modeling transporter representations derived from protein language models and chemically informative substrate fingerprints within a geometry-aware architecture, TraSPIN transcends sequence-only paradigms to establish a structurally informed approach to transporter–substrate inference. Across independent evaluations, the model demonstrated strong predictive performance and robust generalization to transporters with low sequence similarity to the training set. Crucially, this highlights its capacity to operate beyond the narrow homology regimes that constrain traditional annotation strategies.

A fundamental challenge in transporter biology is predicting the specificity for proteins that lack closely related, experimentally characterized homologs [46]. Addressing this need, TraSPIN maintains

high performance for transporters sharing less than 40% sequence identity with the training data. This indicates that the learned representations capture deep functional principles rather than relying on superficial sequence similarity. Furthermore, the model demonstrates remarkable resilience when applied to novel substrates absent from the training set; this suggests that its joint encoding strategy captures transferable biochemical features rather than memorizing known transporter–substrate pairs.

Unlike prior approaches that rely exclusively on primary sequence data [11], TraSPIN incorporates geometric inductive biases and predicted structural context to identify spatially coherent interaction patterns. Attribution analyses and visualization reveal that residues prioritized by the model frequently localize to plausible binding pockets, providing mechanistic insight into substrate recognition. While high-quality protein models are now broadly accessible through advances in structure prediction, effectively distilling functional information from these structures remains a significant challenge. Our results demonstrate that integrating structural priors with expressive sequence embeddings offers a practical and performant intermediate strategy, bridging the gap between purely sequence-based models and computationally intensive, structure-centric approaches.

Our analyses further indicate that predictive performance is sensitive to substrate heterogeneity and the fidelity of existing annotations. Functional classes characterized by vast chemical space or loosely defined substrates exhibit modestly higher misclassification rates, underscoring the importance of rigorous functional curation. Looking forward, the integration of quantitative transport kinetics, subcellular localization data, and biochemical regulatory context may further enhance specificity predictions [11], moving the model toward a more holistic representation of functional transport biology.

From a practical perspective, TraSPIN is particularly suited to streamline transporter de-orphanization efforts. In scenarios where a limited panel of candidate substrates can be experimentally

tested, the model serves as a high-confidence prioritization tool, substantially reducing the experimental burden. Conversely, when applied to large-scale screening of extensive chemical libraries, additional calibration strategies may be required to mitigate false positive rates. By providing systematic predictions for previously uncharacterized members of the solute carrier (SLC) family, TraSPIN demonstrates the scalability and translational potential necessary to map the uncharacterized transporter genome.

In summary, TraSPIN establishes a generalizable and mechanistically informative framework for predicting transporter-substrate specificity. By unifying protein language representations, chemical descriptors, and geometric deep learning, the model advances both predictive performance and biological interpretability. As experimental datasets expand and structure-aware representation learning matures, these integrative approaches will become essential for decoding the complexities of membrane transport. TraSPIN provides a scalable foundation with broad implications for systems biology, drug discovery, and the advancement of precision medicine.

Methods

Dataset for transporter–substrate pair prediction

To construct the dataset for transporter–substrate pair prediction, following [11], we collected experimentally validated transporter–substrate associations from curated databases, including the Gene Ontology (GO) [47] and manually reviewed UniProt entries [48]. Substrates were mapped to standardized SMILES molecular representations [49], and entries corresponding to non-specific molecular entities were excluded. Transporter–proton pairs were further removed to avoid including secondary coupling substrates. After quality control and filtering, the final positive dataset comprised 8,587 transporter–substrate pairs, covering 5,882 distinct transport proteins and 364 unique small-

molecule substrates across diverse transporter families. Since experimentally confirmed negative transporter–molecule pairs are not systematically available, negative samples were generated using a controlled similarity-based sampling strategy as described in [11]. For each positive pair, structurally similar molecules were sampled as non-substrates while excluding candidates associated with highly homologous transporters (>60% sequence identity) to reduce potential false negatives. This resulted in a final dataset of 33,162 labeled transporter–molecule pairs. The dataset was split into 80% training and 20% test sets without overlapping transporter sequences. For downstream analysis, subcellular localization information of eukaryotic transporters was retrieved from UniProt [48] and categorized into cell membrane, organelle membrane, nucleus membrane, or unknown/ambiguous. In addition, transporters in the test set were classified by mechanism of action into channels/pores, electrochemical potential-driven transporters, primary active transporters, and group translocators based on TC annotations when available [30].

Dataset of substrate classification

To construct the dataset for substrate class prediction, substrates were grouped into seven biochemical categories following established transporter classification schemes [30]: cations, anions, sugars, amino acids and oligopeptides, proteins and mRNA, electrons, and other substrates. Transport proteins were labeled according to the class of their associated substrate. To ensure a single-label multi-class setting, transporters associated with multiple substrate classes or lacking sufficient classification information were excluded. This filtering resulted in 11,664 transporter sequences, each assigned to one of the seven substrate classes. As in the binary task, the dataset was split into 80% training and 20% test sets without overlapping sequences.

Transporter protein sequence and structure embedding

The TraSPIN framework takes a transporter protein sequence paired with the SMILES representation [49] of a molecular substrate as the input and the predicted specificity probabilities as the outputs. The transporter sequence $S = \{r_1, r_2, \ldots, r_n\}$, consisting of n residues, is represented as an attributed graph $Trans = (V, A, X)$, which embeds the three-dimensional structure of the input transporter. The nodes are defined by the residue set $V \subseteq S$, where each residue is represented by a feature vector in $X \in \mathcal{R}^{n \times d}$ encoding its properties, while the edges are defined by the adjacency matrix $A \in \mathcal{R}^{n \times n}$, which captures pairwise connectivity based on spatial distance.

For constructing the residue-level features, embeddings are extracted using two pretrained protein language models: ProtT5 [16] and ESM2 [17], yielding $E_1 \in \mathcal{R}^{n \times d_1}$ and $E_2 \in \mathcal{R}^{n \times d_2}$, where $d_1 = 1024$ and $d_2 = 1280$, respectively. The two embeddings enable to capture individual residues and broader protein-level characteristics across the protein sequence. For modeling the transporter's spatial structure, a residue adjacency matrix A is constructed based on the predicted 3D coordinates derived from ESM2 [17].

Overview of the model

The transporter-substrate representation encoder

A Principal Neighbourhood Aggregation (PNA)-based graph encoder [50], a variant of graph convolutional networks, is employed to compute and propagate residue-level representations of transporters. By incorporating multiple aggregators and degree scalers, the encoder enhances message passing among spatially proximal residues and accounts for variations in residue density at functional sites.

Initially, the transporter representations are concatenated and projected to a hidden dimension h using a multilayer perceptron (MLP), producing the initial embedding $E_{ini} \in \mathcal{R}^{n\times \mathrm{h}}$, where $h = 200$ by default:

$$E_{ini} = Norm(MLP([MLP(E_1), MLP(E_2)]))$$

where Norm refers to layer normalization applied to the input features to stabilize optimization and the brackets [] indicate the concatenation operation. The Rectified Linear Unit (ReLU) is used as the activation function.

The initial transporter representation $E_{ini} \in \mathcal{R}^{n\times \mathrm{h}}$, along with the adjacency matrix $A \in \mathcal{R}^{\mathrm{n}\times \mathrm{n}}$, is fed into a PNA module. This module produces residue-level representations $E_{res}^{(i)} \in \mathcal{R}^{n\times d_{\mathrm{i}}}$ through message passing between spatially proximal residues, where d_{i} denotes the embedding dimension of the ith convolutional layer:

$$E_{res}^{(i)} = \mathrm{PNA}\left(A, E_{res}^{(i-1)}\right)$$

For the substrate, the initial embedding $S_{ini} \in \mathcal{R}^{1\times \mathrm{h}}$ is obtained by aggregating atom-level embeddings $X_{atom} \in \mathcal{R}^{\mathrm{m}\times 43}$ into a ligand representation using importance-weighted pooling, followed by concatenation with global molecular features $X_{mol} \in \mathcal{R}^{1\times 1024}$ and projection to a hidden dimension $h = 200$ via an MLP.

$$S_{lig} = softmax(MLP(S_{atom}))^T S_{atom}$$

$$S_{ini} = Norm\left(MLP\left(\left[MLP\left(S_{lig}\right), MLP(S_{mol})\right]\right)\right)$$

The graph-based residue-level partition module

In order to group residues into spatially coherent partitions for structure-aware interaction modeling with molecular embeddings, we applied a graph convolutional network (GCN) [19] to learn an transporter partition assignment matrix $A_{region} \in \mathcal{R}^{\mathrm{k}\times n}$ from the graph defined by the adjacency matrix $A \in \{0,1\}^{n\times n}$ and the residue-level feature matrix $E_{res} \in \mathcal{R}^{n\times h}$. Here, k represents the number of transporter regions for the ith convolutional layer. The transporter partition features $E_{region} \in \mathcal{R}^{\mathrm{k}\times \mathrm{h}}$ is subsequently obtained by pooling residue-level features according to their assigned regions.

$$A_{region}^{(i)} = \mathrm{GCN}\left(A, E_{res}^{(i)}\right)$$

$$E_{region}^{(i)} = A_{region}^{(i)} E_{res}^{(i)}$$

The transporter-substrate interaction and fusion module

Using the substrate molecular representation $S \in \mathcal{R}^{1\times \mathrm{h}}$ and the transporter partition features $E_{region} \in \mathcal{R}^{\mathrm{k}\times \mathrm{h}}$ as inputs, the model infers the propensity of specific transporter regions for transporter-substrate interactions by calculating the relationship between the substrate molecule and various transporter partitions through an attention mechanism. The substrate molecular representation and transporter partition representations are then updated as follows:

$$Att = softmax\left(\frac{SW^{Q} \times (E_{region}W^{K})^{T}}{\sqrt{d_K}}\right)$$

$$\hat{S} = Norm(S + MLP\left((Att \times E_{region}W^{V})\, W^{O}\right))$$

$$\hat{E}_{region} = Norm(E_{region} + MLP((1_k^T S) \odot Att^T))$$

where W^Q, W^K, W^V, and W^o are learnable parameters, d_K is a constant scalar factor, and *Norm* refers to layer normalization.

A message-passing operation is then applied to independently facilitate information exchange within the substrate and transporter substructures. Specifically, substrate representations at both the atomic and global molecular levels, as well as transporter representations at the residue and partition levels, undergo information exchange and updating through a multilayer perceptron (MLP) network as follows:

$$\hat{S}_{atom} = Norm(S_{atom} + MLP(\hat{S}))$$

$$\hat{E}_{res} = E_{res} + MLP((A_{region})^T \hat{S})$$

Through iterative updates of residue-level and atomic-level representations, the final transporter and substrate embeddings are derived using a max-pooling operation. These embeddings are subsequently projected into class probabilities for transporter–substrate interaction classification:

$$\hat{y} = sigmoid(MLP([max(\hat{S}_{atom}), max(\hat{E}_{res})]))$$

Acknowledgements

We thank all providers of public data used in this work for their sharing.

Author's contributions

T.P. conceived the ideas. T.P., J.S. and F.L. designed the experiments. T.P. performed the experiments. T.P., J.L. and J.X. analysed the data and prepared figures. T.P., F.L. and J.S. wrote the manuscript. R.H., X.G., Z.R., Y.H. and Y.B. provided advice on data analyses. J.S. and F.L. supervised the project. All authors contributed ideas to the work and assisted in manuscript editing and revision.

Funding

This work is financially supported by the National Health and Medical Research Council of Australia (grant nos. APP1127948, APP1144652, APP2036864 to J.S.), Australian Research Council (grant no. LP220200614 to J.S.), the Major and Seed Inter-Disciplinary Research projects awarded by Monash University (J.S.), South Australian immunoGENomics Cancer Institute (SAiGENCI) received grant

funding from the Australian Government. F.L. is supported by the Australia National Health and Medical Research Council (NHMRC) Investigator Fellowship (GNT2041439).

Data availability

All raw and benchmark data resources used in this study are publicly available via Zenodo at (https://doi.org/10.5281/zenodo.18753677). The curated datasets were obtained from the following databases: The Gene Ontology (GO) knowledgebase (https://geneontology.org/), Transport classification database (TCDB) (http://www.tcdb.org/), Protein Data Bank (PDB) (https://www.rcsb.org), UniProt (https://www.uniprot.org). All source code required to evaluate the conclusions in the paper is available under the GPL v3.0 license in the GitHub repository https://github.com/BioColLab/TraSPIN.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests.

References

1. Radulovic, M., C. Yang, and H. Stenmark, *Lysosomal membrane homeostasis and its importance in physiology and disease.* Nature reviews Molecular cell biology, 2025: p. 1-17.
2. Galetin, A., et al., *Membrane transporters in drug development and as determinants of precision medicine.* Nature Reviews Drug Discovery, 2024. **23**(4): p. 255-280.
3. Radivojac, P., et al., *A large-scale evaluation of computational protein function prediction.* Nature methods, 2013. **10**(3): p. 221-227.
4. Levental, I. and E. Lyman, *Regulation of membrane protein structure and function by their lipid nano-environment.* Nature Reviews Molecular Cell Biology, 2023. **24**(2): p. 107-122.
5. Sinha, S., B. Eisenhaber, and A.M. Lynn, *Predicting protein function using homology-based methods*, in *Bioinformatics: sequences, structures, phylogeny*. 2018, Springer. p. 289-305.
6. Ye, J., S. McGinnis, and T.L. Madden, *BLAST: improvements for better sequence analysis.* Nucleic acids research, 2006. **34**(suppl_2): p. W6-W9.
7. Attaluri, S., S. McGhee, and M. Stamp, *Profile hidden Markov models and metamorphic virus detection.* Journal in computer virology, 2009. **5**: p. 151-169.
8. Alballa, M., *Predicting transporter proteins and their substrate specificity*. 2020, Concordia University.
9. Alballa, M., F. Aplop, and G. Butler, *TranCEP: Predicting the substrate class of transmembrane transport proteins using compositional, evolutionary, and positional information.* PloS one, 2020. **15**(1): p. e0227683.
10. Saier Jr, M.H., et al., *The transporter classification database: recent advances.* Nucleic acids research, 2009. **37**(suppl_1): p. D274-D278.
11. Kroll, A., et al., *SPOT: A machine learning model that predicts specific substrates for transport proteins.* PLoS Biology, 2024. **22**(9): p. e3002807.
12. Pan, T., et al., *A geometric deep learning framework for genome-wide prediction of enzyme turnover number.* Genome Biology, 2026.
13. Ran, Z., et al., *A scalable equivariant graph network framework for precise protein function prediction.* Genome Biology, 2025. **26**(1): p. 407.
14. Mishra, N.K., J. Chang, and P.X. Zhao, *Prediction of membrane transport proteins and their substrate specificities using primary sequence information.* PloS one, 2014. **9**(6): p. e100278.
15. Khan, A., et al., *Solute carriers: The gatekeepers of metabolism.* Cell, 2025. **188**(4): p. 869-884.
16. Elnaggar, A., et al., *Prottrans: Toward understanding the language of life through self-supervised learning.* IEEE transactions on pattern analysis and machine intelligence, 2021. **44**(10): p. 7112-7127.
17. Bepler, T. and B. Berger, *Learning the protein language: Evolution, structure, and function.* Cell systems, 2021. **12**(6): p. 654-669. e3.
18. Honda, S., S. Shi, and H.R. Ueda, *Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery.* arXiv preprint arXiv:1911.04738, 2019.
19. Wu, F., et al. *Simplifying graph convolutional networks*. in *International conference on machine learning*. 2019. PMLR.
20. Kramer, O., *K-nearest neighbors*, in *Dimensionality reduction with unsupervised nearest neighbors*. 2013, Springer. p. 13-23.

21. Brinkworth, R.I., R.A. Breinl, and B. Kobe, *Structural basis and prediction of substrate specificity in protein serine/threonine kinases.* Proceedings of the National Academy of Sciences, 2003. **100**(1): p. 74-79.
22. Almeida, L.B., *Multilayer perceptrons*, in *Handbook of neural computation*. 2020, CRC Press. p. C1. 2: 1-C1. 2: 30.
23. LaValley, M.P., *Logistic regression.* Circulation, 2008. **117**(18): p. 2395-2399.
24. De Ville, B., *Decision trees.* Wiley Interdisciplinary Reviews: Computational Statistics, 2013. **5**(6): p. 448-455.
25. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
26. Medsker, L.R. and L. Jain, *Recurrent neural networks.* Design and applications, 2001. **5**(64-67): p. 2.
27. Chithrananda, S., G. Grand, and B. Ramsundar, *ChemBERTa: large-scale self-supervised pretraining for molecular property prediction.* arXiv preprint arXiv:2010.09885, 2020.
28. McDermott, M., et al., *A closer look at AUROC and AUPRC under class imbalance.* Advances in Neural Information Processing Systems, 2024. **37**: p. 44102-44163.
29. Mengin-Lecreulx, D., J. Van Heijenoort, and J.T. Park, *Identification of the mpl gene encoding UDP-N-acetylmuramate: L-alanyl-gamma-D-glutamyl-meso-diaminopimelate ligase in Escherichia coli and its role in recycling of cell wall peptidoglycan.* Journal of bacteriology, 1996. **178**(18): p. 5347-5352.
30. Saier Jr, M.H., et al., *The transporter classification database (TCDB): recent advances.* Nucleic acids research, 2016. **44**(D1): p. D372-D379.
31. Gyimesi, G. and M.A. Hediger, *Systematic in silico discovery of novel solute carrier-like proteins from proteomes.* PLoS One, 2022. **17**(7): p. e0271062.
32. Gerhard, D.S., et al., *The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC).* Genome research, 2004. **14**(10B): p. 2121-2127.
33. Hirsch, D., A. Stahl, and H.F. Lodish, *A family of fatty acid transporters conserved from mycobacterium to man.* Proceedings of the National Academy of Sciences, 1998. **95**(15): p. 8625-8629.
34. Carninci, P., et al., *The transcriptional landscape of the mammalian genome.* science, 2005. **309**(5740): p. 1559-1563.
35. Erkens, G.B., et al., *The structural basis of modularity in ECF-type ABC transporters.* Nature structural & molecular biology, 2011. **18**(7): p. 755-760.
36. El-Husseini, A.E.-D., et al., *Synaptic strength regulated by palmitate cycling on PSD-95.* Cell, 2002. **108**(6): p. 849-863.
37. Hillier, L.W., et al., *Generation and annotation of the DNA sequences of human chromosomes 2 and 4.* Nature, 2005. **434**(7034): p. 724-731.
38. Spang, A., E.F. Caceres, and T.J. Ettema, *Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life.* Science, 2017. **357**(6351): p. eaaf3883.
39. Consortium, U., *The universal protein resource (UniProt).* Nucleic acids research, 2007. **36**(suppl_1): p. D190-D195.
40. Pizzagalli, M.D., A. Bensimon, and G. Superti‐Furga, *A guide to plasma membrane solute carrier proteins.* The FEBS journal, 2021. **288**(9): p. 2784-2835.
41. Zhang, Y., S. Newstead, and P. Sarkies, *Predicting substrates for orphan solute carrier proteins using multi-omics datasets.* BMC genomics, 2025. **26**(1): p. 130.
42. Bechtel, S., et al., *The full-ORF clone resource of the German cDNA Consortium.* BMC genomics, 2007. **8**(1): p. 399.

43. Ruder, H.J., L. Loriaux, and M. Lipsett, *Estrone sulfate: production rate and metabolism in man.* The Journal of clinical investigation, 1972. **51**(4): p. 1020-1033.
44. Dunham, A., et al., *The DNA sequence and analysis of human chromosome 13.* Nature, 2004. **428**(6982): p. 522-528.
45. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
46. Montanari, F. and G.F. Ecker, *Prediction of drug–ABC-transporter interaction—Recent advances and future challenges.* Advanced drug delivery reviews, 2015. **86**: p. 17-26.
47. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
48. *UniProt: the universal protein knowledgebase in 2021.* Nucleic acids research, 2021. **49**(D1): p. D480-D489.
49. Weininger, D., *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.* Journal of chemical information and computer sciences, 1988. **28**(1): p. 31-36.
50. Corso, G., et al., *Principal neighbourhood aggregation for graph nets.* Advances in Neural Information Processing Systems, 2020. **33**: p. 13260-13271.

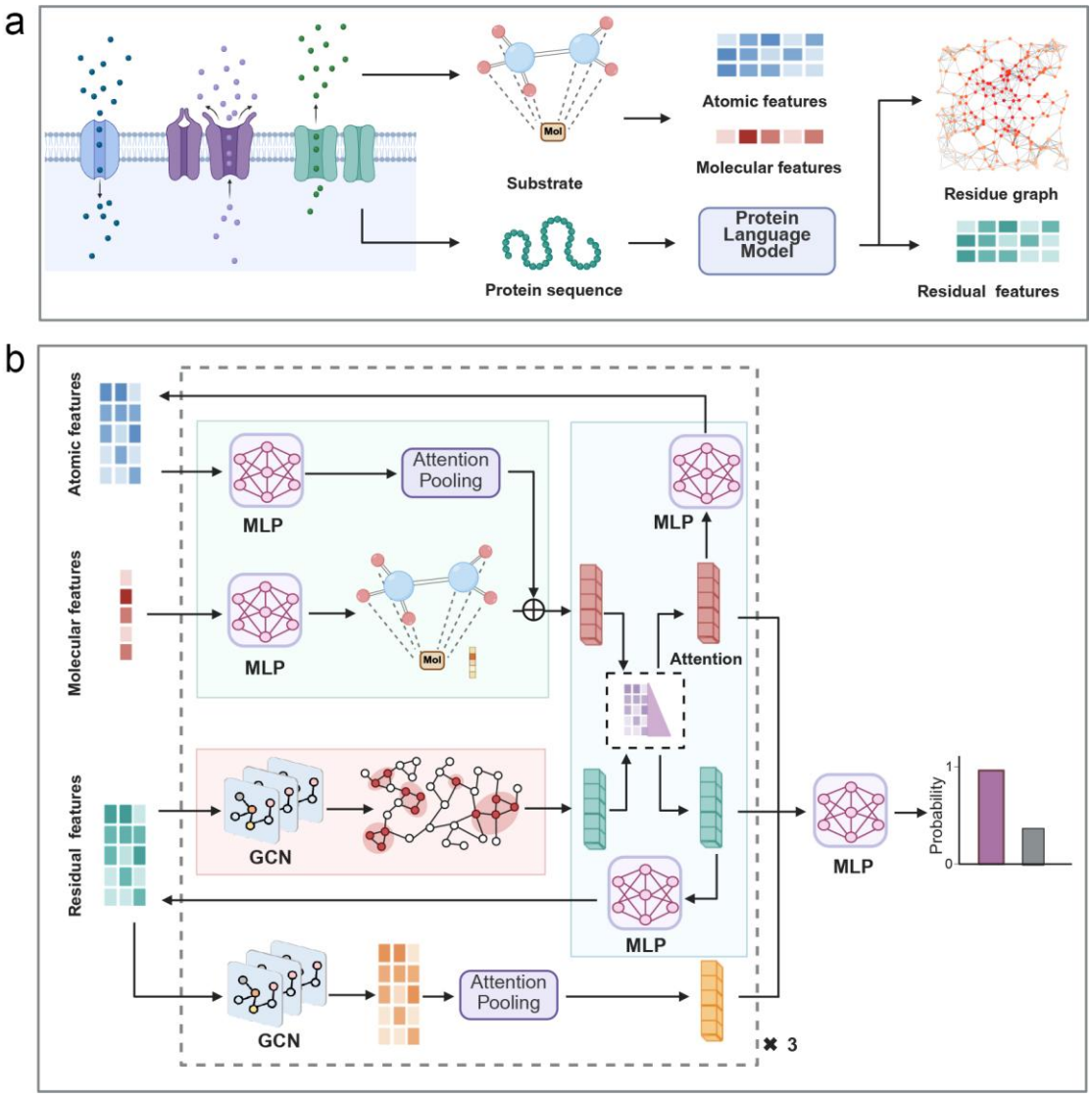


Fig. 1. Overview of the TraSPIN framework. a Model inputs. TraSPIN takes paired transporter amino acid sequences and chemical substrate representations as input. Transporter sequences are encoded using pretrained protein language models to obtain contextualized residue-level embeddings. Substrates are represented from their SMILES strings and corresponding molecular structures. **b** Architecture schematic. TraSPIN integrates three coordinated components. (i) Substrate encoding channel: chemical substrates are represented through dual-level encodings that combine atom-level molecular graph features with pretrained SMILES transformer–derived embeddings. (ii) Transporter encoding channel:

contextual residue embeddings are extracted from primary sequences using large-scale protein language models (ProtT5 and ESM2). Structural inductive bias is introduced by constructing residue graphs from ESM2-predicted contact maps, followed by graph convolution and spatial partitioning to decompose transporters into structurally coherent residue substructures. (iii) Interaction and prediction module: substrate embeddings and residue-partition representations are integrated through an attention-based interaction mechanism that explicitly models transporter–substrate interactions, weighting residue partitions according to their relevance for substrate recognition. The fused representations are projected to scalar probability scores corresponding to predicted substrate specificity.

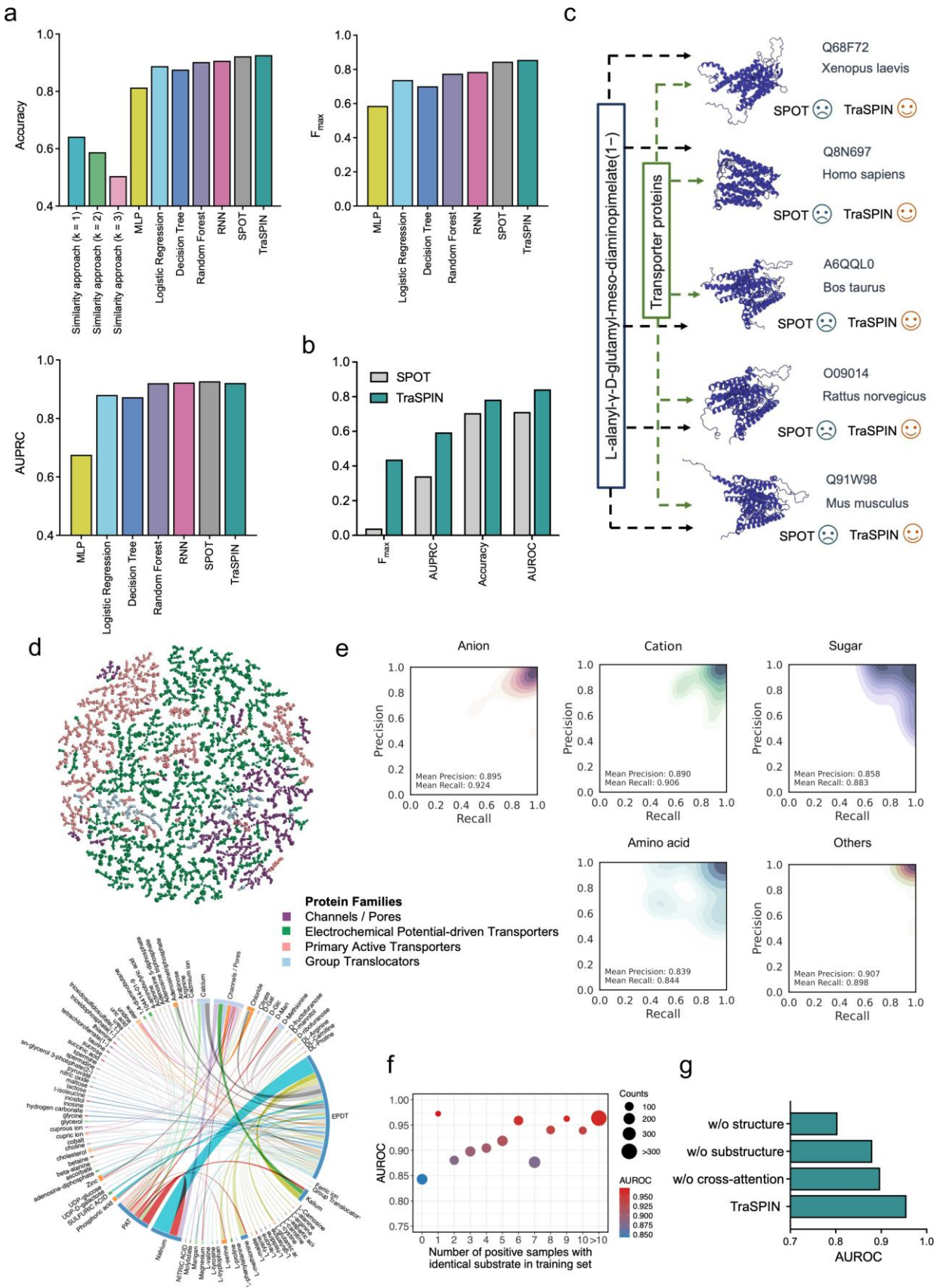
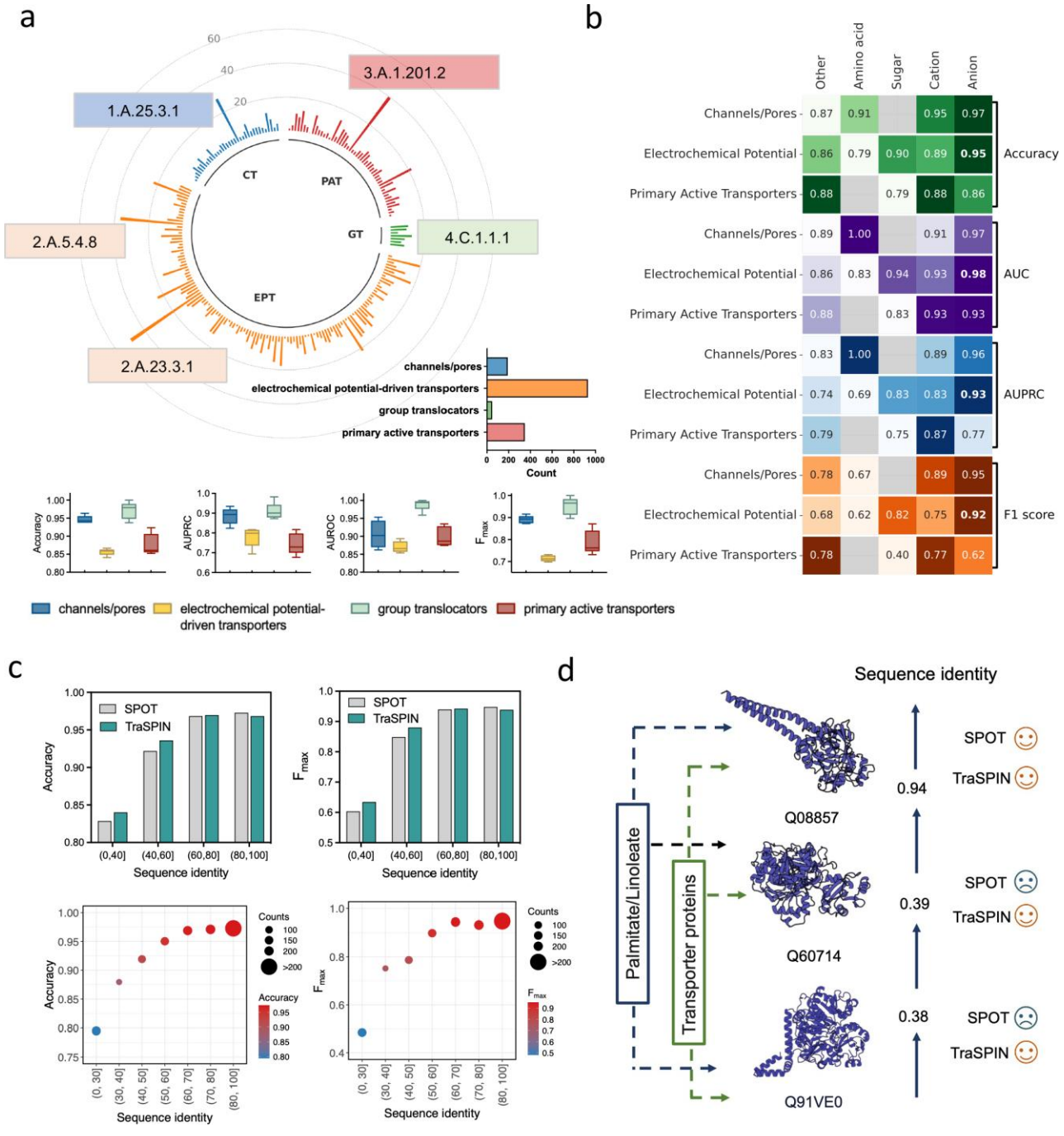

a
Accuracy
Similarity approach (k = 1)
Similarity approach (k = 2)
Similarity approach (k = 3)
MLP
Logistic Regression
Decision Tree
Random Forest
RNN
SPOT
TraSPIN
F_{max}
AUPRC
b
SPOT
TraSPIN
F_{max}
AUPRC
Accuracy
AUROC
c
L-alanyl-γ-D-glutamyl-meso-diaminopimelate(1−)
Transporter proteins
Q68F72
Xenopus laevis
Q8N697
Homo sapiens
A6QQL0
Bos taurus
O09014
Rattus norvegicus
Q91W98
Mus musculus
SPOT
TraSPIN
d
Protein Families
Channels / Pores
Electrochemical Potential-driven Transporters
Primary Active Transporters
Group Translocators
e
Anion
Cation
Sugar
Amino acid
Others
Precision
Recall
Mean Precision: 0.895
Mean Recall: 0.924
Mean Precision: 0.890
Mean Recall: 0.906
Mean Precision: 0.858
Mean Recall: 0.883
Mean Precision: 0.839
Mean Recall: 0.844
Mean Precision: 0.907
Mean Recall: 0.898
f
AUROC
Counts
100
200
300
>300
Number of positive samples with identical substrate in training set
g
w/o structure
w/o substructure
w/o cross-attention
TraSPIN
AUROC

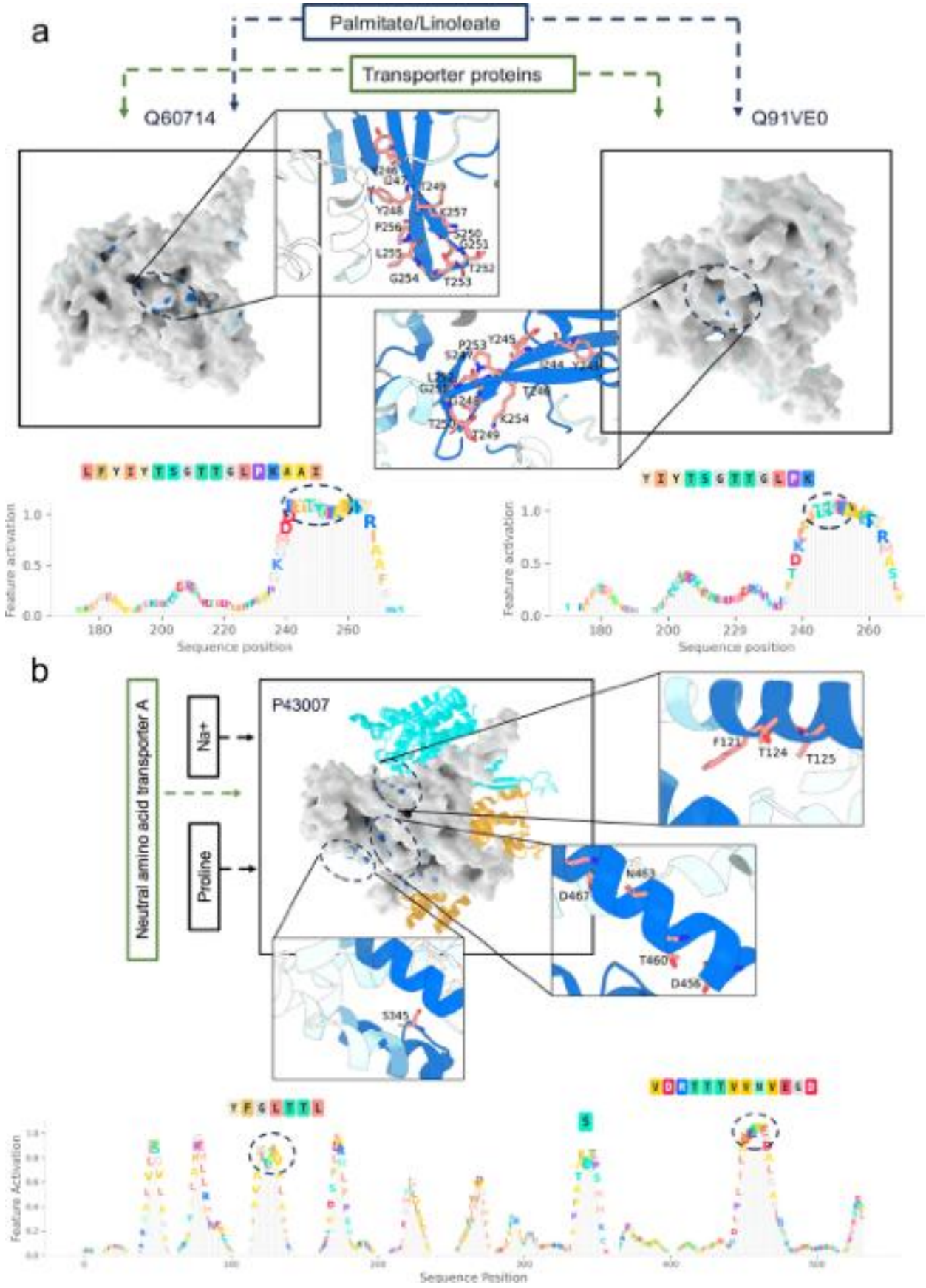
Fig. 2. Comparative evaluation of TraSPIN and baseline models. a Comparison of model performance between TraSPIN and baseline methods, evaluated using Accuracy, and AUPRC. Bars indicate different models, and colours distinguish individual methods. **b** Performance comparison under a stringent evaluation setting in which all substrates present in the test set are absent from the training set. The x-axis indicates different evaluation metrics (AUPRC, Accuracy and AUROC), and the y-axis denotes the corresponding performance values. Grey and teal bars represent the SPOT baseline and TraSPIN, respectively. **c** Case study for the substrate L-alanyl-γ-D-glutamyl-meso-diaminopimelate(1−), which does not occur in the training set. Predicted transporter proteins are shown for five organisms. Smiling and sad face icons indicate correct and incorrect predictions by TraSPIN and the SPOT baseline, respectively. **d** Distribution of transporter mechanistic classes in the dataset, including channels/pores, electrochemical potential–driven transporters, primary active transporters and group translocators. Colours denote different transporter families, and links indicate transporter–substrate associations. **e** In silico evaluation of TraSPIN on five representative substrate classes (anion, cation, sugar, amino acid and others). Precision–recall distributions are shown, where colour intensity reflects the density of predictions. **f** Relationship between predictive performance and substrate frequency in the training set. The x-axis indicates the number of positive training samples sharing identical substrates, and the y-axis denotes AUROC. Each point represents a substrate group, and point size reflects the number of samples in that group. **g** Ablation study evaluating the contributions of different model components. The x-axis denotes AUROC, and the y-axis lists model variants with specific components removed (without structure, without substructure, and without cross-attention).



651

652 **Fig. 3. Performance evaluation of TraSPIN for transporter-substrate specificity prediction and**
653 **homolog-independent assessment. a** Number and proportion of transporter proteins in the constructed
654 dataset across the four major classes defined by the Transporter Classification Database
655 (Channels/Pores, EPDT, Group Translocators, and PAT), along with TraSPIN's prediction performance

in each class. **b** Substrate-level prediction performance within three transporter categories (Channels/Pores, EPDT, and PAT) across five substrate types (Other, Amino Acid, Sugar, Cation, and Anion). Performance is evaluated using Accuracy, AUC, AUPRC, and F1 score. Group Translocators are excluded due to limited amount and substrate diversity. **c** Performance of TraSPIN stratified by sequence identity. Transporter proteins are grouped based on sequence identity to evaluate model generalization under decreasing homology levels, with TraSPIN compared directly to SPOT. **d** Case study of Palmitate- and Linoleate-associated transporter proteins, with each case representing a different level of sequence identity. TraSPIN demonstrates superior prediction performance under low-identity conditions compared with SPOT.


a
Palmitate/Linoleate
Transporter proteins
Q60714
Q91VE0
Feature activation
Sequence position
b
Neutral amino acid transporter A
Na+
Proline
P43007
Feature Activation
Sequence Position

665

Fig. 4. Structural interpretation of functional sites identified by TraSPIN. a Predicted substrate-binding regions of long-chain fatty acid transport proteins Q60714 and Q91VE0 in complex with palmitate or linoleate. Dashed circles indicate the putative binding pockets, with the TraSPIN-identified pocket highlighted in blue. Enlarged structural views display key residues mediating substrate interactions, as annotated in the UniProt database. Sequence feature activation profiles shown below denote amino acid positions with high contribution scores. Notably, residues identified by TraSPIN correspond closely to experimentally annotated binding-site residues. **b** Predicted binding site of the neutral amino acid transporter A (P43007) in complex with Na^+ and proline. The inferred substrate-binding pocket is marked by dashed circles and highlighted in blue, indicating residues with high predicted importance for substrate coordination. Enlarged structural views illustrate key interacting residues based on UniProt annotations. Sequence feature activation profiles indicate high-contribution amino acid positions identified by TraSPIN.

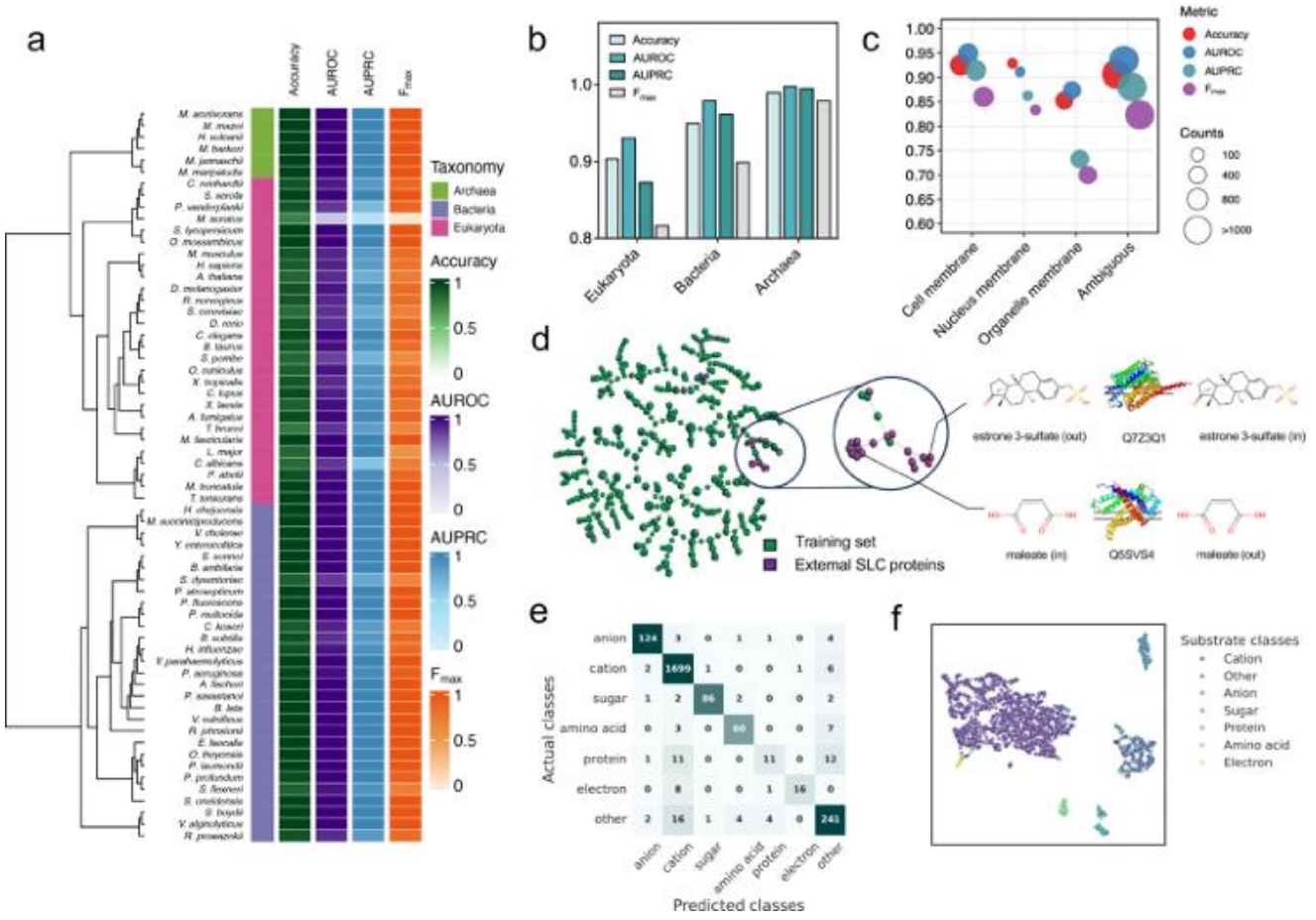


Fig. 5. Evaluation of TraSPIN across phylogenetic distributions, subcellular localizations, transporter families, and substrate class classification. a The phylogenetic tree (left) depicts evolutionary relationships among evaluated organisms, with a sidebar denoting taxonomic domain: Archaea, Eukaryota, and Bacteria. The adjacent heatmap displays four performance metrics: Accuracy, AUROC, AUPRC, and F_{max} . **b** Grouped bar chart comparing Accuracy, AUROC, AUPRC, and F_{max} across the three domains of life. **c** Performance metrics stratified by subcellular localization for eukaryotic transporters. **d** TMAP projection of protein embedding space, colour-coded by dataset origin: green, proteins present in the training set (including SLCs); purple, SLC proteins absent from the training set. Representative SLC structures and example substrate structures are shown alongside. **e** Confusion matrix for seven-class substrate classification; rows and columns represent actual and

690 predicted classes, respectively. **f** Two-dimensional UMAP projection of protein representations learned
691 by TraSPIN, coloured by substrate class.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplymentaryTraSPIN.docx