BREAKTHROUGH REPORT

Structure-guided discovery of protein functions in plants

Jiarong Chen[1,2,3,9], Yanlei Feng[2,9], Yuchan Zhang[1,2,9], Jucan Gao[4,9], Jinda Ou[5], Weiyin Wu[3], Can Li[1], Shuyan Song[1,2], Li Tai[1,2], Mahmudul Hasan Rifat[1], Delara Akhter[1,2,6], Jianping Hu[7,8], Peiqiang Feng[5]*, Xing-Xing Shen[1,3]*, Ronghui Pan[1,2]*

[1]State Key Laboratory of Rice Biology and Breeding, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China.

[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, Zhejiang, China.

[3]Key Laboratory of Biology of Crop Pathogens and Insects of Zhejiang Province, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, Zhejiang, China.

[4]College of Biotechnology and Bioengineering, Zhejiang University of Technology, Hangzhou, Zhejiang, China.

[5]CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, Shanghai, China.

[6]Department of Genetics and Plant Breeding, Sylhet Agricultural University, Sylhet, Bangladesh.

[7]Michigan State University-Department of Energy Plant Research Laboratory, Michigan State University, East Lansing, MI, USA.

[8]Department of Plant Biology, Michigan State University, East Lansing, MI, USA

[9]These authors contributed equally: Jiarong Chen, Yanlei Feng, Yuchan Zhang

*Correspondence. Email: panr@zju.edu.cn (R. Pan), xingxingshen@zju.edu.cn (X.-X. Shen), pqfeng@cemps.ac.cn (P. Feng)

Short title: Discovering Plant Protein Functions via Structure

The author(s) responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plcell/pages/General-Instructions) is: Ronghui Pan (panr@zju.edu.cn).

ABSTRACT

Protein structure serves as a critical bridge between sequence and functional annotation, particularly in establishing functional links among distantly homologous proteins with low sequence similarities. However, systematic protein structure-based functional annotations have been lacking in plants, where functions for a significant portion of the proteomes are still elusive. In this study, we leveraged protein structural data from 17 angiosperms to uncover previously

unannotated protein functions in plants. After structural clustering, we used the plant clusters to query the UniProtKB/Swiss-Prot database (the expertly curated component of UniProtKB), a repository of expertly curated and reliably annotated proteins, and identified structural matches for thousands of plant clusters that were undetectable by sequence-based BLAST searches. We further selected 120 clusters, which are highly reliable in structural quality and alignment and are well-conserved across plant species, and uncovered various protein functions that are potentially widely important in plants. Finally, we experimentally analyzed one plant cluster structurally resembling the yeast peroxisomal peroxin 8 (PEX8) protein and verified that plant PEX8-like proteins can functionally complement yeast *pex8* mutants. Our findings highlight the power of structural comparison in uncovering protein functions in plants.

Keywords: AlphaFold, Angiosperm, Plant genome and proteome, Protein structure, Gene function annotation, Foldseek, PEX8

INTRODUCTION

Elucidating the function of plant proteins is paramount to biological and agricultural research and has provided the knowledge base for crop improvement (Bailey-Serres et al. 2019). In the past decades, despite the vast successes in protein function annotation achieved through sequencing and molecular genetics, many plant proteins still lack functional annotations. By 2024, ~26% of the proteins in the model species Arabidopsis (*Arabidopsis thaliana*) remain functionally unresolved in The Arabidopsis Information Resource (TAIR) database (Reiser et al. 2024). Sequence similarity has been a valuable tool for predicting plant protein functions, particularly through transferring annotations from well-studied non-plant systems like mammals, yeasts, and bacteria. However, this approach often fails when plant proteins lack identifiable sequence homologs in other organisms, highlighting the need for alternative methods to uncover plant protein functions.

Protein structure is closely linked to function and often far more conserved than sequence, thus making structural data valuable for improving homology inference (Illergård et al. 2009; Vanni et al. 2022). The advent of AlphaFold has achieved unprecedented accuracy and simplicity for protein structure prediction (Senior et al. 2020; Jumper et al. 2021; Abramson et al. 2024). The AlphaFold Protein Structure Database (AFDB) provides extensive structural data (Varadi et al. 2024), while tools like FoldSeek enable efficient identification of remote homologs with divergent sequences (Van Kempen et al. 2024). Together these tools help researchers discover protein families and functions (Barrio-Hernandez et al. 2023; Durairaj et al. 2023). This structure-based

approach has been successfully employed in annotating unknow protein functions in specific domains of life, including viruses, Asgard archaea, pathogen effectors, and several other species (Derbyshire and Raffaele 2023; Ruperti et al. 2023; Seong and Krasileva 2023; Köstlbacher et al. 2024; Nomburg et al. 2024), yet its vast potential has not been exploited in plants. Until recently, Yu et al. explicitly proposed and emphasized the great potential of protein structure clustering in functional annotation of plant proteins, discovery of new functions, and even protein design (Yu et al. 2025).

To discover protein functions in plants, we conducted protein structure-based annotation in 17 representative angiosperm species. Angiosperms, commonly known as flowering plants, constitute the largest clade of the plant kingdom and ~90% of terrestrial plants (Christenhusz and Byng 2016), display remarkable morphological and functional diversities and establish themselves as essential components of global ecosystems (Benton et al. 2022). These angiosperm proteins were first clustered based on their structural similarities. Then the clusters were used for sequence and structural alignments against the UniProtKB/Swiss-Prot database, from which we discovered many clusters with significant structural homologies to well-annotated proteins despite lacking detectable sequence similarities. By focusing on conserved protein clusters present in more than 10 plant species, we identified 120 clusters with previously uncharacterized functions. Finally, we provided experimental validation for plant PEX8-like proteins, confirming their functional conservation with the yeast PEX8 counterparts. This study provides a valuable resource and toolbox for uncovering protein functions in plants.

RESULTS

Protein structural clustering

To perform structure-based protein function inference and annotation in plants, we selected 17 representative angiosperm species, including 2 basal species, 6 monocots, and 9 eudicots that consist of 5 rosids and 4 asterids as proxies for the two major eudicot branches (Figure 1A, Supplementary Data Set 1). The selected species all have high completion in genome sequencing, with BUSCO >86.5% (Figure 1A) and nearly fully structure-predicted proteomes in AFDB (Varadi et al. 2024), i.e., 83.7% for sesame and >97.5% for all the others. We obtained a dataset comprising

564,657 protein structures (Figure 1B), ~90% of which exhibit a predicted local distance difference test (pLDDT) score >0.5 (Figure 1C).

We then conducted structure-based clustering of all the 564,657 angiosperm proteins using the FoldSeek algorithm (Van Kempen et al. 2024), which had been used in several recent studies of protein structure clustering (Barrio-Hernandez et al. 2023; Köstlbacher et al. 2024; Nomburg et al. 2024). There are 177,510 clusters of angiosperm protein structures, among which 14.55% (25,825) are non-singleton (with more than one protein) (Figures 1B; Supplementary Data Set 2), a percentage that is similar to that in a previous study of AFDB protein structures using Foldseek (Barrio-Hernandez et al. 2023). To evaluate the quality of the non-singleton clusters, we used FoldSeek to select a representative structure from each cluster and align it to every other member of the cluster to calculate the LDDT and template modelling (TM) scores for each alignment. The median LDDT and TM scores are 0.84 and 0.74, respectively (Figure 1D). Further, all cluster members were predicted for Pfam domains followed by calculation of Pfam consistency for each cluster, which revealed that 67.6% of the non-singleton clusters exhibit 100% Pfam consistency (Figure 1D). Compared to the previous study using Foldseek cluster (Barrio-Hernandez et al. 2023), our LDDT, TM and Pfam consistency scores are slightly higher, supporting high structural homogeneity in the angiosperm non-singleton clusters in this study.

Structure-guided functional annotation

To identify structural homologs of the plant protein clusters from annotated proteins, we first selected a representative protein from each non-singleton cluster, using the Arabidopsis homologs when available or the protein with the highest pLDDT score. Then we employed FoldSeek to perform structural alignment between these representative proteins and the UniProtKB/Swiss-Prot database, which contains manually curated and reviewed UniProtKB entries with high-quality annotations (Bairoch 2000). Alignments for 15,566 of the plant protein clusters were found, and the best-hit match was kept as the final structural alignment for each cluster (Figure 2A).

To exclude functional annotations that could already be resolved through routine BLAST, we conducted parallel BLAST searches against Swiss-Prot and found 14,151 successful alignments (Figure 2A). Unlike our approach for structure matches, where only the best hits were retained, we

retained all detectable sequence matches identified by BLAST. Comparison of structural and sequence alignments identified 3,109 plant protein clusters whose best-hit structural alignments cannot be captured by BLAST (Figure 2A), representing proteins with conserved structural architecture despite extensive sequence deviations from their structural matches.

We further filtered the 3,109 clusters down to 1,292, which show high structural and alignment quality according to pLDDT ($\geqslant$0.7) and TM ($\geqslant$0.5) scores (Figure 2B; Supplementary Data Set 3). Next, we analyzed the number of plant species present in each cluster and focused on the 246 that contain more than 10 plant species (Figure 2C), which in theory carry functional importance broadly in plants.

To focus on the plant protein functions that have not yet been recognized to date, we manually checked whether our structure-based findings coincided at least partially with existing annotations in the full UniProtKB database, TAIR (Reiser et al. 2024), or even the predictions by EGGNOG-mapper (including PFAM domain information) (Cantalapiedra et al. 2021) and HHblits (a software for sensitive protein sequence searching based on the pairwise alignment of Hidden Markov Models or HMM) (Steinegger et al. 2019) (Figure 2D). Through this rigorous filtering process, we ultimately identified 120 conserved clusters, whose functions could not be inferred from sequence-based annotations (i.e., UniProtKB, TAIR, EGGNOG-mapper, or HHblits predictions) and required structural similarities to provide reliable functional insights (Figure 2D, Supplementary Data Set 4).

Discovery of protein functions in plants

Among the 120 selected clusters, 55 displayed the closest structural homology to proteins from other non-plant eukaryotes (animals, fungi, and protists), 32 showed prokaryotic affinities (predominantly bacterial with few archaeal matches), and 33 had their strongest structural matches within plants (Figure 3A, Supplementary Data Set 4). Most of these 33 clusters contain Arabidopsis members, which maintain high structural conservation with another protein in the same species despite significant sequence divergence. For example, AT3G20680 in cluster_1902 lacks identifiable BLAST matches in Arabidopsis but shares striking structural similarity with Arabidopsis LPA3 (low PSII accumulation 3, AT1G73060), a protein predicted to localize to chloroplasts and had been shown to be essential for photosystem II assembly (Järvi et al. 2015)

(Figure 3B). Similarly, AT3G60810 in cluster_24424 was predicted to localize to chloroplasts and structurally resembled AT4G24930, a protein annotated as the chloroplast thylakoid lumenal 17.9 kDa protein (Figure S1A). Moreover, AT1G23110 and AT1G70900 in cluster_1567 mirrored the structural fold of the stress-responsive ACER protein (alkaline ceramidase, AT4G22330) (Wu et al. 2015; Huang et al. 2022) (Figures S1B). These plant genes might have originated from ancient duplication events within plant lineages and subsequently underwent extensive sequence diversification while preserving core structural features.

We further predicted targeting signals for chloroplasts, mitochondria, and peroxisomes for the 120 conserved clusters (see METHODS). Chloroplast targeting signals were identified for 24 clusters, among which 7 were matched with plant proteins and 10 were matched with bacterial proteins in Swiss-Prot (Figure 3A). Considering the endosymbiotic origin of chloroplasts from cyanobacteria and the fact that cyanobacterial proteins are largely underrepresented in Swiss-Prot, we speculated that some of the 10 clusters matched with bacterial proteins would be structurally more like proteins from cyanobacteria than those from other bacteria, if they have an endosymbiotic origin. Thus, we specifically included the available AFDB structures of their cyanobacterial homologs in the structural alignment analysis. In addition to the 2 clusters aligned with cyanobacteria in the original analysis (Figures S2A-B), we identified 3 additional clusters with stronger structural similarities to cyanobacterial homologs than to other bacterial proteins (Figures 3C, S2C-D), suggesting their endosymbiotic origin. An example is cluster_18328, which structurally aligns with *B. subtilis* Lipase EstA, an alkaline-tolerant lipase (Nguyen et al. 2024) (Figure 3C).

Although most of these 120 clusters contain Arabidopsis members, there are 24 exceptions, among which cluster_21741 shows structural similarity to yeast vacuolar protein 8 (VAC8) (Figure 3D), an armadillo repeat protein mediating vacuole inheritance and cytoplasm-to-vacuole protein targeting (Wang et al. 1998). Interestingly, plant VAC8-like genes seem to have an unusual evolutionary history, as they are completely absent in green and streptophyte algae and appeared sporadically in bryophytes before undergoing significant expansion in seed plants, with four subgroups in gymnosperms and five in most angiosperms. Except cluster_21741, the other four angiosperm subgroup members are mostly excluded from AFDB due to their large size, i.e., above the 1,200 amino-acid upper limit of AFDB. Additionally, four angiosperm subgroups, including

cluster_21741, are specifically lost in the core Brassicales (Brassicaceae, Cleomaceae, and Capparaceae), contrasting with their broad conservation in other angiosperms (Figure S3).

We also categorized the 120 clusters using the Gene Ontology (GO) molecular function terms of their best-hit structural matches in the Swiss-Prot database, which had been converted into their corresponding plant-specific GO terms (see METHODS). Enriched GO terms include protein binding, hydrolase activity, transferase activity, nucleotide binding, and others (Figure 4A; Supplementary Data Set 4).

There are 41 clusters associated with GO terms of different enzymatic activities (Figure 4A; Supplementary Data Set 4). For example, cluster_15445 of "transferase activity" is matched with *M. vanbaalenii* PapA5 (phthiocerol/phthiodiolone dimycocerosyl transferase) (Figure 4B), which catalyzes the acylation of diol-containing polyketides for the biosynthesis of phenolic glycolipids (Chavadi et al. 2012).

There are 8 clusters sharing the GO term "nucleotide binding". An interesting example is cluster_12521, which is matched with *S. pombe* PCT1 (Pombe capping enzyme triphosphatase 1), an RNA 5′ triphosphatase (TPase) catalyzing the first step in mRNA capping (Pei et al. 2001) (Figure 4C). Yeast-type RNA triphosphatases are unrelated in mechanism and structure to mammalian-type RNA TPases, such as the human MCE1 (mRNA-capping enzyme 1), a bifunctional enzyme exhibiting RNA TPase activity in the N-terminus and mRNA guanylyltransferase (GTase) activity in the C-terminus (Ramanathan et al. 2016). The two known ARCP1 and ARCP2 (mRNA capping phosphatase 1 and 2) proteins in Arabidopsis are mammalian-type bifunctional MCEs (Ning et al. 2024), yet yeast-type TPases have never been identified in plants. Our findings indicate that angiosperms may have both types of TPases.

Six clusters share the GO term "transporter activity". Cluster_12986 was matched with *D. acidovorans* Omp32 (Outer membrane porin protein 32) (Figure 4D), a major outer membrane protein of the bacteria (Zachariae et al. 2006). Its best match in *E. coli* is OmpC, a general porin on the outer membrane of gram-negative bacteria that forms pores to allow passive diffusion of small molecules (Bölter and Soll 2001). The Arabidopsis protein in this cluster is AT1G11320, which has been shown to localize to the plastid envelope based on fluorescence microscopy and mass spectrometry (MS) evidence (Bouchnak et al. 2019; Trentmann et al. 2020), suggesting that this cluster may represent a porin on the chloroplast envelope. Interestingly, proteins in this cluster

do not align with any eukaryotic proteins and show higher structural and sequence similarities to the cyanobacterial than the *E. coli* counterpart (Figure 4D), suggesting that this protein might have been acquired with the chloroplast from its cyanobacterial ancestor.

The largest GO category is "protein binding", which contains 37 clusters. Cluster_10847, which was predicted to be peroxisomal (Supplementary Data Set 4), is matched with the yeast peroxin 8 (PEX8) protein (Figure 4E), a key peroxisome biogenesis factor whose orthologs had been elusive in plants. Plant PEX8-like proteins have high structural but low sequence similarities with PEX8 proteins from the yeasts *S. cerevisiae*, *Pichia pastoris*, *Hansenula polymorpha*, and *Yarrowia lipolytica* (Figure 4E) (Waterham et al. 1994; Liu et al. 1995; Rehling et al. 2000; Smith and Rachubinski 2001; Jansen et al. 2021). Our findings suggest that this essential peroxin also exists in plants but have evaded sequence-based searches.

Functional validation of plant PEX8

Since our main research interest is in plant organelle biology, we elucidated the function of the identified PEX8-like proteins to confirm the reliability of our structure-based methods in protein function annotation. To this end, we characterized their phylogenetics, subcellular localization, and Arabidopsis mutant phenotypes. While PEX8 can only be found in some species in green and streptophyte algae with limited sequence similarities, they are well conserved in protein sequence and the C-terminal peroxisome targeting signal type 1 (PTS1) peptide in embryophytes (land plants) (Figure 5A). Fusions of a fluorescent protein to the N-terminus of either the full length or the C-terminal 15-aa peptide of PEX8-like proteins from Arabidopsis, rice, and moss *Physcomitrium patens* all localized to peroxisomes, when stably or transiently expressed (Figures S4A-S4B). To analyze its physiological importance, we obtained two T-DNA mutants of Arabidopsis *PEX8* (*AtPEX8*). Neither allele could produce viable homozygotes, while seeds from heterozygous plants showed 1:2:0 segregation of wild type: heterozygous: homozygous (Figure S4C). In heterozygous *Atpex8* plants, ~21% of the seeds in the siliques appeared aborted (Figure S4D) and the aborted embryos were arrested at the heart stage (Figure S4E). The seed and embryo defects of *Atpex8* mutants are similar to those of the previously reported peroxin mutants such as *Atpex10* and *Atpex12* (Schumann et al. 2003; Fan et al. 2005) (Figures S4D-S4E), consistent with the notion that AtPEX8 might be a peroxin.

During the review process of this work, Buck et al reported the identification of AtPEX8 using HHpred (a server running HMM-HMM comparison and also integrating information from predicted secondary structure) (Steinegger et al. 2019) and similar observations regarding its conservation in plants, subcellular localization, and null mutant phenotypes (Buck et al. 2025). Both of our findings support the hypothesis that plant PEX8-like genes are functional equivalents of yeast PEX8s.

We further performed a yeast complementation experiment to validate the above hypothesis, by complementing the yeast mutant with plant PEX8 proteins. We generated a mutant of *Pichia pastoris* lacking *PEX8* and then introduced the different plant *PEX8* genes driven by the *P. pastoris TEF-1* promoter into the mutant (Figures 5B-5C). Disruption of PEX8 compromises yeast cell growth (Figure 5C), consistent with previous reports (Waterham et al. 1994; Liu et al. 1995; Rehling et al. 2000; Smith and Rachubinski 2001; Agne et al. 2003). *PEX8* genes from Arabidopsis, rice and Physcomitrium all rescued the growth defect of *P. pastoris pex8* (Figure 5C), providing strong evidence that these structurally PEX8-like plant proteins can indeed function as PEX8.

Yeast PEX8 proteins are known to be independent of PTS1 for localization and function (Waterham et al. 1994; Liu et al. 1995; Rehling et al. 2000; Smith and Rachubinski 2001; Wang et al. 2004; Zhang et al. 2006; Ma et al. 2009; Deckers et al. 2010; Jansen et al. 2021). Consistently, we confirmed that plant PEX8s can target peroxisomes when PTS1 is blocked or deleted in either plant (Figures S5A-S5B) or yeast cells (Figure S5C). Moreover, when we re-introduced *AtPEX8* driven by its native promoter into heterozygous *Atpex8* plants, the full-length genomic or cDNA sequence, as well as the truncated PTS1-less version, fully complemented the mutant defects (Figures S5D-S5G). Thus, plant and yeast PEX8 proteins share the unique PTS1-independent targeting mechanism, further supporting that they derive from the same ancestral gene despite deviant sequences.

DISCUSSION

Our study highlights the power of structural comparison in uncovering novel protein functions in plants. Here, we employed a structure-based approach that goes beyond the traditional sequence alignment to systematically annotate the function of plant proteins. We identified 1,292

angiosperm protein clusters with significant structural matches in the Swiss-Prot database that were undetectable by BLAST searches due to low sequence similarities. Focusing on the 120 protein clusters that are highly reliable and widely present across plants, we demonstrated many cases where structural similarities revealed potential functional relationships despite low sequence identities. One of such examples is PEX8, a missing plant peroxin we experimentally validated in this study. As proof of concept, we only showcased a few selected examples in this report. For convenient access to our full dataset of non-singleton clusters and their structural and sequence alignments, an online database was generated to allow gene ID-based data retrieval (currently at https://ai-biolab.cn).

Protein structures not only provide functional insights but also reveal intriguing evolutionary patterns. For example, we unveiled proteins with high structural conservation but sequence divergence within the same species such as *Arabidopsis thaliana* (Figures 3B, 3C, & S1) and lineage-specific gene losses such as the absence of VAC8-like genes from core Brassicales (Figure S3). Another example is the discovery of potential ancient origins of some proteins, such as PEX8 that was previously thought to be lineage specific but found in our study to be widely present in eukaryotes and possibly dating back to the most ancient common ancestor of eukaryotes (Figure 5). Structural data thus complements traditional evolutionary studies by uncovering relationships that sequence-based methods alone might miss. However, current limitations remain in conducting reliable phylogenetic analyses based solely on protein structures, particularly for proteins with highly divergent sequences. Our attempt to reconstruct the evolutionary history of PEX8 across eukaryotes, for instance, was complicated by extreme sequence divergence that made it difficult to confidently resolve the phylogenetic relationships. The structure-based tree-building methods also require further development to improve robustness. Establishing reliable structure-based phylogenetic approaches will be crucial for advancing our understanding of protein evolution and function, especially for ancient or rapidly evolving gene families where sequence conservation is low but structural features are preserved.

This study has several limitations that should be considered in future research. First, the relatively small sampling size of angiosperm species may have led to the omission of certain lineage-specific genes, particularly those restricted to specific families (e.g., Poaceae) or orders. This limited taxonomic coverage likely contributed to the observed abundance of singleton clusters

and small-sized clusters containing few proteins. Future studies employing similar approaches should incorporate broader species representation, especially within key lineages, such as monocots and Poaceae, to better understand patterns of gene family diversification in plants. Second, technical constraints of the AlphaFold DB resulted in the exclusion of many large proteins (>1,200 amino acids), potentially introducing bias to our cluster analysis. For example, many members of the VAC8-like protein family were missing from the database. Moreover, certain plant genes currently have ambiguous functional annotations, particularly within large superfamilies. While these genes may be annotated based on broad functional characteristics of their respective superfamilies, they often lack precise, gene-specific functional characterizations. Although such genes were not classified as functionally unknown in this study, they warrant more detailed future investigations to elucidate their specific roles.

METHODS

Acquisition and quality assessment of protein structural data and database construction

Protein structures for the 17 angiosperms (Supplementary Data Set 1) were acquired from the the AlphaFold DB (AFDB) (Varadi et al. 2024), and sequence data were obtained from UniProt Proteome (https://www.uniprot.org/proteomes). Completeness of the proteome data was estimated by BUSCO (Manni et al. 2021) v5.6.1 using the lineage dataset "embryophyte_odb10". Phylogeny of the 17 species was reconstructed using the 1,614 BUSCO genes. Sequences were aligned using MAFFT (Katoh and Standley 2013) v7.525 with the "auto" mode and trimmed using TrimAl (Capella-Gutiérrez et al. 2009) v1.4.rev15. Then, a maximum likelihood (ML) tree was constructed by IQ-TREE (Minh et al. 2020) v2.3.5, with model "LG+G4" and 1,000 ultrafast bootstraps (Hoang et al. 2018). The iTOL website (Letunic and Bork 2024) was used for tree visualization.

Structural data for the Swiss-Prot database (Supplementary Data Set 1) were downloaded from the AFDB. Sequence data for proteins in the Swiss-Prot database were obtained from the UniProt database. Proteins with unclear entries in the annotation, such as those with terms like "uncharacterized", "unknown", "hypothetical", "domain of unknown function (DUF)", or "putative", were excluded. Based on Swiss-Prot, the structural database was constructed using

FoldSeek v8.ef4e90 (Van Kempen et al. 2024) and sequence database was built using BLAST v2.12.0 (Camacho et al. 2009).

Clustering of protein structure

The angiosperm protein structures were initially clustered using FoldSeek "easy-cluster". With the threshold of ≥30% sequence identity and a target coverage rate of ≥50% in the 3D-interaction sequences, we prioritized structural similarities over sequence similarities within the same cluster. The specific parameters used were "--cov-mode 0 --align-type 2 –min-seqid 0.3 -c 0.5".

Calculation of the local distance difference test (LDDT), template modeling (TM) score, predicted local distance difference test (pLDDT), and Pfam consistency

To evaluate structural similarities, the average LDDT and TM scores were calculated for each cluster. For each cluster, the representative structure was matched to the cluster members using "structurealign -e INF -a module" in FoldSeek. The representative for each cluster was selected by FoldSeek during clustering. The alignment LDDT and TM scores were obtained using "format-output lddt,alntmscore". The pLDDT values were extracted and calculated from a PDB file obtained from AFDB, using the PDBParser in BioPython v1.84 (Cock et al. 2009). All proteins within each cluster underwent Pfam prediction using InterProScan (Jones et al. 2014) v5.66-98.0. Only clusters containing at least two annotated sequences were chosen for calculating Pfam consistency, based on a previously published method (Barrio-Hernandez et al. 2023).

Search for structural and sequence alignments in the Swiss-Prot dataset

For each non-singleton cluster, we selected the *Arabidopsis thaliana* protein as the representative if available. Otherwise, the protein structure with the highest confidence was chosen.

Using FoldSeek for structural annotation with Swiss-Prot structure database, an e-value <0.01 was required and parameters "--max-seqs 50,000 -s 9.5 -e 0.01 –alignment-type 2 –cov-mode 0" were used. For sequence searches, e-value <0.01 was required in Blastp command as well, ensuring that the alignment reflects structural but not sequence homology. In the filtering process, TM-align v20190425 (Zhang 2005) was used to further compare the alignment results using the

parameter "-a". The TM-score between representative of cluster and the best hit was calculated to determine whether it was ≥0.5 to ensure significant structural similarities.

Calculating global sequence similarities using the Needleman-Wunsch algorithm

Global sequence similarities for all sequences presented in the figures were calculated using the Needleman-Wunsch algorithm and via the EMBOSS package (v.6.6.0.0). This approach ensures a comprehensive and optimal global alignment across the entire length of both sequences.

Gene annotation

Gene sequences were annotated using EggNOG-Mapper v2.1.9 with the v5 database (Huerta-Cepas et al. 2019; Cantalapiedra et al. 2021) under default parameters. For profile-to-profile annotation, we employed HHblits from the HH-suite 3.3.0 (Steinegger et al. 2019) against the "PDB_mmCIF" database using the parameters "-n 2 -d pdb70 -cpu 2 -E 0.001".

GO term counts

We performed this analysis using the top matches in Swiss-Prot database for the 120 high-confidence cluster representative sequences (Supplementary Data Set 4). Their GO terms on molecular functions were obtained from QuickGO (https://www.ebi.ac.uk/QuickGO) on EMBL's European Bioinformatics Institute website. Using the Python package GOATOOLS v9 (Klopfenstein et al. 2018), all the GO terms were converted into plant-specific GO terms. The list of plant GO terms was downloaded from GOslim (https://www.ebi.ac.uk/QuickGO/slimming) by selecting "predefined GO slims" as "goslim_plant" on the Explore Biology page in QuickGO. The conversion process itself was implemented using the Python script "slim_GO_by_list.py" which is available on GitHub (https://github.com/Chenjiaron/Structure-guided-plant-protein-discovery). Finally, GO terms were counted by number and plotted in R.

Prediction of protein subcellular localization

Prediction of the mitochondrial and chloroplast transit peptides was performed by TargetP v2.0 (Emanuelsson et al. 2000) and based on the experimental localization information on SUBA5 (Hooper et al. 2022). Peroxisome targeting signal type 1 (PTS1) was initially identified by examining the C-terminal tripeptide (Deng et al. 2022) and further confirmed by PredPlantPTS1 (Reumann et al. 2012).

In vivo protein targeting analysis

For protein expression in tobacco and *Arabidopsis* leaves, cDNA sequences were PCR-amplified and cloned into a pCambia1300-YFP vector or a pCambia1300-mVenus vector, using the ClonExpress II One Step Cloning Kit (Vazyme, Nanjing). *Agrobacterium tumefaciens* strain GV3101 (pMP90) was transformed with the binary construct and selected on kanamycin. To detect stable subcellular localization in *Arabidopsis* leaves, *Arabidopsis* plants were transformed by Agrobacterium containing the constructs using floral dip and selected on 40 mg/L hygromycin. To detect the subcellular localization of proteins transiently expressed in tobacco leaves, *Agrobacterium* was infiltrated into tobacco leaves, and fluorescence signals were observed 2 d later.

For peroxisomal localization studies in *Saccharomyces cerevisiae*, the BY4741 laboratory strain was used. Vector pGADT7 expressing YFP fusion proteins and pYES2 expressing the peroxisomal mox3Cerulean fluorescent protein (PEX14p-Cerulean) were co-transferred into BY4741. Both fluorescent proteins were expressed under the control of the constitutive ADH1 promoter. The nuclear localization signal peptide was deleted from pGADT7. Yeast cells harboring both constructs were selected on SC agar medium with 2 % (w/v) glucose in the absence of Leu (pGADT7) and uracil (pYES2). Cells in growing colonies were used for fluorescence signal detection.

Confocal microscopy of plant leaf cells and yeasts was performed using a Fluoview FV3000 confocal laser-scanning microscope (Olympus, Tokyo, Japan). YFP and mVenus were excited with 514 nm lasers and detected at 530–630 nm, while CFP and mox3Cerulean signals were excited with 445 nm lasers and detected at 460–500 nm.

For subcellular localization in Arabidopsis protoplast, PEX8 coding sequences from various species were amplified and cloned into the p2GYW7 vector, which carries the mEGFP coding sequence. The mRFP CDS was amplified by primers with an added C-terminal SKL tripeptide in one of them, and ligated into a modified HBT95 vector to generate the RFP-SKL construct. All constructs were verified by DNA sequencing. Arabidopsis protoplast isolation and transformation assays were carried out as previously described (Shen et al. 2014). Typically, 0.1 ml aliquots of protoplasts were transfected with 10μg of plasmid DNA in each sample. After overnight incubation, protoplasts were examined by a Leica SR5 laser scanning confocal microscope (Germany).

Excitation (ex) and emission (em) parameters for the detection of the different fluorophores are as follows (ex/em): 488 nm/510–550 nm for GFP, 561 nm/590–631 nm for RFP, and 633 nm/680–740 nm for chlorophyll. The pinhole was set at 1 Airy unit. All images were captured using the same settings for direct comparisons.

Generation of *P. pastoris pex8* mutants and complementation analysis

Plasmids (Supplementary Data Set 5) were constructed using Gibson Assembly. Oligonucleotides (TSINGKE Biological Technology, Hangzhou, China) used for gRNA construction, DNA amplification, plasmid assembly, and diagnostic PCR verification are listed in Supplementary Data Set 5. The gRNA sequences were designed using the Benchling CRISPR tool (https://benchling.com/crispr/), and the corresponding gRNA plasmid HZP-sgRNA-IntPpaPEX8 was constructed using a previously reported toolkit (Gao et al. 2022). PEX8 CDS from *A. thaliana*, *P. patens*, and *Rice* was respectively inserted into the helper plasmids to obtain Int33-AtPEX8, Int33-PpPEX8 and Int33-OsPEX8. After plasmid construction and sequencing, the DNA segment was amplified to generate donor DNA. The CRISPR-Cas9 system was employed to integrate genes into *P. pastoris,* using a previous method for transformation (Gao et al. 2022).

Observation of Arabidopsis seed development

Four heterozygous plants from each *pex8* mutant allele and four wild-type self-fertilized plants were analyzed. Developing seeds from 10 siliques from each plant were removed and scored for abnormalities in appearance. More than 200 seeds per plant were scored, and the average frequency of abnormal seeds per heterozygous plant was calculated.

Complementation of Arabidopsis *pex8*

The *AtPEX8* genomic sequence or cDNA sequence with or without the C-terminal PTS1, all driven by the 1.5- kb promoter fragment of *AtPEX8,* was respectively inserted into the binary vector pCAMBIA 1300 digested with *Hind*III and *Xba*I. *Agrobacterium tumefaciens* strain GV3101 (pMP90) was transformed with the resulted constructs and selected on kanamycin. *Arabidopsis PEX8*/*pex8* heterozygous plants were transformed with the constructs via floral dipping. Seeds of the transformants were selected on MS medium with 40 mg/L hygromycin to select mutants containing $PEX8_{pro}$:: $PEX8_{genomic}$, $PEX8_{pro}$:: $PEX8_{cDNA}$ and $PEX8_{pro}$:: $PEX8\Delta PTS1$.

Plant materials and growth conditions

The T-DNA insertion mutants of *Atpex8*-1 (SALK_129140) and *Atpex8*-2 (SALK_032940) were obtained from the Arabidopsis Biological Recourse Center (ABRC; Columbus, OH, USA). Homozygous lines were identified by PCR genotyping, using genomic DNA from seedlings and gene-specific primers (Supplementary Data Set 5). Arabidopsis seeds were plated on 1/2 MS medium and stored at 4°C for 2 d. Then, plants were grown on 1/2 MS medium for 2 weeks followed by transfer to the soil, at 23 °C and under a 16/8h light/dark photoperiod unless otherwise specified in the text.

Embryo dissection and microscopy observations

Seed morphology was observed under a microscope (Nikon AZ100) after siliques were gently dissected with tweezers. Pictures were taken by a Nikon DigiSight DS-Ri1 camera. To analyze embryo development, developing seeds from the same siliques were placed onto a glass side in a drop of transparent liquid (30% glycerol, 2.5 g/ml hydrated trichloroacetaldehyde). Embryonic development was observed for normal and abnormal seeds from the same silique under a Nomarski contrast microscope (Nikon Eclipse Ni) and recorded with a Nikon DS-Ri2 CCD camera.

Phylogenetic analysis of VAC8 gene family

Protein sequences were aligned using MAFFT in "auto" mode and subsequently trimmed with TrimAl. Phylogenetic trees were then inferred with IQ-TREE using the "LG+F+G4" model and 1,000 ultrafast bootstrap replicates. Final tree visualization was conducted in iTOL. The multiple sequence alignment used to generate this phylogeny, along with the resulting tree file, are provided as Supplementary Data File 1 and Supplementary Data File 2, respectively.

Quantitative analysis and data visualization

Quantifications and statistical analyses were performed in R and plotted using the ggplot2 (Wickham 2016) v3.5.1 package. Protein structures were generated by PyMOL (https://pymol.org) for visualization.

Accession numbers

Sequence/Structure data from this article can be found in UniProt/Alphafold DB under accession numbers that are listed in Supplementary Data Set 4.

Data availability

All data are available in the main text or the supplementary materials. Information of the non-singleton clusters is provided on the website https://ai-biolab.cn and can be retrieved by gene IDs. The code used in this study has been uploaded to GitHub (https://github.com/Chenjiaron/Structure-guided-plant-protein-discovery).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32470287, 32500235 & 32200231), the Zhejiang Provincial Natural Science Foundation of China (R26C130007, QN26C020005, & LZ23C020002), the Natural Science Foundation of Hangzhou (2025SZRJJ0918 & 2024SZRYBC130003), the National Key Research and Development Program (2022YFD1401600 & 2024YFD1200401), the Fundamental Research Funds for the Central Universities (226-2025-00163), and fellowship of the China Postdoctoral Science Foundation (2024M762901, 2025T180747, 2025M782775, & 2025M772587).

Author contributions

R.P., X.S., and P.F. conceptualized and supervised the study. R.P., X.S., J.C., Y.F., J.G., Y.Z., and P.F. designed the experiments, analyzed the data, and prepared the figures. J.C., Y.F., D.A., and W.W., conducted bioinformatic analysis. Y.Z., C.L., S.S., L.T., J.G., M.H.R. and J.O. performed the molecular, genetic, protein localization and physiological experiments. R.P., X.S., J.C., Y.F., and J.H. participated in data interpretation and cowrote the manuscript.

Competing interests

Authors declare no competing interests.

REFERENCES

Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024:**630**(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w

Agne B, Meindl NM, Niederhoff K, Einwächter H, Rehling P, Sickmann A, Meyer HE, Girzalsky W, and Kunau W-H. Pex8p: An Intraperoxisomal Organizer of the Peroxisomal Import Machinery. Mol Cell. 2003:**11**(3):635–646. https://doi.org/10.1016/S1097-2765(03)00062-5

Bailey-Serres J, Parker JE, Ainsworth EA, Oldroyd GED, and Schroeder JI. Genetic strategies for improving crop yields. Nature. 2019:**575**(7781):109–118. https://doi.org/10.1038/s41586-019-1679-0

Bairoch A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000:**28**(1):45–48. https://doi.org/10.1093/nar/28.1.45

Barrio-Hernandez I, Yeo J, Jänes J, Mirdita M, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, and Steinegger M. Clustering predicted structures at the scale of the known protein universe. Nature. 2023:**622**(7983):637–645. https://doi.org/10.1038/s41586-023-06510-w

Benton MJ, Wilf P, and Sauquet H. The Angiosperm Terrestrial Revolution and the origins of modern biodiversity. New Phytol. 2022:**233**(5):2017–2035. https://doi.org/10.1111/nph.17822

Bölter B and Soll J. Ion channels in the outer membranes of chloroplasts and mitochondria: Open doors or regulated gates? EMBO J. 2001:**20**(5):935–940. https://doi.org/10.1093/emboj/20.5.935

Bouchnak I, Brugière S, Moyet L, Le Gall S, Salvi D, Kuntz M, Tardif M, and Rolland N. Unraveling Hidden Components of the Chloroplast Envelope Proteome: Opportunities and Limits of Better MS Sensitivity. Mol Cell Proteomics. 2019:**18**(7):1285–1306. https://doi.org/10.1074/mcp.RA118.000988

Buck GC, Weeks AD, Ordner NE, and Bartel B. Identifying and characterizing a missing peroxin-PEX8-in Arabidopsis thaliana. Plant Cell. 2025:koaf166. https://doi.org/10.1093/plcell/koaf166

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, and Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009:**10**(1):421. https://doi.org/10.1186/1471-2105-10-421

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, and Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol. 2021:**38**(12):5825–5829. https://doi.org/10.1093/molbev/msab293

Capella-Gutiérrez S, Silla-Martínez JM, and Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009:**25**(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chavadi SS, Onwueme KC, Edupuganti UR, Jerome J, Chatterjee D, Soll CE, and Quadri LEN. The mycobacterial acyltransferase PapA5 is required for biosynthesis of cell wall-associated phenolic glycolipids. Microbiology. 2012:**158**(5):1379–1387. https://doi.org/10.1099/mic.0.057869-0

Christenhusz MJM and Byng JW. The number of known plants species in the world and its annual

increase. Phytotaxa. 2016:**261**(3):201. https://doi.org/10.11646/phytotaxa.261.3.1

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009:**25**(11):1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Deckers M, Emmrich K, Girzalsky W, Awa WL, Kunau W-H, and Erdmann R. Targeting of Pex8p to the peroxisomal importomer. Eur J Cell Biol. 2010:**89**(12):924–931. https://doi.org/10.1016/j.ejcb.2010.06.019

Deng Q, Li H, Feng Y, Xu R, Li W, Zhu R, Akhter D, Shen X, Hu J, Jiang H, et al. Defining upstream enhancing and inhibiting sequence patterns for plant peroxisome targeting signal type 1 using large-scale *in silico* and *in vivo* analyses. Plant J. 2022:**111**(2):567–582. https://doi.org/10.1111/tpj.15840

Derbyshire MC and Raffaele S. Surface frustration re-patterning underlies the structural landscape and evolvability of fungal orphan candidate effectors. Nat Commun. 2023:**14**(1):5244. https://doi.org/10.1038/s41467-023-40949-9

Durairaj J, Waterhouse AM, Mets T, Brodiazhenko T, Abdullah M, Studer G, Tauriello G, Akdel M, Andreeva A, Bateman A, et al. Uncovering new families and folds in the natural protein universe. Nature. 2023:**622**(7983):646–653. https://doi.org/10.1038/s41586-023-06622-3

Emanuelsson O, Nielsen H, Brunak S, and Von Heijne G. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. J Mol Biol. 2000:**300**(4):1005–1016. https://doi.org/10.1006/jmbi.2000.3903

Fan J, Quan S, Orth T, Awai C, Chory J, and Hu J. The arabidopsis PEX12 gene is required for peroxisome biogenesis and is essential for development. Plant Physiol. 2005:**139**(1):231–239. https://doi.org/10.1104/pp.105.066811

Gao J, Xu J, Zuo Y, Ye C, Jiang L, Feng L, Huang L, Xu Z, and Lian J. Synthetic Biology Toolkit for Marker-Less Integration of Multigene Pathways into Pichia pastoris via CRISPR/Cas9. ACS Synth Biol. 2022:**11**(2):623–633. https://doi.org/10.1021/acssynbio.1c00307

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, and Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol. 2018:**35**(2):518–522. https://doi.org/10.1093/molbev/msx281

Hooper C, Millar H, Black K, Castleden I, Aryamanesh N, and Grasso S. Subcellular Localisation database for Arabidopsis proteins version 5. 2022. https://doi.org/10.26182/8DHT-4017

Huang L-Q, Li P-P, Yin J, Li Y-K, Chen D-K, Bao H-N, Fan R-Y, Liu H-Z, and Yao N. Arabidopsis alkaline ceramidase ACER functions in defense against insect herbivory. J Exp Bot. 2022:**73**(14):4954–4967. https://doi.org/10.1093/jxb/erac166

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019:**47**(D1):D309–D314. https://doi.org/10.1093/nar/gky1085

Illergård K, Ardell DH, and Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009:**77**(3):499–508. https://doi.org/10.1002/prot.22458

Jansen RLM, Santana-Molina C, van den Noort M, Devos DP, and van der Klei IJ. Comparative Genomics of Peroxisome Biogenesis Proteins: Making Sense of the PEX Proteins. Front Cell Dev Biol. 2021:**9**(May):1–22. https://doi.org/10.3389/fcell.2021.654163

Järvi S, Suorsa M, and Aro E-M. Photosystem II repair in plant chloroplasts — Regulation, assisting proteins and shared components with photosystem II biogenesis. Biochim Biophys Acta BBA - Bioenerg. 2015:**1847**(9):900–909. https://doi.org/10.1016/j.bbabio.2015.01.006

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014:**30**(9):1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021:**596**(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

Katoh K and Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013:**30**(4):772–780. https://doi.org/10.1093/molbev/mst010

Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, Mungall CJ, Yunes JM, Botvinnik O, Weigel M, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep. 2018:**8**(1):10872. https://doi.org/10.1038/s41598-018-28948-z

Köstlbacher S, Van Hooff JJE, Panagiotou K, Tamarit D, De Anda V, Appler KE, Baker BJ, and Ettema TJG. Structure-based inference of eukaryotic complexity in Asgard archaea. 2024. https://doi.org/10.1101/2024.07.03.601958

Letunic I and Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 2024:**52**(W1):W78–W82. https://doi.org/10.1093/nar/gkae268

Liu H, Tan X, Russell KA, Veenhuis M, and Cregg JM. PER3, a Gene Required for Peroxisome Biogenesis in Pichia pastoris, Encodes a Peroxisomal Membrane Protein Involved in Protein Import. J Biol Chem. 1995:**270**(18):10940–10951. https://doi.org/10.1074/jbc.270.18.10940

Ma C, Schumann U, Rayapuram N, and Subramani S. The Peroxisomal Matrix Import of Pex8p Requires Only PTS Receptors and Pex14p. Mol Biol Cell. 2009:**20**(16):3680–3689.

https://doi.org/10.1091/mbc.e09-01-0037

Manni M, Berkeley MR, Seppey M, Simão FA, and Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021:**38**(10):4647–4654. https://doi.org/10.1093/molbev/msab199

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, and Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020:**37**(5):1530–1534. https://doi.org/10.1093/molbev/msaa015

Nguyen VDH, Huynh TNP, Nguyen TTT, Ho HH, Trinh LTP, and Nguyen AQ. Expression and characterization of a lipase EstA from *Bacillus subtilis* KM-BS for application in bio-hydrolysis of waste cooking oil. Protein Expr Purif. 2024:**215**:106419. https://doi.org/10.1016/j.pep.2023.106419

Ning K, Li X, Yan J, Liu J, Gao Z, Tang W, and Sun Y. Heat Stress Inhibits Pollen Development by Degrading mRNA Capping Enzyme ARCP1 and ARCP2. Plant Cell Environ. 2024. https://doi.org/10.1111/pce.15178

Nomburg J, Doherty EE, Price N, Bellieny-Rabelo D, Zhu YK, and Doudna JA. Birth of protein folds and functions in the virome. Nature. 2024:**633**(8030):710–717. https://doi.org/10.1038/s41586-024-07809-y

Pei Y, Schwer B, Hausmann S, and Shuman S. Characterization of Schizosaccharomyces pombe RNA triphosphatase. Nucleic Acids Res. 2001:**29**(2):387–396. https://doi.org/10.1093/nar/29.2.387

Ramanathan A, Robb GB, and Chan S-H. mRNA capping: biological functions and applications. Nucleic Acids Res. 2016:**44**(16):7511–7526. https://doi.org/10.1093/nar/gkw551

Rehling P, Skaletz-Rorowski A, Girzalsky W, Voorn-Brouwer T, Franse MM, Distel B, Veenhuis M, Kunau W-H, and Erdmann R. Pex8p, an Intraperoxisomal Peroxin of Saccharomyces cerevisiae Required for Protein Transport into Peroxisomes Binds the PTS1 Receptor Pex5p. J Biol Chem. 2000:**275**(5):3593–3602. https://doi.org/10.1074/jbc.275.5.3593

Reiser L, Bakker E, Subramaniam S, Chen X, Sawant S, Khosa K, Prithvi T, and Berardini TZ. The Arabidopsis Information Resource in 2024. GENETICS. 2024:**227**(1):iyae027. https://doi.org/10.1093/genetics/iyae027

Reumann S, Buchwald D, and Lingner T. PredPlantPTS1: A Web Server for the Prediction of Plant Peroxisomal Proteins. Front Plant Sci. 2012:**3**. https://doi.org/10.3389/fpls.2012.00194

Ruperti F, Papadopoulos N, Musser JM, Mirdita M, Steinegger M, and Arendt D. Cross-phyla protein annotation by structural prediction and alignment. Genome Biol. 2023:**24**(1):113. https://doi.org/10.1186/s13059-023-02942-9

Schumann U, Wanner G, Veenhuis M, Schmid M, and Gietl C. AthPEX10, a nuclear gene essential for peroxisome and storage organelle formation during Arabidopsis embryogenesis. Proc

Natl Acad Sci U S A. 2003:**100**(16):9626–9631. https://doi.org/10.1073/pnas.1633697100

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020:**577**(7792):706–710. https://doi.org/10.1038/s41586-019-1923-7

Seong K and Krasileva KV. Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. Nat Microbiol. 2023:**8**(1):174–187. https://doi.org/10.1038/s41564-022-01287-6

Shen J, Fu J, Ma J, Wang X, Gao C, Zhuang C, Wan J, and Jiang L. Isolation, culture, and transient transformation of plant protoplasts. Curr Protoc Cell Biol. 2014:**63**:2.8.1-17. https://doi.org/10.1002/0471143030.cb0208s63

Smith JJ and Rachubinski RA. A Role for the Peroxin Pex8p in Pex20p-dependent Thiolase Import into Peroxisomes of the Yeast Yarrowia lipolytica. J Biol Chem. 2001:**276**(2):1618–1625. https://doi.org/10.1074/jbc.M005072200

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, and Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019:**20**(1):473. https://doi.org/10.1186/s12859-019-3019-7

Trentmann O, Mühlhaus T, Zimmer D, Sommer F, Schroda M, Haferkamp I, Keller I, Pommerrenig B, and Neuhaus HE. Identification of Chloroplast Envelope Proteins with Critical Importance for Cold Acclimation. Plant Physiol. 2020:**182**(3):1239–1255. https://doi.org/10.1104/pp.19.00947

Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, and Steinegger M. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024:**42**(2):243–246. https://doi.org/10.1038/s41587-023-01773-0

Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, Delmont TO, Duarte CM, Eren AM, Finn RD, et al. Unifying the known and unknown microbial coding sequence space. eLife. 2022:**11**:e67667. https://doi.org/10.7554/eLife.67667

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. 2024:**52**(D1):D368–D375. https://doi.org/10.1093/nar/gkad1011

Wang X, McMahon MA, Shelton SN, Nampaisansuk M, Ballard JL, and Goodman JM. Multiple Targeting Modules on Peroxisomal Proteins Are Not Redundant: Discrete Functions of Targeting Signals within Pmp47 and Pex8p. Mol Biol Cell. 2004:**15**(4):1702–1710. https://doi.org/10.1091/mbc.e03-11-0810

Wang Y-X, Catlett NL, and Weisman LS. Vac8p, a Vacuolar Protein with Armadillo Repeats, Functions in both Vacuole Inheritance and Protein Targeting from the Cytoplasm to Vacuole. J Cell

Biol. 1998:**140**(5):1063–1074. https://doi.org/10.1083/jcb.140.5.1063

Waterham HR, Titorenko VI, Haima P, Cregg JM, Harder W, and Veenhuis M. The Hansenula polymorpha PER1 gene is essential for peroxisome biogenesis and encodes a peroxisomal matrix protein with both carboxy- and amino-terminal targeting signals. J Cell Biol. 1994:**127**(3):737–749. https://doi.org/10.1083/jcb.127.3.737

Wickham H. ggplot2: Elegant Graphics for Data Analysis 2nd ed. (Springer: Cham). https://doi.org/10.1007/978-3-319-24277-4

Wu J, Li J, Liu Z, Yin J, Chang Z, Rong C, Wu J, Bi F, and Yao N. The Arabidopsis ceramidase ACER functions in disease resistance and salt tolerance. Plant J. 2015:**81**(5):767–780. https://doi.org/10.1111/tpj.12769

Yu M, Wu J, Zhao C, and Qiu J-L. Exploring plant protein functions through structure-based clustering. Trends Plant Sci. 2025:**30**(10):1111–1118. https://doi.org/10.1016/j.tplants.2025.03.014

Zachariae U, Klühspies T, De S, Engelhardt H, and Zeth K. High Resolution Crystal Structures and Molecular Dynamics Studies Reveal Substrate Binding in the Porin Omp32 *. J Biol Chem. 2006:**281**(11):7413–7420. https://doi.org/10.1074/jbc.M510939200

Zhang L, Leon S, and Subramani S. Two Independent Pathways Traffic the Intraperoxisomal Peroxin PpPex8p into Peroxisomes: Mechanism and Evolutionary Implications. Mol Biol Cell. 2006:**17**.

Zhang Y. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005:**33**(7):2302–2309. https://doi.org/10.1093/nar/gki524

Figure Legends

Figure 1. Structure-based clustering of angiosperm proteins.

(A) Analysis of the 17 selected angiosperm species. The three bar graphs following the phylogenetic tree show the BUSCO assessments of proteome completeness, number of available protein structures in their proteomes, and distribution of structure prediction confidence scores (pLDDT), respectively. **(B)** Illustration of the clustering process of angiosperm protein structures. A total of 564,657 protein structures from 17 flowering plant species were retrieved from AlphaFold DB and clustered into 177,510 clusters. After removing singleton clusters, 25,825 clusters containing at least two members were kept. **(C)** Cumulative distribution of structure prediction confidence scores (pLDDT) in the 564,657 protein structures. The x-axis represents pLDDT values, and the y-axis represents the percentage of structures with pLDDT values greater than the corresponding value on the x-axis. **(D)** Distribution of purities of the 25,825 non-singleton clusters quantified by Pfam consistency, template modeling score (TM-score), and local distance difference test (LDDT). The graphs show a median TM-score of 0.74, a median LDDT of 0.84, and 67.6% of clusters exhibiting 100% Pfam consistency.

Figure 2. Structure-based annotation of non-singleton clusters.

(A) Alignment process for the 25,825 non-singleton clusters. Clusters containing proteins that are annotated in Swiss-Prot were removed, followed by selection of a representative protein for each remaining cluster on which both FoldSeek structural alignments and BLASTp sequence searches against the Swiss-Prot database were performed. By comparing the structural and sequence alignment results, we categorized clusters into three groups: clusters without structural alignment results, clusters with structural alignment results that can also be identified through sequence BLAST, and clusters with structural alignment results that cannot be identified through sequence BLAST. **(B)** Distribution of the 3,109 clusters with structural alignment results unidentifiable by sequence BLAST. X-axis, pLDDT score; y-axis, TM-score (TMalign). The selected clusters (orange) meet the following criteria: (1) TM-score ≥0.5 to indicate significant structural similarity; and (2) pLDDT score ≥0.7 to ensure a high-confidence level in the structural predictions. **(C)** Number of species in the 1,292 selected clusters. The x-axis represents the number of species per cluster, and the y-axis indicates the number of clusters. Protein count distribution for the 1,292 successfully annotated clusters is indicated at the top of the bar graphs. **(D)** Manual cross-validation against multiple reference databases to exclude representatives/clusters with existing functional annotations at least partially overlapping with the functional annotations of their structural matches.

Figure 3. Distribution of the 120 conserved clusters and examples of potential novel functions.

(A) Distribution of the best matches of the 120 conserved clusters in different types of species and the predicted organellar locations of the proteins. **(B)** Example of an uncharacterized Arabidopsis protein matched with another Arabidopsis protein (LPA3). **(C)** Example of the cyanobacterial-like proteins. Structural comparison of Cluster_18328, the Lipase EstA of *Bacillus subtilis*, and a cyanobacterial homologue is shown. **(D)** Example of clusters absent from Arabidopsis. The representative of this cluster, a rice protein, is aligned with the vacuolar protein 8 in yeast. The red and blue numbers in all the panels indicate TM-score and Seq-id, respectively, and numbers in parentheses are UniProt accessions of the proteins corresponding to the displayed structures.

Figure 4. Examples of potentially unannotated molecular functions in the plant clusters.

(A) Top 10 GO terms for the molecular functions of the 120 angiosperm clusters. The GO terms were predicted by QuickGO and then converted to plant-specific GO terms through GOslim. **(B)** Cluster_15445 is matched with the *M. vanbaalenii* phthiocerol/phthiodiolone dimycocerosyl transferase. **(C)** Cluster_12521 is matched with the *S. pombe* RNA TPase. **(D)** Cluster_12986 is matched with *D. acidovorans* OmpC. Further comparison revealed a closer structural similarity to the cyanobacterial (e.g., *Gloeomargarita lithophora*) homolog. **(E)** Comparison of protein structures of the Arabidopsis PEX8-like protein from cluster_10847 with PEX8 proteins from 4 different yeasts, namely, *Pichia pastoris*, *Saccharomyces cerevisiae*, *Hansenula polymorpha*, and *Yarrowia lipolytica*. The red and blue numbers in all the panels indicate TM-score and Seq-id, respectively. The numbers in parentheses are UniProt accessions of the proteins corresponding to the displayed structures.

Figure 5. Sequence, structural and functional analysis of plant PEX8.

(A) Potential PEX8 homologs in land plants, algae, and yeasts, with conserved sequence motifs shown. The heatmaps represent the sequence (Seq ID) and structural (TM score) similarities between Arabidopsis PEX8 and other PEX8 homologs. Sequences at the far right are the PTS1-containing 15 amino acids at the C-terminus of PEX8 homologs, where the PTS1 tripeptides are indicated in pink. **(B)** Protein structures of PEX8 in *Arabidopsis thaliana, Oryza sativa* and *Physcomitrella patens*. **(C)** Growth analysis of wild-type and modified strains of *P. pastoris*. The cell culture was diluted to different concentrations before being inoculated onto YNO plates. YNO medium (0.1% oleic acid, 0.05% Tween 40, 0.1% yeast extract and 0.67% yeast nitrogen base without amino acids) was used. Genotype of the mutant *P. pastoris* is *GS115-HIS4::Cas9-*Δ*PpPEX8*.

Figure 1

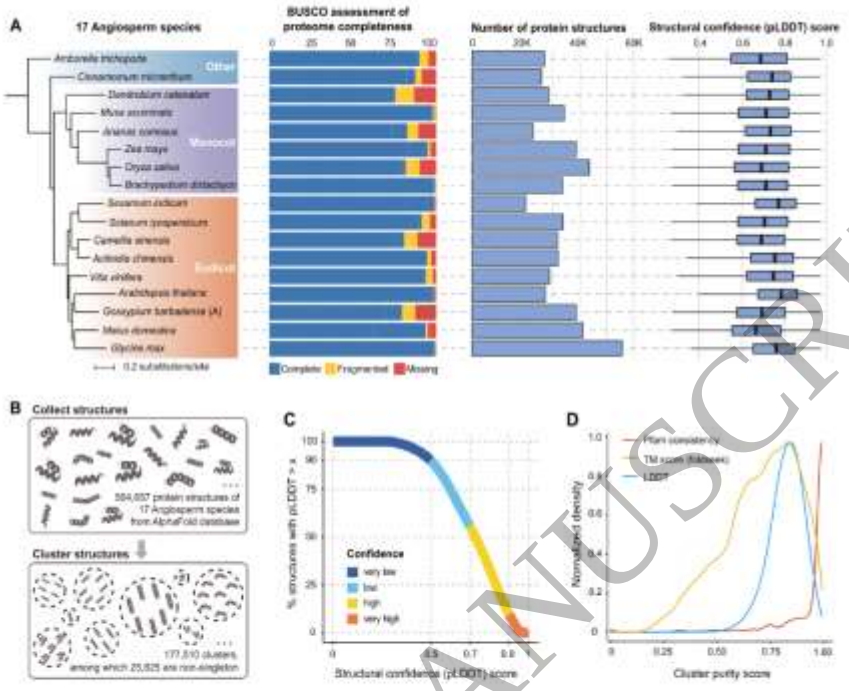

Figure 1
210x297 mm (x DPI)

Figure 2

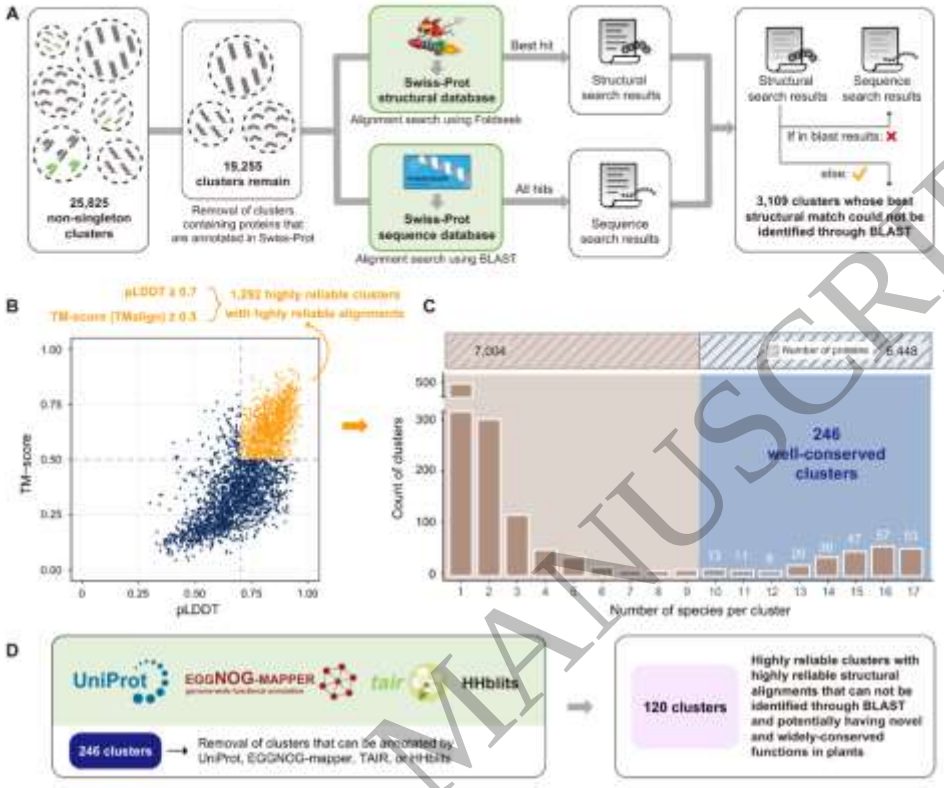

Figure 2
210x297 mm (x DPI)

Figure 3

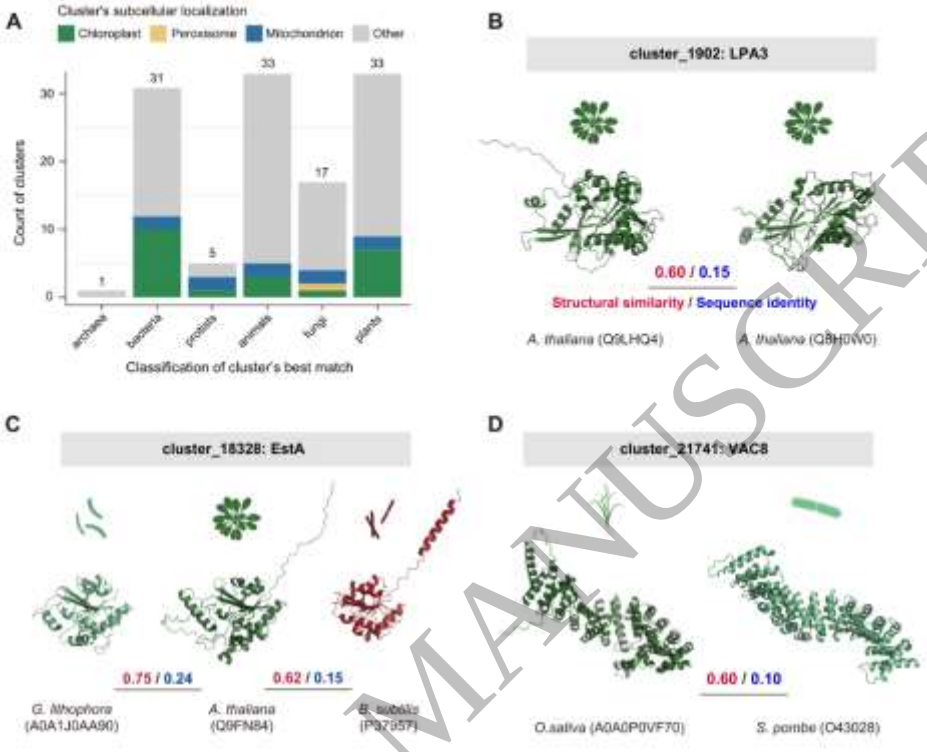

Figure 3
210x297 mm (x DPI)

Figure 4

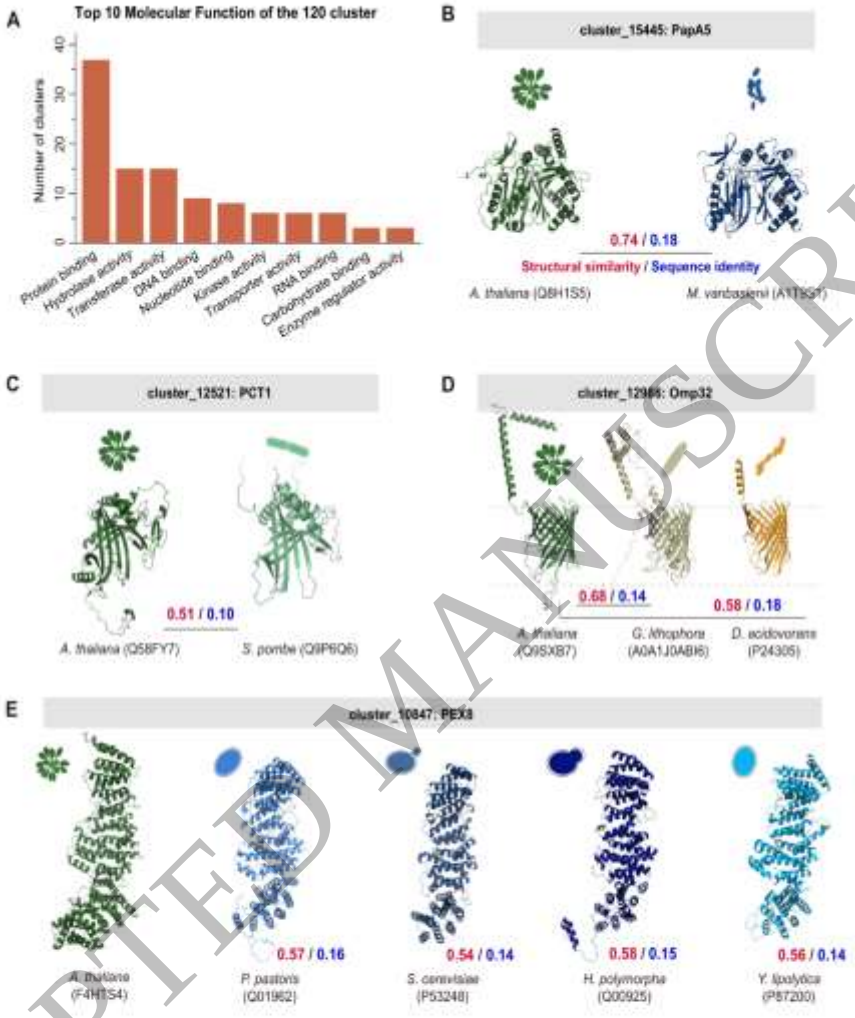

Figure 4
210x297 mm (x DPI)

Figure 5

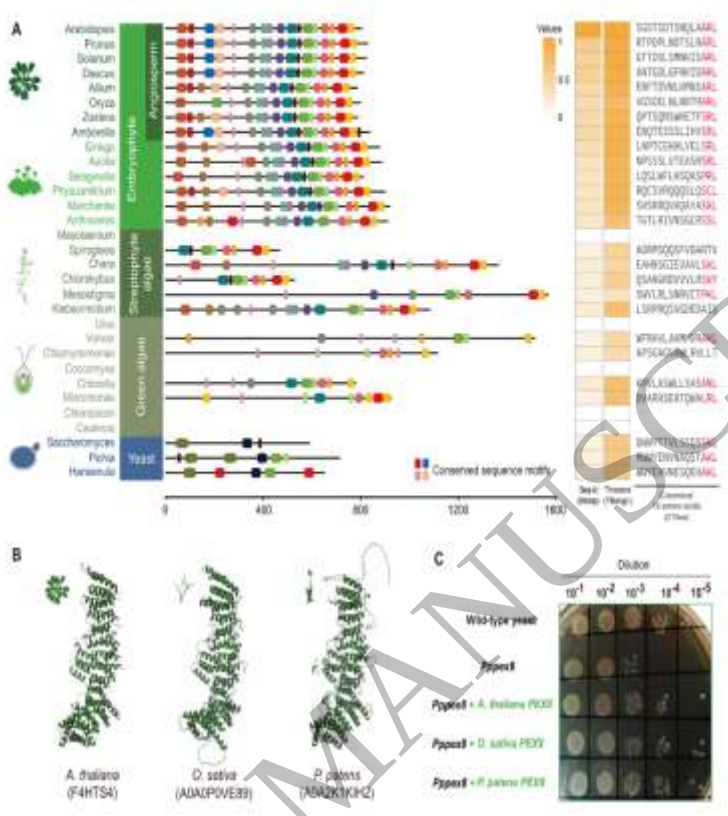

Figure 5
210x297 mm (x DPI)