Misannotated domains in plant databases: lessons from 'PIPLC Y-box-containing' proteins

Lucas Amokrane, Sébastien Aubourg, Eric Ruelland

AUTHOR VERSION OF Lucas Amokrane, Sébastien Aubourg, Eric Ruelland. Misannotated domains in plant databases: lessons from 'PIPLC Y-box-containing' proteins. *Trends in Plant Science*, 2026 , DOI: 10.1016/j.tplants.2026.04.005

Misannotated Domains in Plant Databases: Lessons from "PIPLC Y-box–Containing" Proteins

Amokrane Lucas (0009-0006-1497-4510), Université de Technologie de Compiègne, Unité Génie Enzymatique & Cellulaire, UMR CNRS 7025, F-60203 Compiègne, France

Aubourg Sébastien (0000-0002-0695-4767), INRAE, Institut Agro, Université Angers, IRHS, SFR QUASAV, F-49000, Angers, France

Ruelland Eric (0000-0002-7877-4941), Université de Technologie de Compiègne, Unité Génie Enzymatique & Cellulaire, UMR CNRS 7025, F-60203 Compiègne, France

Correspondence: eric.ruelland@utc.fr (E. Ruelland)
https://gec-upjv.utc.fr/

KEYWORDS

Automatic annotation; Phosphoinositide-dependent phospholipase C (PI-PLC); Protein annotation bias; Functional genomics

ABSTRACT (114 WORDS)

Automated protein annotation pipelines increasingly shape our understanding of gene function, yet their internal biases often go unchallenged. In this opinion article, the domain label "PIPLC Y-box–containing" is used as a case study to

illustrate how misleading annotations can become embedded and propagated across plant protein databases. A lightweight analytical pipeline that exposes inconsistencies between domain names and actual protein structure or function is presented, revealing deeper flaws in domain-based inference and the challenges of automated curation. This example underscores the risks of overinterpretation, threshold drift, and naming inertia. We advocate for greater vigilance in interpreting functional predictions and for integrating structural and evolutionary context to improve the reliability of plant protein annotations.

Main text (2210 words) does not include text in boxes, tables, figure legends, abstract, or references.)

Annotation Bias in Plant Genomics

Automated **annotation pipelines** (see Glossary) are indispensable in modern genomics, yet their outputs are rarely scrutinized. As genomic databases expand [1–4], **domain** annotations—a cornerstone of functional prediction—are often treated as factual, despite their potential to generate errors and distort our understanding of protein function.

Automated pipelines [5] typically begin with **gene models**—predicted coding sequences (CDS) derived from genomic features such as sequence similarities, transcription evidences, open reading frames and splice sites (Figure 1A). Translated protein sequences are then compared to existing entries via similarity searches (e.g., BLASTP) [6] and scanned against domain databases using Hidden Markov Models (HMMs; e.g. **Pfam, PANTHER**) [7–9] or Position-Specific Scoring Matrices (PSSMs; e.g., **PROSITE, CDD**) [10,11] (Figure 1B). HMMs capture full domain architectures by modeling position-specific amino acid probabilities and allowing insertions or deletions [12], whereas PSSMs focus on short, well-conserved motifs, providing faster but less context-aware detection of functional sites [13]. Domains are assigned based on similarity scores and statistical significance thresholds, derived from curated alignments and training sets [8,14], and detected domains are assembled into architectures which are used to predict molecular function [15,16]. Protein names and function descriptors (often using Gene Ontology classification [17]) are then assigned either by transferring information from the closest characterized or predicted homologs (homology-based propagation) or by applying rule-based templates derived from detected domain architectures (e.g., “X-domain–containing protein”) [5]. When multiple domains are detected, annotation systems assign precedence according to curated hierarchies that favor functionally defining domains (e.g., catalytic over regulatory or structural). Canonical combinations with known architectures (e.g.,

kinase + regulatory domain) generate standardized names, whereas ambiguous or low-confidence configurations default to generic templates such as “X-domain–containing protein.” This automated prioritization ensures consistency but can also amplify upstream domain misannotations [18,19]. In practice, every database label represents a cascade of probabilistic decisions—from gene prediction to homology inference, domain detection and **functional annotation**—each step carrying potential bias if statistical thresholds, domain co-occurrence, or biological context are not rigorously evaluated [20–23].

In large annotation systems, these layers of inference can mask the uncertainty of underlying sequence matches. To illustrate how such uncertainty propagates into functional assignments, we analyzed plant proteins annotated as “PIPLC Y-box–containing.” This case reveals how low-confidence domain assignments can spread across databases, exposing systemic flaws in domain-based inference and underscoring the need for greater rigor in annotation practices.

What is a PI-PLC Y-box?

Phosphoinositide-dependent phospholipases C (PI-PLCs) hydrolyze phosphatidylinositol 4,5-bisphosphate (PIP_2) into diacylglycerol (DAG) and inositol 1,4,5-trisphosphate (IP_3), triggering diverse signaling pathways [13–15]. In plants, DAG is further converted to phosphatidic acid (PA), a lipid mediator in stress and pathogen responses. Canonical PI-PLCs comprise four conserved regions: an N-terminal EF-hand domain, the catalytic **PIPLC_X_box** and **PIPLC_Y_box** domains, and a C-terminal **C2 domain** [16–19] (Figure 2A). Unlike animal PI-PLCs, plant isoforms lack additional domains (e.g., PH, SH2, SH3) [20,21]. The X-box and Y-box domains are structurally interdependent, forming a TIM barrel–like active site only when combined in the NH_2–X–Y–COOH arrangement (Figure 2B) [22,23]. Important residues for PI-PLC catalytic activity are at the interface between X-box and Y-box (Figure 2C) [24]. Thus, neither the X-box nor the Y-box functions independently; treating them as separate domains ignores their structural interdependence and risks false-positive assignments. Despite this, **UniProtKB**

[5] annotates a subset of plant proteins as “PIPLC Y-box domain–containing,” implying the presence of a standalone catalytic Y-box domain. Although the X and Y regions likely co-evolved to form a single catalytic entity, we cannot entirely exclude that they originated as distinct ancestral modules. This evolutionary uncertainty reinforces the need to examine what lies behind the “PIPLC Y-box domain–containing” label—whether it represents a genuine, independent structural unit or merely a statistical echo of the Y-region embedded within full PI-PLC enzymes.

To clarify whether this annotation reflects a genuine domain or a misleading artifact, we focused on Viridiplantae proteins explicitly named “PIPLC Y-box domain-containing protein.” In the following sections, these sequences are compared to *bona fide* PI-PLCs and the nature of their so-called Y-box regions is dissected.

Do “Y-box” Annotations Reflect Real Domains?

To investigate the meaning of the “Y-box” label in annotated proteins, we searched the UniProtKB database (release 2024_03) using the keyword 'PIPLC Y-box domain-containing protein’. This search returned 1,889 entries, of which 1,085 had exactly this protein name. These entries spanned all domains of life: Archaea (10), Bacteria (382), Eukaryotes (614), and viruses (10), with an additional 69 entries from metagenomes. We focused our analysis on the 179 entries from the Viridiplantae taxonomic group. The most represented genera were *Gossypium* (cotton, 39 entries), *Oryza* (rice, 22), and *Populus* (poplar, 9). Interestingly, although five entries were attributed to the genus *Arabidopsis*, none were from *Arabidopsis thaliana*. We then examined whether these 179 entries contained a *bona fide* PIPLC_Y_DOMAIN. To do so, we first queried UniProtKB for all Viridiplantae proteins annotated with the Prosite PIPLC_Y_domain (PS50008), regardless of their protein name. This yielded 2,675 entries**,** from which we defined two groups for comparison: Group 1, corresponding to proteins with a detected PIPLC_Y_DOMAIN and named “PIPLC Y-box domain-containing protein” (*n* = 82)

and Group 2 corresponding to proteins with a detected PIPLC_Y_DOMAIN and named “Phosphoinositide-PLC” (or similar naming) (*n* = 2,276). The mean Prosite confidence score for Group 1 was 9, while for Group 2 it was 34. In addition, the predicted domain lengths were shorter and more variable in Group 1 (mean 47 aa) than in Group 2 (mean 84 aa). These observations indicate that the so-called “Y-box” regions in Group 1 proteins are likely spurious or truncated predictions rather than *bona fide* catalytic Y-domains.

Revealing the True Identity of Misannotated Proteins

To investigate the functional identity of proteins annotated as “PIPLC Y-box domain–containing” we applied a generalizable strategy combining sequence clustering (multiple alignment), domain analysis, and **phylogenetic analysis**. Our goal was to determine whether these proteins genuinely belong to the PI-PLC family or whether their annotations resulted from partial domain matches. We first examined the distribution of protein lengths among the 179 *Viridiplantae* entries and then focused on the sequences corresponding to the three major peaks (Figure 3A). These entries were further analyzed based on their physicochemical properties (such as pI, % of tiny residues, % of aliphatic residues; % of charged residues…) retrieved from EMBOSS Pepstats [25], revealing three major groups and several outliers (Figure 3B).

From these, we selected a representative protein—A0A0E0FMM1 from *Oryza nivara*—to illustrate our workflow (Figure 3B). We performed a BLASTP search restricted to Viridiplantae using UniProt reference proteomes using A0A0E0FMM1 as query. Among the closest matches, none was labeled as PI-PLCs, but surprisingly some were labeled as “plant mobile domain–containing” proteins (**PMD,** [26–28]). To clarify these relationships, we aligned the top hits and constructed a phylogenetic tree (Figure 3C). The analysis showed that A0A0E0FMM1 and its close homologs consistently clustered within the PMD-containing protein family.

We then focused on the region annotated as a “PIPLC_Y_DOMAIN” (InterPro IPR001711) in A0A0E0FMM1 and its detected homologs. These predicted domains were only 23–24 residues long**,** significantly shorter than the canonical 117-residue Y-box. Sequence alignment confirmed that they were highly conserved internally but showed only 31% identity with the AtPLC5 Y domain. Structural modeling of a representative sequence (A2X6N8) using I-TASSER placed the predicted “Y-box” mainly overlapping the PMD domain (Figure 3D). Together, these results show that the so-called “PIPLC_Y_DOMAIN” of A0A0E0FMM1 and similar “PIPLC Y-box domain–containing” proteins correspond to a subregion of the PMD, and not to a Y-domain in the PI-PLC sense.

We applied this same workflow to other representative sequences from the major clusters and outliers. In each case, the closest matches belonged to known protein families, unrelated to PI-PLCs, including AUGMIN8-like proteins and TAB2-like proteins. AUGMIN8-like proteins are named due to their homology with AUGMIN subunit 8, a subunit of a complex involved in microtubule organization, dynamics, and branching [29]. These proteins are essential for cell division, elongation, and morphogenesis in plants, regulating cortical microtubule reorientation and tubulin polymerization [30]. They are characterized by the presence of a **QWRF** domain (InterPro IPR007573), which is a conserved motif found in AUGMIN subunit 8-like proteins [31]. The QWRF domain is critical for their function in microtubule-associated processes [32]. TAB2-like proteins are named for their homology with TAB2 (Translational Activator of psaB), a protein involved in regulating chloroplast gene expression and photosynthetic activity [33]. Specifically, TAB2 facilitates the translation of *psaB* mRNA, which is essential for photosystem I assembly. TAB2-like proteins contain a Tab2-like domain (InterPro IPR009472; [34]), which is associated with their role in chloroplast function and photosynthetic regulation. For the sequences annotated as “PIPLC Y-box domain-containing” that showed homology with AUGMIN8-like or TAB2-like proteins, the predicted “Y-box” regions overlapped with portions of established domain families (QWRF or TAB2-like).

These overlaps indicate that the Y-box signal reflects low-confidence, partial matches to existing domains produced by domain prediction tools with relaxed thresholds, rather than identifying a distinct, functional domain.

True PI-PLCs and partial fragments: a minority of cases

One cluster (peak 2 proteins from Figure 3A) contained sequences related to PI-PLCs, but their alignments covered only a limited region of PI-PLCs. Most aligned with the N-terminal portion of the *Arabidopsis thaliana* PI-PLC5 PIPLC_Y domain, while one sequence (UniProt B9GIH3) also extended into part of the adjacent C2 domain. None displayed the complete EF-hand–X–Y–C2 domain organization typical of functional PI-PLCs. This situation highlights the difficulties inherent in annotating pseudogenes and constructing accurate gene models. Non-functional genes affected by deletions, frameshifts, or premature stop codons can retain detectable similarities with their functional homologs, misleading prediction software to reconstruct coherent gene structures.

In conclusion of our case study, among the 179 *Viridiplantae* entries annotated as “PIPLC Y-box domain–containing,” detailed phylogenetic and structural analyses revealed that 84 corresponded to TAB2-like proteins, 9 to AUGMIN8-like isoforms, and 7 to plant mobile domain (PMD)–containing proteins, while only a few (≤ 5) represented genuine PI-PLCs. These numbers are not exhaustive, as our goal was not to reassign all sequences but to illustrate recurrent cases of misannotation within distinct protein families.

The Role of Context in Annotation Errors

The distinction between domains and motifs is often blurred in annotation pipelines, yet these entities differ both structurally and functionally: domains are independently folding units, whereas motifs are short conserved signatures that usually act in a specific structural context. Our case study highlights how

annotation errors can arise when short sequence motifs are interpreted outside their structural or functional context. Many sequences in our dataset matched only short regions corresponding to fragments of the PI-PLC Y domain or unrelated plant motifs (e.g., PMD, QWRF, or TAB2-like regions). These local similarities were nevertheless treated as evidence of a full “domain,” leading to enzyme-level annotations. In reality, such cases often reflect threshold effects in sequence detection rather than true domain homology. Motif-based resources such as PROSITE are highly reliable when used for their intended purpose—detecting conserved catalytic or binding sites—but extrapolating these short signatures to infer whole-domain architectures can lead to overinterpretation [10]. Re-evaluating such hits in a structural framework, whenever possible, helps prevent cascading annotation errors. In some cases, apparent similarities may result from convergent compositional bias rather than a shared evolutionary origin, reflecting the limited sequence diversity of short conserved motifs [35].

The Propagation of Errors and Its Consequences

This analysis highlights a deeper issue: domain labels are not neutral descriptors. The “Y-box” designation has been applied to proteins that clearly belong to well-characterized families with no mechanistic, functional or evolutionary links to PI-PLCs. Misannotations driven by weak domain similarity propagate into databases and literature through a snowball effect (function inferred from mislabeled proteins in a recursive way), obscuring the true identity and function of these proteins. Without critical review, such errors accumulate, misleading downstream studies and confounding comparative genomics. Once introduced, such misannotations tend to propagate across databases.

Mitigating Misannotation: The Need for Context and Expertise

Our findings underscore weaknesses in current annotation processes and highlight the need for caution when interpreting automated labels. To improve

accuracy, annotations should systematically integrate: physicochemical properties and predicted structures (e.g., AlphaFold models), phylogenetic context (homologous sequences), annotation confidence scores (e.g., UniProt's metrics), where low scores should trigger expert review. Structural annotation issues—such as gene fusions, splitting, or pseudogenes—can also contribute to erroneous domain architectures. For example, truncated or degraded gene models may retain isolated motifs that are misinterpreted as functional domains.

Concluding Remarks

This case study underscores the limitations of current protein annotation systems, particularly their reliance on low-confidence domain matches [20] and insufficient structural validation [36]. Such errors can obscure biological function and mislead downstream research, as demonstrated here and elsewhere [21]. To address these challenges, we advocate for annotation pipelines that integrate predicted structural data, domain architecture, and phylogenetic context [22,23]. Context-aware scoring—such as penalizing isolated domain detections or incorporating phylogenetic validation—could further reduce misclassification [23].

Beyond domain identification, predicting biological function remains a significant challenge. Reliable inference requires a combination of approaches—sequence similarity searches, motif detection, ortholog definition, transcriptomic profiling, and genomic context—yet even these strategies are incomplete without experimental validation [37]. Extending automated frameworks toward integrative, evidence-weighted models is essential for improving functional predictions in plants.

Ultimately, high-quality annotation depends not only on advanced algorithms but also on continuous expert curation. Resources like UniProtKB/Swiss-Prot demonstrate the value of combining computational predictions with literature-based validation and community expertise [5]. This philosophy is reflected in initiatives such as the *Arabidopsis thaliana* genome annotation update (TAIR12), where collective expertise refines structural and functional annotations. Such

efforts underscore the enduring importance of expert input and data integration in ensuring biological accuracy, even as automated prediction scales up.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used ChatGPT (GPT-5, OpenAI) to assist in language editing, rephrasing, and structural refinement of the text. After using this tool, the author carefully reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Giani, A.M. *et al.* (2020) Long walk to genomics: History and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* 18, 9–19
2. Exposito-Alonso, M. *et al.* (2020) The Earth BioGenome project: opportunities and challenges for plant genomics and conservation. *Plant J.* 102, 222–229
3. Lewin, H.A. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* 115, 4325–4333
4. Mukherjee, S. *et al.* (2025) Genomes OnLine Database (GOLD) v.10: new features and updates. *Nucleic Acids Res.* 53, D989–D997
5. The UniProt Consortium *et al.* (2025) UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* 53, D609–D617
6. Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
7. Mistry, J. *et al.* (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419
8. Blum, M. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354
9. Mi, H. *et al.* (2021) PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* 49, D394–D403
10. Sigrist, C.J.A. *et al.* (2012) New and continuing developments at PROSITE. *Nucleic Acids Res.* 41, D344–D347
11. Lu, S. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268
12. Paysan-Lafosse, T. *et al.* (2025) The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.* 53, D523–D534
13. Yu, R. *et al.* (2024) Utilizing profile hidden Markov model databases for discovering viruses from metagenomic data: a comprehensive review. *Brief. Bioinform.* 25, bbae292

14. Finn, R.D. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285
15. Suskiewicz, M.J. *et al.* (2023) Updated protein domain annotation of the PARP protein family sheds new light on biological function. *Nucleic Acids Res.* 51, 8217–8236
16. Hulo, N. *et al.* (2007) The 20 years of PROSITE. *Nucleic Acids Res.* 36, D245–D249
17. The Gene Ontology Consortium *et al.* (2023) The Gene Ontology knowledgebase in 2023. *GENETICS* 224, iyad031
18. Koehorst, J.J. *et al.* (2017) Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. *F1000Research* 5, 1987
19. Goll, J. *et al.* (2010) The Protein Naming Utility: a rules database for protein nomenclature. *Nucleic Acids Res.* 38, D336–D339
20. Kress, A. *et al.* (2023) Real or fake? Measuring the impact of protein annotation errors on estimates of domain gain and loss events. *Front. Bioinforma.* 3, 1178926
21. Schnoes, A.M. *et al.* (2009) Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput. Biol.* 5, e1000605
22. Asim, M.N. *et al.* (2025) Protein Sequence Analysis landscape: A Systematic Review of Task Types, Databases, Datasets, Word Embeddings Methods, and Language Models. *Database* 2025, baaf027
23. Lin, B. *et al.* (2024) A comprehensive review and comparison of existing computational methods for protein function prediction. *Brief. Bioinform.* 25, bbae289
24. Amokrane, L. *et al.* (2024) Phospholipid Signaling in Crop Plants: A Field to Explore. *Plants* 13, 1532
25. Rice, P. *et al.* (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16, 276–277
26. Ikeda, Y. *et al.* (2017) Arabidopsis proteins with a transposon-related domain act in gene silencing. *Nat. Commun.* 8, 15122
27. Nicolau, M. *et al.* (2020) The plant mobile domain proteins MAIN and MAIL1 interact with the phosphatase PP7L to regulate gene expression and silence transposable elements in Arabidopsis thaliana. *PLOS Genet.* 16, e1008324
28. Pélissier, T. *et al.* (2025) Plant mobile domain protein–DNA motif modules counteract Polycomb silencing to stabilize gene expression. *Nat. Plants* DOI: 10.1038/s41477-025-02127-1
29. Gonzalez, J.P. *et al.* (2023) The role of intrinsic disorder in binding of plant microtubule-associated proteins to the cytoskeleton. *Cytoskeleton* 80, 404–436
30. Hotta, T. *et al.* (2012) Characterization of the Arabidopsis augmin complex uncovers its critical function in the assembly of the acentrosomal spindle and phragmoplast microtubule arrays. *Plant Cell* 24, 1494–1509
31. Li, Z. *et al.* (2021) Plasmodesmata-Dependent Intercellular Movement of Bacterial Effectors. *Front. Plant Sci.* 12, 640277

32. Ma, H. *et al.* (2021) Arabidopsis QWRF1 and QWRF2 Redundantly Modulate Cortical Microtubule Arrangement in Floral Organ Growth and Fertility. *Front. Cell Dev. Biol.* 9, 634218
33. Barneche, F. *et al.* (2006) ATAB2 is a novel factor in the signalling pathway of light-controlled synthesis of photosystem proteins. *EMBO J.* 25, 5907–5918
34. Dauvillee, D. (2003) Tab2 is a novel conserved RNA binding protein required for translation of the chloroplast psaB mRNA. *EMBO J.* 22, 6378–6388
35. Biegert, A. and Söding, J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci.* 106, 3770–3775
36. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
37. Majidian, S. *et al.* (2025) Quest for Orthologs in the era of Data Deluge and AI: Challenges and Innovations in Orthology Prediction and Data Integration. *J. Mol. Evol.* DOI: 10.1007/s00239-025-10272-6

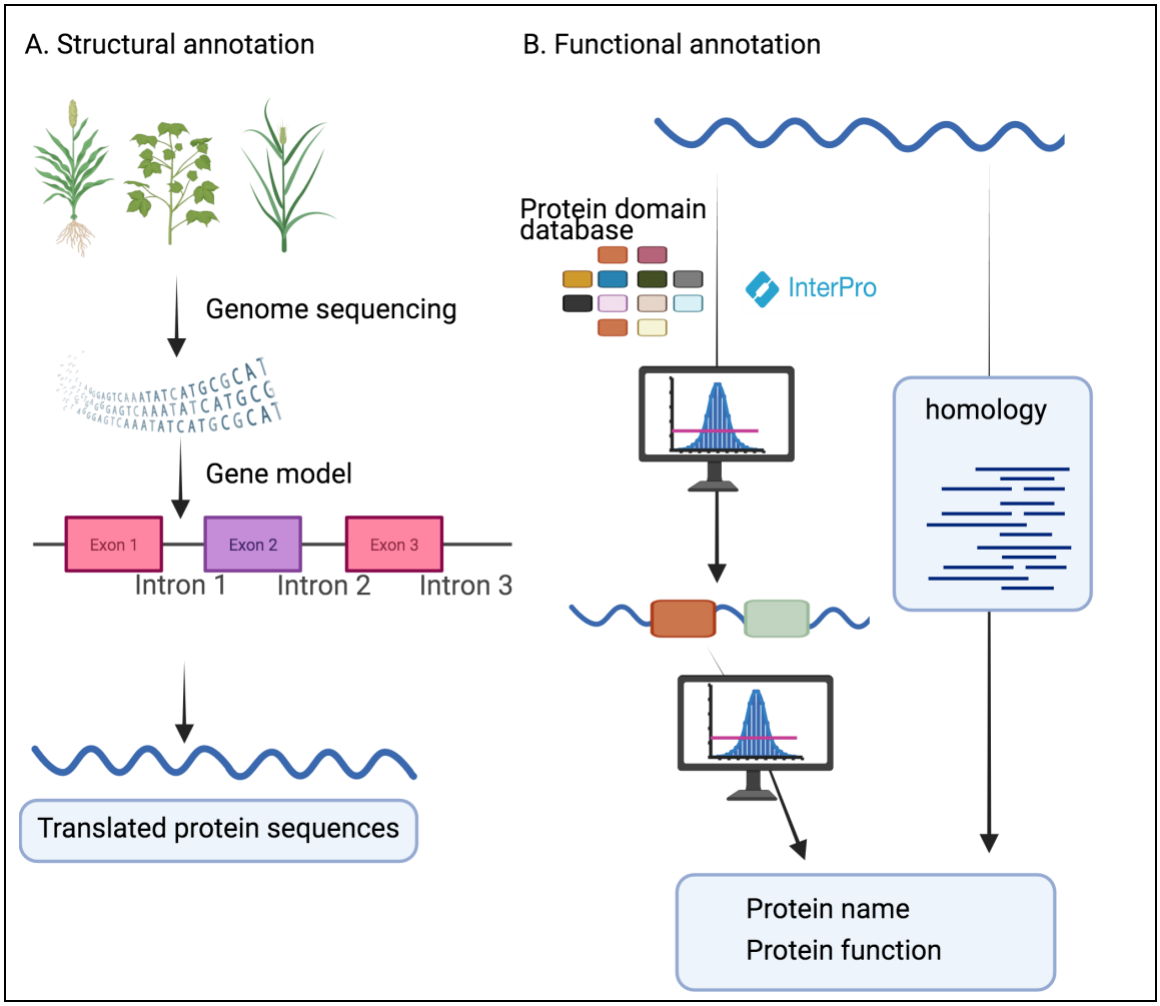

Figure 1. Overview of automated protein annotation pipelines. (A) Gene models are inferred from genomic features such as open reading frames and splice sites, producing coding sequences (CDS); the corresponding protein sequence is then deduced. **(B)** Proteins are compared to reference databases through two complementary routes: homology searches, which transfer information from the closest characterized proteins, and domain detection (e.g., InterPro), which defines domain architectures. Both types of evidence are integrated to infer protein names and functions during automated annotation. Figure created with biorender.com.

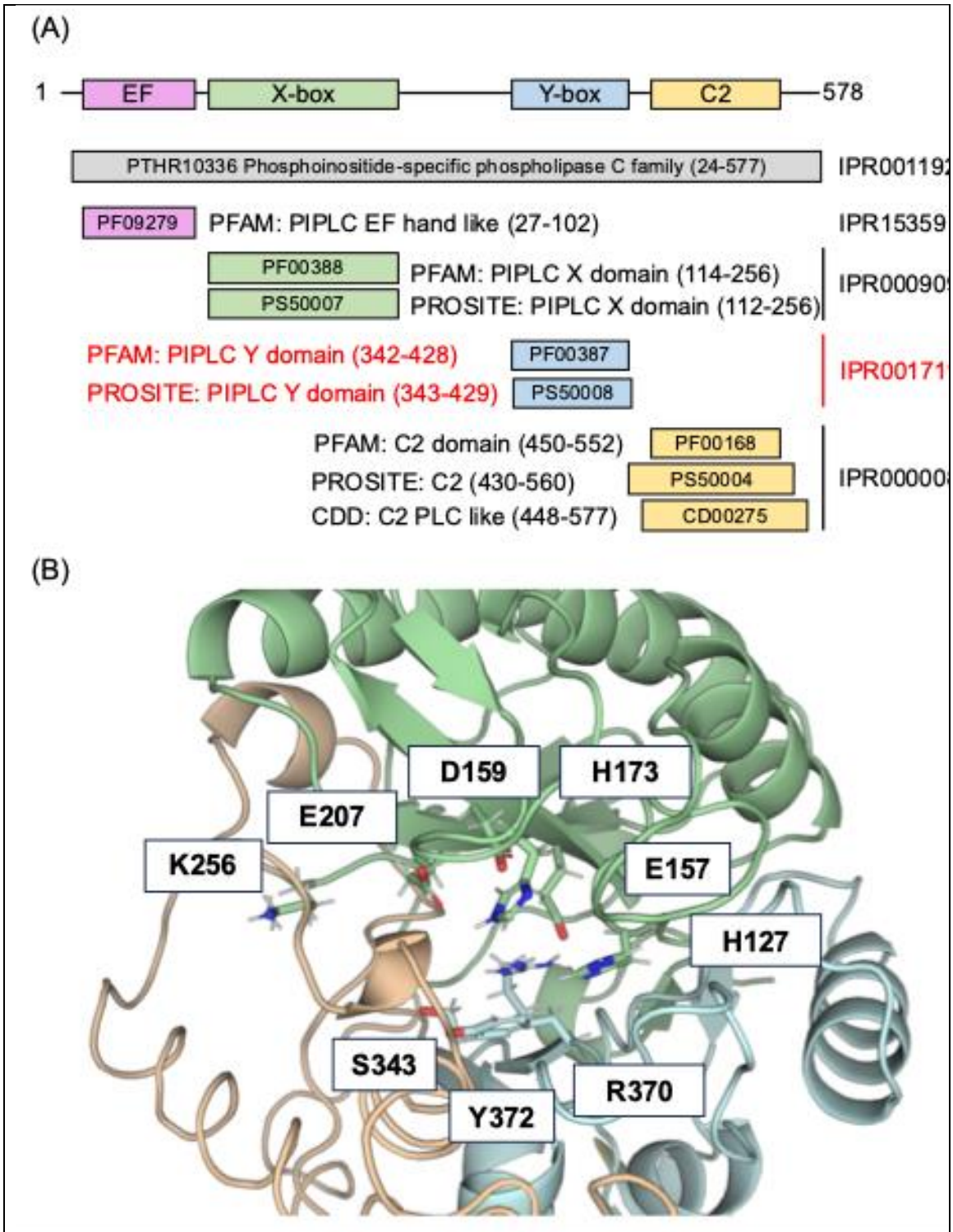

Figure 2. Domain organization and catalytic interface based on the example of *Arabidopsis thaliana* phosphoinositide-dependent phospholipase C 5 (AtPLC5). (A) Domain architecture of AtPLC5 (578

aminoacids) showing the canonical arrangement of an N-terminal EF-hand (pink), the catalytic X_box (green) and Y_box (blue) regions, and a C-terminal C2 domain (yellow). The structure is based on the domains as defined by PFAM. Each domain appears in different name, with different start and end positions, in the different domain databases. These names and positions are shown. Each domain has also an InterPro identifier indicated on the right of the panel. The PANTHER classification (PTHR10336) applies to the entire PI-PLC family, as PANTHER models protein families and subfamilies rather than individual domains. **(B)** Close-up view of the predicted 3D structure of AtPLC5 generated by I-TASSER, focusing on the catalytic interface between the X- and Y-boxes. Domains are colored as follows PIPLC_X_box (green), PIPLC_Y_box (blue). Key catalytic residues (H127, E157, D159, H173, E207, K256, S343, R370, Y372) cluster at the X/Y-box interface.

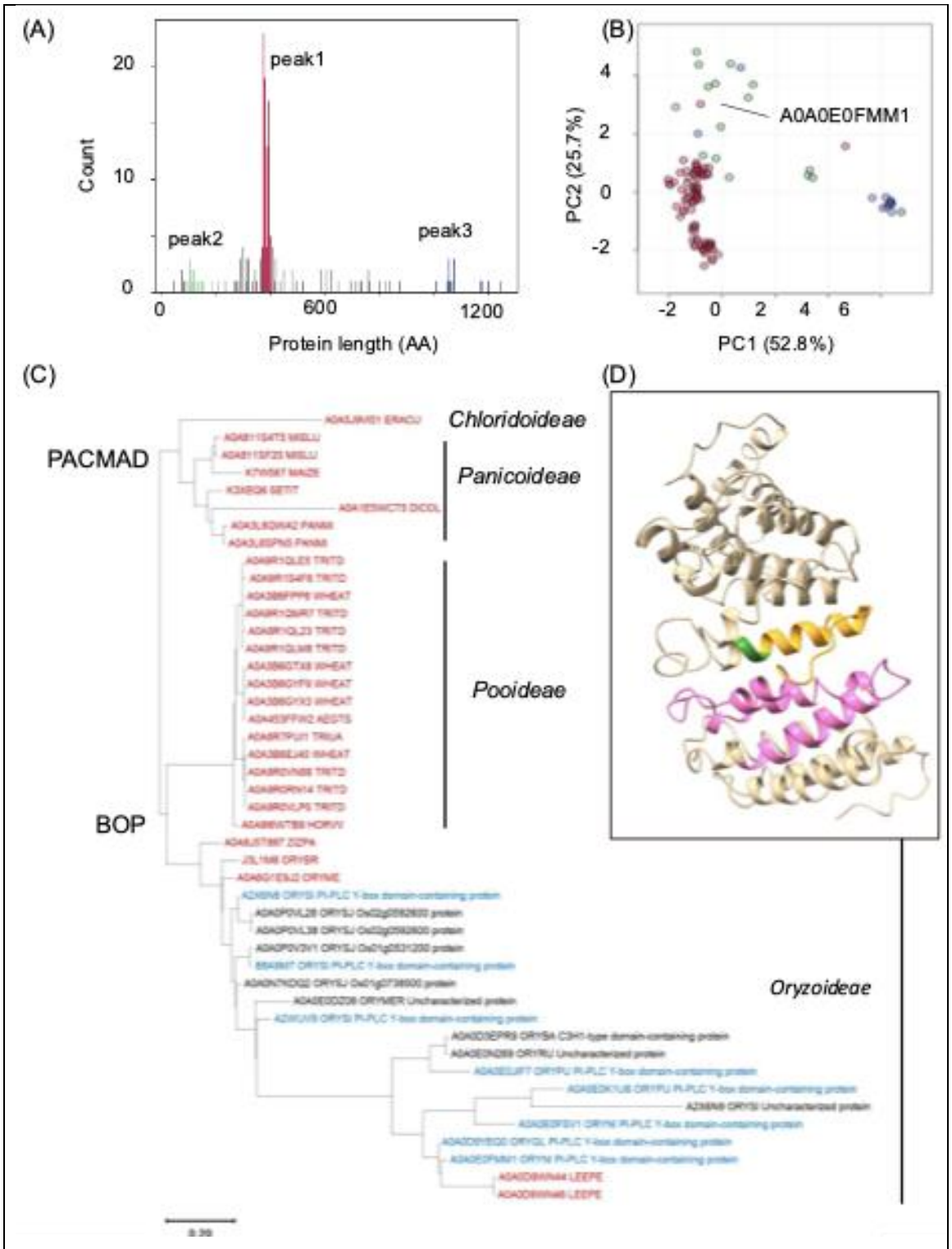

Figure 3. Workflow and case study for re-evaluating proteins annotated as "PIPLC Y-box domain-containing." (A) Distribution of sequence lengths of 179 Viridiplantae entries **(B)** Sequence diversity of entries from peaks 1, 2 and

3 defined in (A) analyzed with EMBOSS Pepstats and visualized by principal-component analysis. **(C)** Phylogenetic tree of Viridiplantae homologs indicating that *Oryza nivara* A0A0E0FMM1 and its closest relatives coherently cluster within the plant mobile domain (PMD) family across grass subfamilies (Oryzoideae, Panicoideae, Pooideae, etc.). The phylogeny follows the established division of grasses into the PACMAD and the BOP clades. In blue, «PIPLC Y-box domain-containing proteins», in red proteins with PMD in their names. **(D)** Predicted structural model of a representative homolog (A2X6N8) generated by I-TASSER (model 1; C-score = –1.53), showing the predicted PMD (pink and orange) and the short region annotated as “PIPLC_Y_DOMAIN” (orange and green). The Y-domain region is overlapping the PMD fold (orange).

GLOSSARY (444 words)

Annotation pipeline: Computational workflow that assigns biological meaning to sequences by predicting genes and functional elements. It integrates homology searches, domain detection using probabilistic models such as Hidden Markov Models (HMMs), and rule-based naming schemes. Incomplete or low-confidence matches can propagate annotation errors.

C2 domain: Conserved C-terminal domain of phosphoinositide-specific phospholipase C (PI-PLC) enzymes, mediating calcium-dependent phospholipid binding. Accessions: InterPro, IPR000008; Pfam, PF00168; PROSITE PS50004; CDD cd00275.

Domain: Independently folding region of a protein, typically associated with a specific structural or functional role (e.g., catalysis or binding). In contrast, a *motif* is a short, conserved sequence pattern that contributes to a function (such as binding or regulation) but does not form an autonomous structural unit. Automated

pipelines sometimes blur this distinction, leading to the misclassification of motifs as domains and to overinterpretation of weak matches.

Domain misannotation: Incorrect domain assignment caused by weak or fragmentary similarity, threshold bias, or pseudogene remnants, leading to spurious functional inference.

Functional annotation: Computational assignment of predicted biological roles—encompassing molecular functions, pathways, and standardized protein names—based primarily on sequence similarity and conserved domains. Most automated annotations rely on statistical inference rather than experimental validation.

Gene model: Predicted structure of a gene, including exons, introns, and untranslated regions, derived from genomic features such as open reading frames and splice sites.

InterPro / Pfam / PANTHER / PROSITE / CDD: Major databases providing complementary models for domain and family classification. InterPro integrates data from multiple sources (including Pfam, PROSITE, PANTHER, and CDD) to assign unified domain identifiers and confidence scores. Each resource relies on distinct statistical models, alignment strategies, and training sets, so the same domain may be defined with slightly different boundaries across databases, leading to minor discrepancies in start and end positions. PANTHER differs from the others by classifying protein families and subfamilies rather than individual domains.

PIPLC_X_box and PIPLC_Y_box: Catalytic regions of PI-PLC enzymes that together form the active site; neither functions independently. X-box accesions: InterPro, IPR000909; Pfam, PF00388; PROSITE, PS50007. Y-box accesions: InterPro, IPR001711; Pfam, PF00387; PROSITE PS50008.

PMD (Plant Mobile Domain): Aminotransferase-like domain (Accessions: InterPro, IPR019557; Pfam, PF10536) found in plant proteins such as MAIN and

MAIL1, involved in transposon silencing, genome stability, and developmental regulation.

Protein structure prediction: Computational inference of three-dimensional protein conformation from a protein amino acid sequence (e.g., AlphaFold, I-TASSER).

QWRF: Domain present in AUGMIN8-like proteins that participate in microtubule organization and cell morphogenesis. The QWRF (InterPro, IPR007573; Pfam, PF04511) is thought to mediate interactions with tubulin and other microtubule-associated factors.

UniProtKB: Comprehensive protein knowledgebase combining manually reviewed (Swiss-Prot) and automatically annotated (TrEMBL) records. The Viridiplantae Swiss-Prot section (release 2025_03) includes ≈ 41 900 entries (< 0.2 % of UniProtKB).

Figure Legends (250 words per legend)

Figure 1. Overview of automated protein annotation pipelines. (A) Gene models are inferred from genomic features such as open reading frames and splice sites, producing coding sequences (CDS); the corresponding protein sequence is then deduced. **(B)** Proteins are compared to reference databases through two complementary routes: homology searches, which transfer information from the closest characterized proteins, and domain detection (e.g., InterPro), which defines domain architectures. Both types of evidence are integrated to infer protein names and functions during automated annotation. Figure created with biorender.com.

Figure 2. Domain organization and catalytic interface based on the example of *Arabidopsis thaliana* phosphoinositide-dependent phospholipase C 5 (AtPLC5). (A) Domain architecture of AtPLC5 (578 aminoacids) showing the canonical arrangement of an N-terminal EF-hand (pink), the catalytic X_box (green) and Y_box (blue) regions, and a C-terminal C2 domain (yellow). The structure is based on the domains as defined by PFAM. Each domain appears in different name, with different start and end positions, in the different domain databases. These names and positions are shown. Each domain has also an InterPro identifier indicated on the right of the panel. The PANTHER classification (PTHR10336) applies to the entire PI-PLC family, as PANTHER models protein families and subfamilies rather than individual domains. **(B)** Close-up view of the predicted 3D structure of AtPLC5 generated by I-TASSER, focusing on the catalytic interface between the X- and Y-boxes. Domains are colored as follows PIPLC_X_box (green), PIPLC_Y_box (blue). Key catalytic residues (H127, E157, D159, H173, E207, K256, S343, R370, Y372) cluster at the X/Y-box interface.

Figure 3. Workflow and case study for re-evaluating proteins annotated as "PIPLC Y-box domain-containing." (A) Distribution of sequence lengths of 179 Viridiplantae entries **(B)** Sequence diversity of entries from peaks 1, 2 and 3 defined in (A) analyzed with EMBOSS Pepstats and visualized by principal-component analysis. **(C)** Phylogenetic tree of Viridiplantae homologs indicating

that *Oryza nivara* A0A0E0FMM1 and its closest relatives coherently cluster within the plant mobile domain (PMD) family across grass subfamilies (Oryzoideae, Panicoideae, Pooideae, etc.). The phylogeny follows the established division of grasses into the PACMAD and the BOP clades. In blue, «PIPLC Y-box domain-containing proteins», in red proteins with PMD in their names. **(D)** Predicted structural model of a representative homolog (A2X6N8) generated by I-TASSER (model 1; C-score = –1.53), showing the predicted PMD (pink and orange) and the short region annotated as “PIPLC_Y_DOMAIN” (orange and green). The Y-domain region is overlapping the PMD fold (orange).