nature portfolio

Peer Review File

A functional atlas of transposon-encoded products and their integration into host networks

Corresponding Author: Dr Leandro Quadrana

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This study presents an integrative analysis of transposable elements (TEs) in Arabidopsis thaliana, combining long- and short-read transcriptomics, regulatory network inference, deep proteomics, and structural modeling. The authors report that TE-derived transcripts and proteins are embedded in host regulatory circuits, interact with host factors, and in some cases, display evidence of domestication or functional co-option. The work aims to provide a comprehensive atlas of TE products and proposes that TEs act as active drivers of genome innovation. Overall, the integrative approach is ambitious and potentially impactful. I have the following minor suggestions for improvement:

1. For terminological clarity, it would be more precise to use “protein-coding genes” rather than “cellular genes” throughout the manuscript, as this avoids ambiguity and aligns with established genomic nomenclature.

2. In Figure 2A, please specify what the colors represent in the co-expression networks.

3. In Figure 2I, please provide the corresponding p-value for the shown comparison.

4. In Figure 2C, rather than summarizing transcription factor binding sites (TFBSs) by TE classes, it would be more informative to display the distribution of TFBSs per individual TE in a heatmap format. This would better illustrate TE-specific regulatory associations and strengthen the related conclusions.

5. Several key terms are used repeatedly but lack explicit definitions or references. For example:

- "Active TEs": How were these defined? Based on transcript abundance, epigenetic marks, protein evidence, or other criteria?
- "DNA methylation-sensitive TFs": Please provide the methods or criteria used to identify these transcription factors, along with appropriate references.

6. In lines 103–106, the authors state that "En/Spm DNA transposons are mainly induced by methylation loss, whereas Ty1/Copia, Ty3/Gypsy, and MuDR elements are activated under environmental stress." However, no supporting data or citations are provided. Please either include relevant experimental results and statistical validation or cite prior studies that substantiate these claims.

Reviewer #2

(Remarks to the Author)
In this manuscript, Borreda et al. started by presenting a new atlas of transposable elements (TEs) transcripts of Arabidopsis thaliana using newly produced ONT cDNA long reads from plants defective in multiple layers of epigenetic silencing mechanisms (epigenetic mutants), under control or stress conditions. Using a large collection of RNAseq covering diverses Wt tissues and growth conditions, the authors next built and explored a transcriptional regulatory network. Interestingly, they identified that most co-expression clusters are integrating both celular- and TE-genes, highlighting their potential interaction. They focused on possible role of Transcriptional Factors (TFs), and predicted thousand co-expressed TF-TE pairs. They explored the effect of lost in DNA methylation in co-expressed clusters by using RNAseq of epigentics mutants. Importantly,

they experimentally demonstrated the role of DNA methylation in the regulation of the co-expression of one TE-TF pair. They further generated new proteomic datasets from Wt and the epigenetic mutant ddm1, that they combined with public proteomic datasets covering multiple tissues and developmental stages. They provided a new TE protein atlas composed of about four hundreds TE proteins of diverse functional categories.
They further clustered modeled predicted TE proteins and cellular gene proteins and identified 3 cellular gene proteins that could orignate from TE domestication and co-option events. The authors also explored the ability for several TE proteins to form higher-odrder complexes. Finaly, using an experimentally benchmarked in silico method, they predicted how TE assessory protein candidates could interact with celular proteins.

The study is highly original and provides important insights about how TE transcripts and their encoded-proteins are part of the transcriptonal and protein network of plants, by interacting with cellular genes and proteins. The study provide an interactive web portal for a part of the presented data.

The feedback below is intended to strengthen the study.

Majors comments :

Missing document : I could not find the Suplementary Tables among the recieved documents.

Overall : Authors should make it clearer when it is subject to "TE-transcript units" (meaning distinct transcriptional units), and "TE-transcript isoforms" (meaning different versions of a TE-transcript unit). For instance, the Figure 1E indicates "Novel TE-transcript", but the associated legend indicates "TE-transcript isoform". Figure S2B also indicates the word "isoform".
The others chapters of the manuscript indicate "TE-genes" as distinct transcriptional units, mixing several nomenclatures, which make it difficult to follow/understand the results.

Lines 107-109 : "Together, we detect expression for the greatest number of Arabidopsis TEs so far, with over 60% of the TE transcripts detected here being absent from previous splicing-aware, transcriptome-guided, or predicted TE-gene annotations".

- Did the authors identified 60% more TE-transcript units or TE-transcript isoforms?

- If the 60% are related to "TE-transcript units", how the authors explain a such important difference compared to previous studies?

- Could the ONT cDNA methodology or FLAIR methodology have overestimated the number of isoforms per TE-transcript units?

- The authors must indicate the number of novel TE-transcript units they identified compare to previous studies by intersecting the TE annotations, that is an important information.

Line 238 : I think this is not indicated in the manuscript how the authors indentified the 2,008 high-quality TE-transcripts (also shown in Figure 3A).

Lines 288 – 289 : "Most of these proteins had high or very high-confidence structural predictions (pLDDT > 70) (Figure S9)."
To my understanding, the authors should rewrite the sentence using "a bit more than half of these proteins", instead of "Most of these proteins". In Figure S9, it is clear that they are about half modeled proteins with pLDDT inferior to 70.

Line 448 : "We found no evidence for global TE reactivation in any Arabidopsis developmental stage or stress condition. Instead, we identified distinct TE subsets with divergent transcriptional programs, suggesting that reactivation occurs in specifie cellular niches (Figure S5D). "
The authors must discuss of this point with the literature, for exemple the article : doi: 10.1016/j.cell.2008.12.038 , which already explored this subject using several Arabidospis tissues and identified that TEs are globally reactivated in pollen vegetative nucleus.

Minors :

Overall : I found several Figures wrongly cited in the manuscript, affecting the quality of the work. They are also several figures that are difficult to understand/read because they are too small or of bad quality, such as Figure S5C, D and E.

Lines 51 - 55 : "However, their diversity, function, and interactions with host proteins are largely unknown. This knowledge gap is largely due to the lack of reliable functional annotations of TEs. In fact, most TE sequences in a genome are thought to be fossilized copies of once active elements and typically have lost all coding potential."
The authors must add relevant litterature citations that has addressed this subject. Such as : doi: 10.1038/s41576-020-0251-y

Line 72 : The authors should rewrite "contains 32,000 annotated TE" to "contains about 32,000 annotated TE".

Reviewer #3

(Remarks to the Author)
The present manuscript describes a fascinating attempt to identify and characterise transposon-encoded products on an organism-wide scale in Arabidopsis, largely through plants defective in DDM1. As written, the manuscript implies that evidence for hundreds of TE-encoded proteins has been uncovered from proteomics analysis, that these proteins have extensive multimerization capacity, and that they are likely to be functionally integrated in host cellular processes.

These are intriguing implications. However they remain largely hypothetical, as they are based on in silico modelling of proteins that are implied from ORF-containing TE transcripts, most of which are not experimentally observed in the present study. This lack of broad-scale experimental validation of proteins undermines the study's overall impact.

For example, the section 'A hidden diversity of TE-encoded protein domains' describes numerous examples of protein structures modelled using AlphaFold2, from which well thought out and interesting interpretations are made. However it is unclear how many, if any, of these structures are associated with proteins experimentally identified from proteomics experiments. Similarly, it is unclear if any of the 18 proteins selected for AlphaFold-Multimer modelling (in the 'Structural modelling of TE multiprotein complexes' section) were uncovered from data associated with the present study, or if they were all sourced from references 42 and 43.

To clarify what proteins were identified, the confidence of these identifications, and the source of the data from which the identifications were made (i.e. data collected from the present study or data sourced from reference 28), substantially more detail needs to be included in the supplementary information. Crucially, the criteria from which proteins have been identified also needs to be substantially tightened, and it is unclear how many proteins will remain after appropriate filtering criteria have been applied (more below).

Specific points that require addressing:

1. As written, the study has treated proteins identified from single peptides as bona fide TE-encoded products. This is a lax criterion for protein identification in a standard proteomics study. For a study designed to uncover novel proteins, which requires substantially higher levels of identification stringency, this criterion is inappropriate unless additional validation criteria are included (or unless proteins identified from single peptides are explicitly treated as putative and lacking in strong experimental evidence).

To address this, it is suggested that, at a minimum, only proteins identified from 2 or more unique peptides are considered identified.

If proteins identified from single peptides are to be included, it is strongly suggested that additional validation criteria for these individual peptides are required. One option would be to collect spectra from synthetic peptides and demonstrate equivalence to spectra collected from plant samples. Another would be to ensure that the single peptides from which proteins are identified meet stringent q-value thresholds across multiple independent proteomics datasets, and that these spectra cannot be otherwise explained by modified peptides uncovered from open modification searches. An example of such a pipeline can be found here:
https://www.sciencedirect.com/science/article/pii/S1535947624000094

2. Using the present identification criteria, the authors state that 426 proteins derived from 372 TEs were identified. Information pertaining to the basis of these protein identifications is lacking. Currently, the only information available to the reader is found in Table S6, which simply includes a 'Found in MS?' column.

If criteria outlined by the leading proteomics journal, Molecular and Cellular Proteomics, are used, the protein identifications should be expanded to include the following: all peptide sequences assigned, precursor change and mass/charge, all modifications observed, score(s), and count of the number of distinct peptide sequences assigned to each protein.

The employed MaxQuant and DIA-NN search parameters should also be included in the methods.

3. Moreover, from Table S6, only 46 proteins listed as True in the 'Found in MS?' column contain a value in the te_id column. This is far below the 426 stated by the authors in the body text. Is this because most of the protein identifications came from data sourced from reference 28 rather than the present data? If so, can the authors comment on this result, particularly in relation to the fact that the reference 28 data does not come from plants defective in DDM1?

If the above protein identification thresholds are applied, it is entirely possible that the total number of genuinely identified proteins could be in the single digits when only searching proteomics data from the present study. While validation of any TE-encoded protein would be of interest, identification of so few TE-encoded proteins would unfortunately appear to be well below the ambition and scope of the manuscript as currently written.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thanks to the authors for well addressing my concerns. I have no further comments.

Reviewer #2

(Remarks to the Author)
The authors have successfully addressed all my comments. I have no further concerns.

Reviewer #3

(Remarks to the Author)
The authors have thoroughly addressed the comments regarding the proteomics data analysis. They have made well considered alterations to the both the main body of the manuscript and the supplementary information, which reflect the changes to the reported number of protein identifications.

A minor observation is that there now appear to be some small inconsistencies in the numbers of high confidence proteins that are reported. In the text of the revised manuscript, the authors report 135 identified proteins, 115 of which were found in ddm1 plants. From my assessment of Supplementary Data 9, I can count 139 identified proteins, 107 of which were found in ddm1 plants from the present study.

Similarly, I count 196 proteins in Supplementary Data 8, which when added to the above 139 proteins makes 335 proteins, which is below the 243 proteins stated in line 260.

Finally, in the abstract, the authors state that proteomics analyses confirm the production of hundreds of TE-encoded proteins. This would now be more appropriately stated as 'over one hundred'.

Aside from the above minor observations, the overall proteomics analysis has been substantially improved and appropriately reflects the underlying data.

Reviewer #4

(Remarks to the Author)
Overall the work appears very interesting with a number of valuable observations and validations. However, I have several comments that should be taken into consideration as for the usage of AlphaFold for global analyses. In this review, I will specifically focus on the sections of the article dedicated to the use of AlphaFold to characterize molecular structures and interactions. In the manuscript, three main applications are presented i) the use of AlphaFold for structural classifications, ii) for stoichiometry estimation of homo-oligomers and iii) for interactome screening.

Structural classification

The use of Foldseek to identify potential structural similarities is appropriate. To assign similarities to the predicted structures of TE ORFs, the authors employed a relatively relaxed E-value threshold of 0.01, which is higher than the default value of 0.001. This choice should be justified and the manuscript would benefit from a more explicit statement clarifying this relaxed parameter selection.

The authors also used LDDT metrics to estimate structural distances between TE ORFs and other proteins which is somewhat unconventional. Large-scale structural clustering is more commonly performed using TM-scores (see for example 10.1038/s41586-023-06622-3). While LDDT is primarily a local measure of structural accuracy, TM-scores are better suited for assessing global structural similarity in a properly normalized manner. The manuscript does not appear to reference key studies describing or benchmarking LDDT, and it would be useful to include such citations, as well as studies comparing different structural similarity metrics. Although the authors would likely reach similar conclusions for most TE ORFs using TM-scores, the assignment of specific cases could be improved. Validating their results using TM-scores computed with standard tools such as US-align or GTalign would align the analysis with common practices and strengthen the robustness of their conclusions.

Stoichiometry estimation

The authors performed a careful analysis of the potential homomerization properties of the detected TE-encoded proteins. They used the ipTM score as a metric to estimate the most likely stoichiometry of the predicted homomultimers. A recent benchmarking study (10.1101/2025.03.10.642518v2) has evaluated the potential of ipTM for predicting protein oligomeric states and was recently updated. The authors could refer to this work to support their methodological choice.

As explained in the next section, in supplementary material, the author should provide a detailed list of the scores calculated for these dimers and oligomers, especially the average of the ipTM scores over the 5 generated models which account for the robustness of the prediction. This important for readers to judge the reliability of the predictions. For instance, testing the interaction between VANB and VANC either by AF2 or AF3 did not convince me from the scores that an interaction can be reliably predicted between the two.

They could also note that distinguishing oligomeric states is not always straightforward and has been addressed using alternative strategies, such as in 10.1016/j.cell.2024.01.022. In that study, the authors compared the geometry of a predicted dimer in isolation with that observed within cyclic oligomers, showing that the assembly whose geometry best matches the free dimer is likely to represent the correct oligomeric state.

Interactome screening

The section describing the AlphaFold2-Multimer screen lacks sufficient methodological detail and would benefit from substantial clarification. Several elements presented in Figure 6 are not adequately described in the Methods section, which makes it difficult to assess the robustness of the analysis. In addition, the web portal provides only limited information regarding the predicted scores. As illustrated by resources such as Predictomes (https://predictomes.org/), access to multiple complementary metrics is essential to properly evaluate the reliability of high-scoring interactions.

Relying solely on the maximum ipTM score is problematic in large-scale interactome screens and can lead to a substantial number of false positives, as discussed in 10.1016/j.molcel.2025.01.034. This issue can arise for multiple reasons. For instance, stochastic effects may lead to spurious interactions: for large prey proteins, a single model among the five generated may display a high ipTM score, while the interaction is not consistently reproduced across models, indicating low reliability. Second, certain bait proteins may exhibit non-specific binding behavior (for example, due to amphipathic helices embedded in disordered regions) which can artificially inflate interaction scores. The bait VANB appears to fall into this category. For instance, when I attempted to reproduce two randomly selected interactions reported in the web portal (AT3G54230.2 and AT5G50780.1, with reported ipTM values of 0.75 and 0.73, respectively), the distribution of ipTM scores across the five models was markedly inconsistent, supporting concerns about robustness. For the five models generated between VANB and AT3G54230.2, the ipTM scores were [0.52, 0.32, 0.31, 0.23, 0.20]. In contrast, for VANB and AT5G50780.1, the scores were [0.70, 0.58, 0.55, 0.32, 0.23]. The fact that only one model yields an ipTM > 0.5 suggests that the interaction with AT3G54230.2 is highly unlikely, whereas for AT5G50780.1, the presence of three models with ipTM > 0.5 may indicate a more reliable signal. However, AT5G50780.1 contains a coiled-coil domain that can pair with that of VANB, while VANB itself is predicted to form a homodimer through this same coiled-coil, raising the possibility of non-specific or competing interactions.

To mitigate these issues, it would be essential to report not only the maximum ipTM score, but also the full set of scores across the five models, along with their average. This typically provides a more reliable indicator of interaction robustness. In addition, the list of interface residues, as well as their potential overlap with homodimer interfaces for high-scoring pairs, should be reported. Further strengthening the analysis would require the inclusion of complementary metrics, such as PAE maps (for example, provided in GIF format) and alternative interface-focused scores, including 10.1101/2025.02.10.637595 (IPSAE) or 10.1093/bioinformatics/btaf107 (actifpTM). Providing this information through the web portal would substantially enhance the practical utility of the predictions and help reduce biases arising from the use of overly permissive ipTM thresholds in Figures 6D and 6F.

Finally, Figures 6A and 6B refer to a systematic benchmarking procedure used to define the ipTM threshold for calling positive interactions, yet this procedure is not described in the Methods section, which represents by itself a concern. Key aspects remain unclear, including how the 600 positive heterodimers were selected and how random pairs were generated. Beyond this, the main issue is that just using random pairs is not a valid approach as carefully stated in 10.1016/j.molcel.2025.01.034. The authors could draw inspiration from this benchmarking study to improve this analysis. In particular, it seems important to restrict both positive and negative sets to comparable subsets (e.g., “C+” categories), as balancing these classes typically reveals that a higher ipTM threshold is required for reliable discrimination in large-scale interactome studies.

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
Thanks to the authors for clarifying the protein counts and successfully addressing my concerns. I have no further comments.

Reviewer #4

(Remarks to the Author)
The authors have properly taken into account my comment and improved the methodological aspects related to the use of AlphaFold.

We thank the reviewers for their positive appraisal of our manuscript and the points raised, which we have answered as detailed below (in blue). We believe that as a result, the revised manuscript is substantially improved and we are grateful for this.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This study presents an integrative analysis of transposable elements (TEs) in Arabidopsis thaliana, combining long- and short-read transcriptomics, regulatory network inference, deep proteomics, and structural modeling. The authors report that TE-derived transcripts and proteins are embedded in host regulatory circuits, interact with host factors, and in some cases, display evidence of domestication or functional co-option. The work aims to provide a comprehensive atlas of TE products and proposes that TEs act as active drivers of genome innovation. Overall, the integrative approach is ambitious and potentially impactful.

We thank Reviewer #1 for the constructive comments on our manuscript.

I have the following minor suggestions for improvement:

1. For terminological clarity, it would be more precise to use “protein-coding genes” rather than “cellular genes” throughout the manuscript, as this avoids ambiguity and aligns with established genomic nomenclature.

We agree that the key distinction is TE-derived versus non-TE genes. In the framework of our paper, the term “protein-coding genes” could be interpreted as excluding TE genes, whereas our results show that many TEs also encode proteins. We have therefore replaced “host (non-TE) genes” by “cellular genes” throughout and defined this terminology explicitly early on in the Results section (line 93). In sections focused on translation and proteomics, we refer to proteins encoded by TE genes versus cellular genes, making clear that both classes can encode proteins.

2. In Figure 2A, please specify what the colors represent in the co-expression networks.

The legend was shared between Figure 2A and Figure 2D. Since this was not clear from the figure, we have included the legend below each of the subfigures to increase clarity.

3. In Figure 2I, please provide the corresponding p-value for the shown comparison.

We generated three independent 35S::HSFA2 lines per background (Col-0 and *ddm1*) and quantified ONSEN induction by RT-qPCR (n = 3 lines, each with 3 replicates). This experiment shows a statistically significant difference in HSFA2-induced *ONSEN* expression in *ddm1* compared to wild-type backgrounds (P-value=0.017, two-tailed t-test). We have updated Figure 2I with these new results and provided the p-value of these comparisons.

4. In Figure 2C, rather than summarizing transcription factor binding sites (TFBSs) by TE classes, it would be more informative to display the distribution of TFBSs per individual TE in a heatmap format. This would better illustrate TE-specific regulatory associations and strengthen the related conclusions.

We thank the reviewer for this suggestion. We previously had this information on the manuscript, but given that Arabidopsis harbors over 30,000 TEs, the total size of the heatmap largely exceeded the size of a full-page figure. We have now included a new Supplementary Figure including the total TFBS distribution summarized per TE family so as to fit a single figure (Supplementary Figure 7). In addition, we now include a supplementary table with the list of TFBS over each TE (Supplementary Data 5) as well as included this information on the TE-atlas website to facilitate its exploration.

5. Several key terms are used repeatedly but lack explicit definitions or references. For example:

- "Active TEs": How were these defined? Based on transcript abundance, epigenetic marks, protein evidence, or other criteria?

We thank the reviewer for this comment. Indeed, TEs can be considered active based on many different criteria. We now avoid the term ‘active’ unless explicitly qualified as transcriptionally active (transcriptomic) or transpositionally active (new insertions, literature).

- "DNA methylation-sensitive TFs": Please provide the methods or criteria used to identify these transcription factors, along with appropriate references.

We apologize for this missing information. We have now expanded the methodology section (Lines 215-222 and 721-726) to indicate that DNA methylation sensitivity of TF was calculated based on the differential occupancy of TFs on amplified DNA (i.e. no DNA methylation) vs native DNA (i.e. with endogenous DNA methylation) as measured by DAPseq experiments reported in O’Malley et al 2016.

6. In lines 103–106, the authors state that "En/Spm DNA transposons are mainly induced by methylation loss, whereas Ty1/Copia, Ty3/Gypsy, and MuDR elements are activated under environmental stress." However, no supporting data or citations are provided. Please either include relevant experimental results and statistical validation or cite prior studies that substantiate these claims.

We thank the reviewer for this comment. That specific statement was derived from Figure 1D, where we show the proportion of expressed TEs on control conditions, and the extra TEs reactivated upon stress, as a proportion of the total TE number. This procedure was unclear on the figure caption and we have now updated it.
We have also included a new supplementary figure (Supplementary Figure 3) in the manuscript showing the proportion of TEs expressed on either control or stress conditions. Moreover, we used a χ^2 test to formally test whether the difference in TE reactivation is statistically significant across superfamilies and included this data into the supplementary figure.

Reviewer #2 (Remarks to the Author):

In this manuscript, Borreda et al. started by presenting a new atlas of transposable elements (TEs) transcripts of Arabidopsis thaliana using newly produced ONT cDNA long reads from plants defective in multiple layers of epigenetic silencing mechanisms (epigenetic mutants), under control or stress conditions. Using a large collection of RNAseq covering diverses Wt tissues and growth conditions, the authors next built and explored a transcriptional regulatory network. Interestingly, they identified that most co-expression clusters are integrating both celular- and TE-genes, highlighting their potential interaction. They focused on possible role of Transcriptional Factors (TFs), and predicted thousand co-expressed TF-TE pairs. They explored the effect of lost in DNA methylation in co-expressed clusters by using RNAseq of epigentics mutants. Importantly, they experimentally demonstrated the role of DNA methylation in the regulation of the co-expression of one TE-TF pair. They further generated new proteomic datasets from Wt and the epigenetic mutant ddm1, that they combined with public proteomic datasets covering multiple tissues and developmental stages. They provided a new TE protein atlas composed of about four hundreds TE proteins of diverse functional categories.
They further clustered modeled predicted TE proteins and cellular gene proteins and identified 3 cellular gene proteins that could orignate from TE domestication and co-option events. The authors also explored the ability for several TE proteins to form higher-odrder complexes. Finaly, using an experimentally benchmarked in silico method, they predicted how TE assessory protein candidates could interact with celular proteins.

The study is highly original and provides important insights about how TE transcripts and their encoded-proteins are part of the transcriptonal and protein network of plants, by interacting with cellular genes and proteins. The study provides an interactive web portal for a part of the presented data.

We thank very much reviewer #2 for the enthusiastic and positive appraisal of our work and the many points raised, which have been very useful to improve the manuscript significantly.

The feedback below is intended to strengthen the study.

Majors comments :

Missing document : I could not find the Suplementary Tables among the recieved documents.

-

We apologize for the inconvenience. We have re-uploaded all Supplementary Tables to the submission portal and verified their availability. In addition, the full supplementary dataset is available through the Zenodo repository as well as through the TE-atlas website.

Overall : Authors should make it clearer when it is subject to "TE-transcript units" (meaning distinct transcriptional units), and "TE-transcript isoforms" (meaning different versions of a TE-transcript unit). For instance, the Figure 1E indicates "Novel TE-transcript", but the associated legend indicates "TE-transcript isoform". Figure S2B also indicates the word "isoform".
The others chapters of the manuscript indicate "TE-genes" as distinct transcriptional units, mixing several nomenclatures, which make it difficult to follow/understand the results.

We apologize for the ambiguity of the terms. To avoid misunderstandings we now refrain from using transcriptional units in the manuscript to refer to TE-genes. We now refer to “TE-transcript isoforms” or “TE-genes” only.

Lines 107-109 : "Together, we detect expression for the greatest number of Arabidopsis TEs so far, with over 60% of the TE transcripts detected here being absent from previous splicing-aware, transcriptome-guided, or predicted TE-gene annotations".

- Did the authors identified 60% more TE-transcript units or TE-transcript isoforms?

We apologize for the unclear wording. We specifically referred to TE-transcript isoforms. Following the previous comment, we now homogenize the nomenclature and clarify this point on the text (Line 113).

- If the 60% are related to "TE-transcript units", how the authors explain a such important difference compared to previous studies?

The 60% refer to TE-transcript isoforms, many of them are derived from novel TE-genes. The increase in the number of TE-transcript isoforms detected in our work compared to previous efforts could be accounted for by the distinctive combination of higher sequencing depth, epigenetic backgrounds, and environmental stimulus used in our work. Specifically, the long-read sequencing depth of Panda and Slotkin is more than five times lower than the one used here (7M vs 38M), they only include control conditions, and their resulting mean transcript length is much lower than ours (990bp vs 1560, Supplementary Figure 2). Berthelier et al, produced ~5M reads from *ddm1* plants, and ~4M reads from *met1* plants, which has a lower number of reactivated TEs compared to *ddm1*, and did not include ONT sequencing of environmentally stressed plants either. We have now clarified these sources of discrepancy on the text (lines 110-115).

- Could the ONT cDNA methodology or FLAIR methodology have overestimated the number of isoforms per TE-transcript units?

Overestimation of transcript isoforms is indeed a source of concern for transcriptomic-based annotations. To limit such biases, we used very restrictive thresholds to call an isoform, including ≥5 ONT reads supporting transcript start-stops and splicing junction combinations; each junction further supported by ≥5 Illumina reads. In total, 94.9% of identified splicing junctions of cellular genes were previously reported in TAIR10 or splicing-aware annotations, supporting the robustness of the transcript annotation pipeline. We also show that most TE-genes produce a single isoform (Figure 1G), indicating limited inflation due to over annotation of TE-transcript isoforms.

- The authors must indicate the number of novel TE-transcript units they identified compare to previous studies by intersecting the TE annotations, that is an important information.

We have included a supplementary figure depicting the number of TE-genes identified by our work and in previous studies (Supplementary Figure 2C).

Line 238 : I think this is not indicated in the manuscript how the authors indentified the 2,008 high-quality TE-transcripts (also shown in Figure 3A).

Thanks for pointing out this lack of clarity. We agree that the basis for the 2,008 “high-quality TE transcripts” subset was not sufficiently explained, and that subsetting the set of TE-transcripts may lead to unnecessary biases. To remove any ambiguity, we have revised this section to no longer rely on that subset and now report all downstream analyses using the full set of TE transcripts identified in our study across epigenetic mutant backgrounds under control and stress conditions (5,732 TE transcript isoforms). Consequently, all predicted protein products derived from these transcripts are now included without additional filtering, and the reported numbers of predicted ORFs and globular domains have been updated accordingly. These changes further reinforce the main conclusions of the section while making the analysis more transparent and comprehensive. The results section (Lines 307-315), Figure 4, and Supplementary Figure S12 have been updated accordingly.

Lines 288 – 289 : "Most of these proteins had high or very high-confidence structural predictions (pLDDT > 70) (Figure S9)."
To my understanding, the authors should rewrite the sentence using "a bit more than half of these proteins", instead of "Most of these proteins". In Figure S9, it is clear that they are about half modeled proteins with pLDDT inferior to 70.

To avoid qualifying the magnitude of the proportion and increase precision, we now report directly the exact percentage.

Line 448 : "We found no evidence for global TE reactivation in any Arabidopsis developmental stage or stress condition. Instead, we identified distinct TE subsets with divergent transcriptional programs, suggesting that reactivation occurs in specifie cellular niches (Figure S5D). "
The authors must discuss of this point with the literature, for exemple the article : doi: 10.1016/j.cell.2008.12.038 , which already explored this subject using several Arabidospis tissues and identified that TEs are globally reactivated in pollen vegetative nucleus.

We thank the reviewer for raising this important point and for pointing us to Slotkin et al. (2009). We agree that pollen is a central context for TE regulation, and we have expanded the Discussion to explicitly compare our results with this prior work.

In Slotkin et al., TE derepression in pollen was inferred primarily from microarray measurements and reporter lines focusing on a limited set of annotated elements. In our study, we quantified TE expression using splicing-aware, genome-wide NGS datasets across multiple tissues and conditions (Supplementary Data 2), applying a consistent operational definition of an “expressed TE-gene” as TPM ≥ 5 (Supplementary Figure 6A). Under this definition, we do not observe evidence for genome-wide TE reactivation in

whole pollen samples. Instead, we consistently detect subset-specific TE transcription in male reproductive samples, with a restricted set of TE-genes showing robust expression and/or strong relative enrichment compared to other tissues (Supplementary Figures 6A, 10B and 10C).

To further exclude that the absence of a “global” signal is an artifact of short-read quantification on repetitive sequences, we analyzed ONT cDNA datasets from pollen and leaves in Col-0. We specifically compared (i) pollen-expressed genes and TEs highlighted in Slotkin et al. (2009) and (ii) the subset of pollen-enriched TE-genes identified in our atlas. This long-read analysis supports the same conclusion: rather than broad TE derepression, pollen expression is dominated by a limited subset of TE-genes (new Supplementary Figure 10F). Finally, consistent with the transcriptomic results, reanalysis of tissue proteomics does not reveal a marked increase in either the fraction or the number of TE-derived proteins detected in pollen relative to other tissues (Supplementary Figure 10D and Supplementary Data 7).

We believe these results can be reconciled with Slotkin et al. by considering methodological and biological differences. For example, microarray-based quantification and normalization can emphasize modest changes in expression of repetitive sequences and may be sensitive to cross-hybridization, while reporter lines are informative but may not fully reflect endogenous TE expression levels and can be influenced by genetic background, which in the case of Slotkin et al were performed using the non-reference Accession Ler-0. Importantly, our conclusion is restricted to the absence of a genome-wide TE upregulation above a robust detection limit. We cannot exclude that broader, low-level derepression occurs below the detection or quantification limits of standard short- and long-read transcriptomic approaches. We have added a short paragraph to the Discussion acknowledging the lack of global TE reactivation in pollen (lines 472-482).

Minors :

Overall : I found several Figures wrongly cited in the manuscript, affecting the quality of the work. There are also several figures that are difficult to understand/read because they are too small or of bad quality, such as Figure S5C, D and E.

We have thoroughly revised the manuscript to correct any misreferenced figure. We have also updated the quality of Supplementary Figures.

Lines 51 - 55 : "However, their diversity, function, and interactions with host proteins are largely unknown. This knowledge gap is largely due to the lack of reliable functional

annotations of TEs. In fact, most TE sequences in a genome are thought to be fossilized copies of once active elements and typically have lost all coding potential."
The authors must add relevant litterature citations that has addressed this subject. Such as : doi: 10.1038/s41576-020-0251-y

We now cite this and other literature (Maumus and Quesneville, 2014; Feschotte and Pritham 2007) that support the statement (Lines 56-63).

Line 72 : The authors should rewrite "contains 32,000 annotated TE" to "contains about 32,000 annotated TE".

Done

Reviewer #3 (Remarks to the Author):

The present manuscript describes a fascinating attempt to identify and characterise transposon-encoded products on an organism-wide scale in Arabidopsis, largely through plants defective in DDM1. As written, the manuscript implies that evidence for hundreds of TE-encoded proteins has been uncovered from proteomics analysis, that these proteins have extensive multimerization capacity, and that they are likely to be functionally integrated in host cellular processes.

These are intriguing implications. However they remain largely hypothetical, as they are based on in silico modelling of proteins that are implied from ORF-containing TE transcripts, most of which are not experimentally observed in the present study. This lack of broad-scale experimental validation of proteins undermines the study's overall impact.

For example, the section 'A hidden diversity of TE-encoded protein domains' describes numerous examples of protein structures modelled using AlphaFold2, from which well thought out and interesting interpretations are made. However it is unclear how many, if any, of these structures are associated with proteins experimentally identified from proteomics experiments. Similarly, it is unclear if any of the 18 proteins selected for AlphaFold-Multimer modelling (in the 'Structural modelling of TE multiprotein complexes' section) were uncovered from data associated with the present study, or if they were all sourced from references 42 and 43.

We thank the reviewer for this important critique. We agree that the original presentation did not sufficiently distinguish between (i) computationally predicted TE-encoded proteins derived from our transcript-informed TE gene annotation and (ii) the subset of proteins with

direct proteomic evidence. We have therefore revised the manuscript to clearly separate these two layers of evidence.

First, we clarify that proteins encoded by all TE-encoded ORFs (n = 6,835) identified in our TE transcript annotation were modeled using AlphaFold2 (Lines 306-308).

Second, for the multimer analyses, we focused on 18 TE-encoded proteins from 11 full-length elements from our transcript-informed TE annotation spanning both Class I and Class II superfamilies. These elements are strongly expressed in epigenetic mutants and have independent evidence of mobility (experimental mobilization assays or recent activity in natural populations–reported in references 42 and 43 in the original manuscript), enriching for TEs encoding functional transposition machinery. These TEs are indeed targets of active investigation in many laboratories worldwide, and we therefore believe that this subset would be of the highest relevance for the transposon community. We now state this rationale explicitly (lines 364-368) and provide the corresponding TE IDs in Supplementary Data 10.

Finally, we revised the wording throughout the Results and Discussion to avoid implying experimental validation where we present predictions. In particular, structural and interaction analyses are framed as hypotheses-generating.

To clarify what proteins were identified, the confidence of these identifications, and the source of the data from which the identifications were made (i.e. data collected from the present study or data sourced from reference 28), substantially more detail needs to be included in the supplementary information. Crucially, the criteria from which proteins have been identified also needs to be substantially tightened, and it is unclear how many proteins will remain after appropriate filtering criteria have been applied (more below).

We agree and have implemented two major changes. (i) We tightened protein identification criteria and explicitly defined a high-confidence TE protein set. (ii) We substantially expanded the supplementary information to provide the evidence required to assess each identification and its provenance.

Specifically, we now identify peptides using stringent q-value thresholds (see point 1 below), and define high-confidence TE proteins as those supported by ≥2 independent peptides per MS run. Proteins supported by a single peptide are now reported as putative and are not used for the main quantitative analyses and conclusions.

To improve transparency, we expanded the Supplementary Tables to include, for each TE protein: peptide sequences, precursor m/z and charge, run identifiers, confidence metrics

(scores/q-values), and the dataset source (new datasets generated here versus reanalyzed public datasets). We also updated the Methods to report the MaxQuant and DIA-NN parameters used.

Specific points that require addressing:

1. As written, the study has treated proteins identified from single peptides as bona fide TE-encoded products. This is a lax criterion for protein identification in a standard proteomics study. For a study designed to uncover novel proteins, which requires substantially higher levels of identification stringency, this criterion is inappropriate unless additional validation criteria are included (or unless proteins identified from single peptides are explicitly treated as putative and lacking in strong experimental evidence).

To address this, it is suggested that, at a minimum, only proteins identified from 2 or more unique peptides are considered identified.

If proteins identified from single peptides are to be included, it is strongly suggested that additional validation criteria for these individual peptides are required. One option would be to collect spectra from synthetic peptides and demonstrate equivalence to spectra collected from plant samples. Another would be to ensure that the single peptides from which proteins are identified meet stringent q-value thresholds across multiple independent proteomics datasets, and that these spectra cannot be otherwise explained by modified peptides uncovered from open modification searches. An example of such a pipeline can be found here:
https://www.sciencedirect.com/science/article/pii/S1535947624000094

We thank the reviewer for raising this key point. We agree that multi-peptide identification is the most robust methodology in the context of novel protein discovery. We therefore revised our analysis and presentation as follows.

1. We increased the stringency of peptide and protein-group calling by applying a conservative q-value cutoff of 0.001 at the protein-group level in DIA-NN (PG.Q.Value, Global.Q.Value, and Lib.PG.Q.Value), and we report these thresholds explicitly in the Methods (lines 802-805).

2. We now define a high-confidence set of proteins requiring ≥2 distinct peptides per protein, and we use this set for all subsequent analyses and statements about experimentally detected TE proteins (including Figure 3 and related text). Importantly, to avoid ambiguity due to homology between TE-encoded proteins and host proteins, we

excluded any peptides shared between TE proteins and host proteins and retained only TE-specific peptides for TE protein calling (lines 805-809; Figure 3; Supplementary Data 9).

3. To benchmark specificity and address the risk of false assignment in repetitive sequence space, we performed multiple orthogonal analyses:

-Negative-control benchmarking: We compared the identification of TE-proteins in LC-MS datasets derived from epigenetic mutants with strong TE reactivation versus wild-type samples, where TEs are generally repressed and show comparatively low expression levels. We found that high-confidence TE proteins are strongly enriched in epigenetic mutant datasets with robust TE transcription, and largely absent from matched wild-type datasets. Indeed, using the ≥2 unique peptide criterion (not shared with host protein groups), we identify 117 TE proteins in TE-reactivating mutant datasets versus 16 across all wild-type proteomic datasets (public and generated here) (lines 261-264; Supplementary Figure 11B; Supplementary Data 9), supporting that missidentification of peptides derived from TE-proteins is limited.
-Expression-consistency benchmarking: To estimate an empirical upper bound rate of false assignment, we queried our LC-MS/MS runs for matches to in silico-derived peptide sequences from TEs with no evidence of transcription in any sample (i.e. epigenetic mutants). Among the 15M candidate peptide sequences, we detected matches for only 78 peptides, leading to an empirical false-assignment rate of ~0.005%. Conversely, peptides assigned to TE proteins overwhelmingly map to TE genes with evidence of transcriptional activity (95.5%), further supporting their validity of the proteomic detections (Supplementary Figure 11).

Last, we have thoroughly revised the manuscript to ensure that the term ‘detected’ is exclusively used for proteins in the high-confidence set (≥2 unique peptides); all others are described as ‘putative and reported for completeness (Supplementary Data 8) but not considered for further analyses.

2. Using the present identification criteria, the authors state that 426 proteins derived from 372 TEs were identified. Information pertaining to the basis of these protein identifications is lacking. Currently, the only information available to the reader is found in Table S6, which simply includes a ‘Found in MS?’ column.

If criteria outlined by the leading proteomics journal, Molecular and Cellular Proteomics, are used, the protein identifications should be expanded to include the following: all peptide sequences assigned, precursor change and mass/charge, all modifications observed, score(s), and count of the number of distinct peptide sequences assigned to each protein.

The employed MaxQuant and DIA-NN search parameters should also be included in the methods.

We agree and have expanded the evidence reporting. We now provide, for each TE protein identification: (i) all assigned peptide sequences, (ii) precursor m/z and charge, (iii) confidence metrics (scores and q-values), and (iv) the number of distinct unique peptides supporting each protein, together with the run identifiers and dataset provenance. We also added a column indicating the number of peptides supporting the identification of these proteins (Supplementary Data 8 and 9). In addition, we restricted identifications to TE-specific peptides by removing any peptide sequence that could be assigned to both TE-encoded and host-encoded protein groups (lines 805-809).

We updated the Methods to report the DIA-NN version and the key search parameters (enzyme specificity, missed cleavages, fixed/variable modifications, FDR control, library settings, match-between-runs settings where applicable, and protein grouping rules). Raw files and processed outputs are now available in PRIDE under PXD067692 and PXD066483, and the full parameter files are deposited alongside these records.

3. Moreover, from Table S6, only 46 proteins listed as True in the 'Found in MS?' column contain a value in the te_id column. This is far below the 426 stated by the authors in the body text. Is this because most of the protein identifications came from data sourced from reference 28 rather than the present data? If so, can the authors comment on this result, particularly in relation to the fact that the reference 28 data does not come from plants defective in DDM1?

We thank the reviewer for flagging this inconsistency. This was due to an error in the originally submitted Supplementary Table: the initial Table S6 (now in Source Data related to Fig. 3A) reflected only a subset of early DDA runs and did not include the full set of DIA acquisitions and additional mutant datasets, nor did it provide consistent TE identifiers across all entries.

We have now corrected this in three ways:
1. Updated Supplementary Data and figure: Table S6 (now found in the Source Data file, as Source Data Fig. 3A) and the corresponding figure have been replaced with complete results including all DIA runs and all experimental conditions, with consistent TE identifiers. We now specify that this data is exclusively gathered from our datasets.

2. We now report protein detection using two tiers. Under the revised, stringent definition of high-confidence (≥2 unique peptides; q ≤ 0.001), we identify 135 TE proteins with robust proteomic support across the analyzed datasets, with the vast majority of them (117 out of

135) being identified in proteomic data from our epigenetic mutant samples (lines 261-267, Figure 2A, and Supplementary Figure 11B). We also provide a table (Supplementary Data 8) with a larger putative set of proteins supported by a single peptide, clearly labeled as such, which is not used for the core quantitative conclusions.

3. We revised the main text and figure legends to ensure that reported numbers consistently refer to the high-confidence set of proteins, and to avoid implying experimental validation where we present predictions.

If the above protein identification thresholds are applied, it is entirely possible that the total number of genuinely identified proteins could be in the single digits when only searching proteomics data from the present study. While validation of any TE-encoded protein would be of interest, identification of so few TE-encoded proteins would unfortunately appear to be well below the ambition and scope of the manuscript as currently written.

We appreciate this concern. After applying the reviewer-recommended stringent thresholds and performing additional MS acquisitions and re-analyses, we confirm our conclusions do not rely on single-digit identifications. Using the conservative high-confidence definition above, we identify 135 TE-encoded proteins with robust proteomic support across the analyzed datasets, with the vast majority (117/135) detected in epigenetic mutant samples generated in this study (*ddm1*).

As explained above, two orthogonal benchmarks support that this set is not inflated by false assignments: (i) these high-confidence TE proteins are strongly enriched in epigenetic mutant samples with robust TE transcription and are almost completely absent from wild-type samples, and (ii) peptide assignments overwhelmingly map to expressed TE-genes rather than non-expressed TEs. Together, these analyses indicate that the number of genuine TE protein identifications in our study is well above single digits while remaining conservative and specificity-controlled.

At the same time, we agree with the reviewer that proteomic validation cannot match the scale of transcript-informed ORF predictions. We therefore revised the manuscript to separate (a) the predicted TE-encoded proteome and structure-informed hypotheses from (b) the subset of experimentally detected TE proteins. All quantitative analyses and claims regarding proteome-based detection now rely on the set of 135 high-confidence TE-proteins, while the single-peptide set is reported as putative and provided for completeness (Figure 3A; Supplementary Data 8; lines 261-264).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

Thanks to the authors for well addressing my concerns. I have no further comments.

We are pleased that all concerns of reviewer 1 have been well addressed.

Reviewer #2 (Remarks to the Author):

The authors have successfully addressed all my comments. I have no further concerns.

We are pleased that all concerns of reviewer 2 have been successfully addressed

Reviewer #3 (Remarks to the Author):

The authors have thoroughly addressed the comments regarding the proteomics data analysis. They have made well considered alterations to the both the main body of the manuscript and the supplementary information, which reflect the changes to the reported number of protein identifications.

A minor observation is that there now appear to be some small inconsistencies in the numbers of high confidence proteins that are reported. In the text of the revised manuscript, the authors report 135 identified proteins, 115 of which were found in ddm1 plants. From my assessment of Supplementary Data 9, I can count 139 identified proteins, 107 of which were found in ddm1 plants from the present study.

Similarly, I count 196 proteins in Supplementary Data 8, which when added to the above 139 proteins makes 335 proteins, which is below the 243 proteins stated in line 260.

We thank the reviewer for carefully checking the Supplementary Data and for pointing out these apparent discrepancies. We have revised the text and tables to make the counting rules explicit and to ensure consistency across the manuscript.

High-confidence TE proteins (Supplementary Data 9). In the main text we first report 135 high-confidence TE proteins (≥2 TE-specific peptides within a sample) based on the *ddm1* and wild-type datasets (Results, lines 260–263). We then include additional proteomes from *ddm1rdr2* and *ddm1rdr6*, which increases the total number of high-confidence TE proteins to 139 (Results, lines 285-287). This matches Supplementary Data 9 (139 total). We also clarify that 115/135 high-confidence TE proteins are detected in ddm1 in the initial

ddm1/WT comparison (Results, lines 288-291), and we have made the “sample” annotation in Supplementary Data 9 explicit so that ddm1 detections can be counted unambiguously.

Putative vs high-confidence sets (Supplementary Data 8 vs 9). Supplementary Data 8 lists the putative TE proteins (including single-peptide support), whereas Supplementary Data 9 lists the high-confidence subset. Consequently, these tables are not additive, and summing counts across the two tables leads to double-counting of the high-confidence proteins that are already included in the putative set. We have updated the headers and Methods to clarify this hierarchy and avoid ambiguity (Methods, lines 819-822).

Finally, in the abstract, the authors state that proteomics analyses confirm the production of hundreds of TE-encoded proteins. This would now be more appropriately stated as 'over one hundred'.

We agree that “hundreds” should be revised. We now state “over a hundred of high-confidence TE-encoded proteins”

Aside from the above minor observations, the overall proteomics analysis has been substantially improved and appropriately reflects the underlying data.

We are pleased to hear that the concerns raised in the proteomics are now solved and that the reviewer considers the results to be sound.

Reviewer #4 (Remarks to the Author):

Overall the work appears very interesting with a number of valuable observations and validations. However, I have several comments that should be taken into consideration as for the usage of AlphaFold for global analyses. In this review, I will specifically focus on the sections of the article dedicated to the use of AlphaFold to characterize molecular structures and interactions. In the manuscript, three main applications are presented i) the use of AlphaFold for structural classifications, ii) for stoichiometry estimation of homo-oligomers and iii) for interactome screening.

We thank Reviewer #4 for the positive evaluation of our work and for the constructive comments. We agree that strengthening the methodological description and robustness of the AlphaFold-based analyses will improve the manuscript, and we have revised the text, figures, and supporting data accordingly as detailed below.

Structural classification

The use of Foldseek to identify potential structural similarities is appropriate. To assign similarities to the predicted structures of TE ORFs, the authors employed a relatively relaxed E-value threshold of 0.01, which is higher than the default value of 0.001. This choice should be justified and the manuscript would benefit from a more explicit statement clarifying this relaxed parameter selection.

The choice of an E-value threshold of 0.01 (instead of the default 0.001) was intentional. Our aim was to identify global structural similarity between TE-encoded domains and reference proteins, and we therefore required substantial alignment coverage by retaining only hits spanning ≥80% of the query (marked by the parameters -c 0.8 and --cov-mode 2). These parameters differ from the default Foldseek behavior, which does not enforce query coverage constraints and can retrieve short local matches that are more prone to spurious similarity (e.g., microhomology). In this context, using a slightly relaxed E-value increases sensitivity to more distant homologs while the stringent coverage filter limits short, likely artifactual hits. We now state and justify these parameter choices explicitly in the Methods section (lines 865-868). In addition, for downstream manual curation of key TE–host evolutionary relationships, we now report the corresponding Foldseek E-values explicitly in Figure 4.

The authors also used LDDT metrics to estimate structural distances between TE ORFs and other proteins which is somewhat unconventional. Large-scale structural clustering is more commonly performed using TM-scores (see for example 10.1038/s41586-023-06622-3). While LDDT is primarily a local measure of structural accuracy, TM-scores are better suited for assessing global structural similarity in a properly normalized manner. The manuscript does not appear to reference key studies describing or benchmarking LDDT, and it would be useful to include such citations, as well as studies comparing different structural similarity metrics. Although the authors would likely reach similar conclusions for most TE ORFs using TM-scores, the assignment of specific cases could be improved. Validating their results using TM-scores computed with standard tools such as US-align or GTalign would align the analysis with common practices and strengthen the robustness of their conclusions.

We agree that TM-score is the standard metric for assessing global structure similarity, yet its calculation is computationally more expensive than that of LDDT (10.1038/s41587-023-01773-0). LDDT-based distances were therefore used as a computationally efficient proxy at the scale of hundreds of millions of comparisons. We additionally applied restrictive filters on reciprocal coverage, where only hits that span 90% of the query length and 90% of the target length were retained. This approach allowed us

to scale up and perform 897 million comparisons. We have clarified the use of LDDT over TM-score in the Methods section (lines 869-871). We also included a reference for a study showing the correlation of TM-scores and LDDT scores (10.1093/bioinformatics/bty760).

Importantly, for the specific TE-protein homology cases highlighted in Figure 4 (VANC, ASY2, etc), we performed manual curation and reported the RMSD metric to illustrate structural similarity. To further support these observations, we have now validated the key structural homology assignments using TM-scores computed with USalign, and we report these TM-scores in the revised Figure 4. We additionally verified the same relationships using GTalign (web server) as an orthogonal check. These validation steps are now described in the Methods (lines 872-874).

Stoichiometry estimation

The authors performed a careful analysis of the potential homomerization properties of the detected TE-encoded proteins. They used the ipTM score as a metric to estimate the most likely stoichiometry of the predicted homomultimers. A recent benchmarking study (10.1101/2025.03.10.642518v2) has evaluated the potential of ipTM for predicting protein oligomeric states and was recently updated. The authors could refer to this work to support their methodological choice.

As explained in the next section, in supplementary material, the author should provide a detailed list of the scores calculated for these dimers and oligomers, especially the average of the ipTM scores over the 5 generated models which account for the robustness of the prediction. This important for readers to judge the reliability of the predictions. For instance, testing the interaction between VANB and VANC either by AF2 or AF3 did not convince me from the scores that an interaction can be reliably predicted between the two.

We thank the reviewer for these comments. We now cite (line 383) a recent benchmarking study evaluating the use of ipTM for inferring oligomeric state (10.1101/2025.03.10.642518v2) and clarify our use of ipTM in the Methods (line 906). In the previous version, the prediction of TE-protein heterodimerization in Figure 5A was based on the number of close-by residues. In the revised manuscript, we now display the mean ipTM across models together with its standard deviation, rather than relying on a single best model. We similarly updated panels that previously reported only the best-model ipTM for higher-order assemblies to report mean ipTM across models, consistent with Figure 5A (Fig. 5B and 5D; Methods, line 906). The per-model ipTM scores for dimers and oligomers, together with the corresponding summary statistics, are now provided in Source Data for Figures 5A, 5B and 5D.

We also added a new inset in Figure 5A showing the distribution of mean ipTM values for 1M random sets of 18 cellular protein homodimers, providing a reference expectation. This analysis shows that TE proteins more frequently yield high-confidence homodimer predictions than expected from random cellular protein sets, supporting an increased propensity of TE proteins to self-assemble.

Finally, we agree that inferring oligomeric states from predictions is not definitive. We therefore revised the Results/Discussion to frame stoichiometry assignments as working hypotheses rather than proof, and we explicitly highlight that experimental validation will be required (lines 544-547)

They could also note that distinguishing oligomeric states is not always straightforward and has been addressed using alternative strategies, such as in 10.1016/j.cell.2024.01.022. In that study, the authors compared the geometry of a predicted dimer in isolation with that observed within cyclic oligomers, showing that the assembly whose geometry best matches the free dimer is likely to represent the correct oligomeric state.

We thank the reviewer for highlighting that inferring oligomeric states from structure prediction is not always straightforward and that alternative strategies have been proposed, including comparing dimer geometry in isolation with that embedded within cyclic oligomers (10.1016/j.cell.2024.01.022). In our study, we implemented cross-stoichiometry comparisons of ipTM values to infer the most likely oligomeric assemblies, an approach supported by recent benchmarking (10.1101/2025.03.10.642518). Notably, this framework identifies pentameric and hexameric configurations for the GAG protein as the most likely assemblies (Supplementary Figure 16), consistent with recently resolved structures of related GAG proteins forming viral like particles structures (10.1016/j.cell.2025.02.003). Nevertheless, we agree that ipTM-based inference does not constitute definitive proof of stoichiometry; we therefore revised the Results and Discussion to explicitly present these assignments as working hypotheses and to emphasize the need for experimental validation (lines 390-392 and lines 544-547)

Interactome screening

The section describing the AlphaFold2-Multimer screen lacks sufficient methodological detail and would benefit from substantial clarification. Several elements presented in Figure 6 are not adequately described in the Methods section, which makes it difficult to assess the robustness of the analysis. In addition, the web portal provides only limited information regarding the predicted scores. As illustrated by resources such as Predictomes

(https://predictomes.org/), access to multiple complementary metrics is essential to properly evaluate the reliability of high-scoring interactions.

Relying solely on the maximum ipTM score is problematic in large-scale interactome screens and can lead to a substantial number of false positives, as discussed in 10.1016/j.molcel.2025.01.034. This issue can arise for multiple reasons. For instance, stochastic effects may lead to spurious interactions: for large prey proteins, a single model among the five generated may display a high ipTM score, while the interaction is not consistently reproduced across models, indicating low reliability.

Second, certain bait proteins may exhibit non-specific binding behavior (for example, due to amphipathic helices embedded in disordered regions) which can artificially inflate interaction scores. The bait VANB appears to fall into this category. For instance, when I attempted to reproduce two randomly selected interactions reported in the web portal (AT3G54230.2 and AT5G50780.1, with reported ipTM values of 0.75 and 0.73, respectively), the distribution of ipTM scores across the five models was markedly inconsistent, supporting concerns about robustness. For the five models generated between VANB and AT3G54230.2, the ipTM scores were [0.52, 0.32, 0.31, 0.23, 0.20]. In contrast, for VANB and AT5G50780.1, the scores were [0.70, 0.58, 0.55, 0.32, 0.23]. The fact that only one model yields an ipTM > 0.5 suggests that the interaction with AT3G54230.2 is highly unlikely, whereas for AT5G50780.1, the presence of three models with ipTM > 0.5 may indicate a more reliable signal. However, AT5G50780.1 contains a coiled-coil domain that can pair with that of VANB, while VANB itself is predicted to form a homodimer through this same coiled-coil, raising the possibility of non-specific or competing interactions.

To mitigate these issues, it would be essential to report not only the maximum ipTM score, but also the full set of scores across the five models, along with their average. This typically provides a more reliable indicator of interaction robustness.

We thank the reviewer for raising this important point. We agree that relying on a single maximum ipTM value can inflate false positives in large-scale screens, and that insufficient reporting of model-to-model variability limits interpretability. We therefore revised both the analysis and its reporting. First, we expanded the Methods to fully describe the benchmarking procedure and the construction of positive and matched negative reference sets (Lines 919-923). Second, we now report the per-model ipTM values and their mean for each TE–host pair in the Source Data associated with Figure 6, allowing readers to directly assess robustness across models. Third, we implemented ipSAE as a complementary interface-focused metric and now use it as the primary criterion for calling interactions, based on benchmarking against experimentally supported Arabidopsis

heterodimers (Fig. 6B; Methods, lines 919-927). Finally, the web portal has been updated to display these score summaries and to provide PAE maps for high-scoring TE–host pairs to facilitate assessment of confidence and interface consistency.

We also note in the Results that for some baits TE–host interfaces may overlap with homodimerization surfaces, raising the possibility of non-specific or competing interactions (Lines 438-441), and we provide the corresponding interface annotations (Supplementary Fig. 19 and associated Source Data).

In addition, the list of interface residues, as well as their potential overlap with homodimer interfaces for high-scoring pairs, should be reported. Further strengthening the analysis would require the inclusion of complementary metrics, such as PAE maps (for example, provided in GIF format) and alternative interface-focused scores, including 10.1101/2025.02.10.637595 (IPSAE) or 10.1093/bioinformatics/btaf107 (actifpTM). Providing this information through the web portal would substantially enhance the practical utility of the predictions and help reduce biases arising from the use of overly permissive ipTM thresholds in Figures 6D and 6F.

We thank the reviewer for emphasizing the importance of reporting additional metrics and interface-level information. To improve the reporting of our results, we now present the homodimerization and TE-host PPI surfaces in Supplementary Figure 19, and modify the results section to discuss the possibility of non-specific or competing interactions (lines 438-441). The list of interface residues and their potential overlaps with homodimer interfaces are now provided as Source Data associated with Supplementary Figure 19. We also include the PAE maps of high-scoring TE-host PPI pairs on the website.

Following the reviewer's suggestions, we have implemented ipSAE both in the benchmarking and in the PPI screening analyses. In our benchmarking, ipSAE discriminates experimentally supported pairs from matched random pairs more consistently than ipTM, and we therefore use ipSAE as the primary metric for calling interactions in the revised manuscript (Fig. 6B–C; Results and Methods, lines 424-427 and 924-927).
Using ipSAE, we found that the main conclusions presented in Figure 6 remain unchanged (ipSAE vs ipTM) and are robust to threshold variation (ipSAE 0.1-0.25). To reflect these changes, we moved several panels from the previous Figure 6 to a new Supplementary Fig. 18, and we updated Figure 6 to present the corresponding analyses based on ipSAE, with a description of the concordance between metrics and thresholds included in the Results section (lines 432-433).

Finally, Figures 6A and 6B refer to a systematic benchmarking procedure used to define the ipTM threshold for calling positive interactions, yet this procedure is not described in the

Methods section, which represents by itself a concern. Key aspects remain unclear, including how the 600 positive heterodimers were selected and how random pairs were generated. Beyond this, the main issue is that just using random pairs is not a valid approach as carefully stated in 10.1016/j.molcel.2025.01.034. The authors could draw inspiration from this benchmarking study to improve this analysis. In particular, it seems important to restrict both positive and negative sets to comparable subsets (e.g., "C+" categories), as balancing these classes typically reveals that a higher ipTM threshold is required for reliable discrimination in large-scale interactome studies.

We apologize for the lack of methodological detail in the previous version. In that version, the negative set consisted of protein pairs sampled at random from the complete Arabidopsis thaliana proteome, and we agree that this approach could generate an imbalanced benchmark.

We have now revised both the benchmarking design and the Methods description. Specifically, we expanded the positive reference set by extracting experimentally supported heterodimers from Arabidopsis Interactome 2.0 (10.3390/plants11030350; available at https://www.arabidopsis.org/download/list?dir=Proteins%2FProtein_interaction_data%2FInteractome2.0), retaining only pairs supported by experimental evidence (Y2H).This yields 1,596 positive protein pairs, and the selection criteria are now described explicitly in the Methods (Lines 919-920).

To ensure that the negative set is directly comparable to the positive set, we now generate the negative set by randomly pairing proteins drawn from the same pool of proteins present in the positive set, yielding 1,596 matched random pairs. This strategy controls protein composition, avoiding potential biases introduced by sampling from the full proteome. In addition, benchmarking and thresholds are now based on mean scores across models used per pairs. A full description of this approach is now presented in the methods section (lines 920-921).