bio-protocol

Open Access

Supplementary information for Predicting Proteomics Data From Transcriptomics Data With MintOmics

Akshata Hegde[1, 2], Yanli Wang[1, 2], Frimpong Boadu[1, 2], Ryan M. Finnerty[1, 3], Wipawee Winuthayanon[1, 3] and Jianlin Cheng[1, 2, *]

[1]Roy Blunt NextGen Precision Health, University of Missouri, Columbia, MO, USA
[2]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA
[3]Department of Obstetrics, Gynecology and Women's Health, School of Medicine, University of Missouri, Columbia, MO, USA
*For correspondence: chengji@missouri.edu

Reference directory and command sequence

```
mintomics/
  requirements.txt
  README.md
  Dataset/
    Collaboration_data.csv
    Labels_orig.csv
    Mouse_TFs1
    genetoprotein.csv
    Data_cpm/
    Labels_proc_log10_minmax/
    Diff_data/
    Diff_labels/
  src/preprocess/Data_preprocessing.py
  src/dataset/PrepareDataset.py
  src/dataset/prepare_test_data.py
  src/model/Model.py
  Training.py
  Inference.py
  Result_analysis.py
  verify_workflow.py
  Trainings/<save_dir>/
 intermediate_files/
```

1. Run Data_preprocessing.py from Dataset/ to construct the RNA and protein matrices. Install the recorded Python environment from requirements.txt before execution.
2. Use the canonical gene-to-protein mapping file Dataset/genetoprotein.csv (tab-delimited, two columns, no header) and compare the retained mapping counts with Table S3. For another organism, provide an organism-appropriate mapping.
3. Configure the paired-unit file lists, view counts, mask fraction, target-abundance threshold, output dimension, and random seed. The --abundance_threshold value is explicitly set for separate 0.6 and 0.8 runs.

4. Run Training.py from Dataset/ with --batch_size 1, --abundance_threshold 0.6 or 0.8, and seed 42; retain the checkpoint with the lowest validation loss.
5. Run Inference.py with the selected checkpoint, an unmasked held-out view, and the same target-abundance threshold used during training.
6. Export the prediction and TF-protein attention tables and prepare the contrast-specific intermediate files before executing Result_analysis.py.

Reproducibility: Python 3.11; PyTorch 2.6.0+cu124; Lightning 2.6.6; torchmetrics 1.9.0; AlmaLinux 8.10; CUDA 12.4; NVIDIA driver 570.211.01; two NVIDIA A100 80GB PCIe GPUs; seed 42; source revision GitHub release (tag: mintomics): Release Mintomicsv1.0, https://github.com/BioinfoMachineLearning/mintomics/releases/tag/mintomics (accessed October 2, 2026). The pinned package set is supplied in requirements.txt.
Tensor verification (batch size 1): input (1, 21190, 7); attention matrix (1, 21190, 21190); pooled representation (1, 7); protein logits (1, 3060); masked gene-to-protein index array (2, 400).
Masking and scaling: TF flags are assigned by membership while preserving the ordered RNA gene list; each masked view starts from np.array(dat, copy=True). The six expression columns undergo within-matrix 99.9th-percentile clipping/rescaling, and the seventh TF indicator remains binary (0/1) for development, validation, diagnostic, and blind-inference inputs.
Representative outputs: training checkpoints are written beneath Trainings/<save_dir>/; figures and companion data examples are provided under figures/representative/. On the reported A100 system, the measured epoch runtime was approximately 31 min (about 6 iterations/s), and a held-out forward pass peaked at approximately 4.06 GB of GPU memory. These measurements depend on the hardware and input dimensions.

Minimal input examples
The examples below show the required orientation and identifiers. The number and names of RNA measurement columns may differ between studies, but their order must be identical across paired units.

```
# Dataset/Data_cpm/Data_example.csv
gene_id,IA_1,IA_2,IA_3,IU_1,IU_2,IU_3
Actb,12.183,12.004,12.311,11.982,12.106,12.221
Pax2,4.230,4.511,4.407,5.012,4.882,5.106
# Dataset/Labels_proc_log10_minmax/Labels_example.csv
Accession,normalized_abundance
P60710,0.9342
Q9Z2V5,0.4127
```

Note: Mouse_TFs1 is the filename in the original supplementary inventory; the main manuscript uses Mouse_TFs descriptively. The configured loader must point to the actual deposited TF catalog. Original source/release/licence metadata for this file have not been supplied and therefore are not assigned here.

```
# Dataset/genetoprotein.csv (tab-delimited; no header)
ACTB   P60710
PAX2   Q9Z2V5

# pairing.tsv (illustrative; tab-delimited)
```

paired_unit_id	rna_file	protein_file	biological_group	biological_source	replicate_structure	batch	role
control	Data_control.csv	Labels_control.csv	control	mouse oviduct	six RNA regional measurements; condition-level protein vector	study_batch	development

2.5_dpc	Data_2_5preg.csv	Labels_2_5preg.csv	2.5_dpc	mouse oviduct	six RNA regional measurements; condition-level protein vector	study_batch	held_out

Independent biological units in this reference application: three development conditions and one held-out condition. The 12,000/3,000/2,000 computational views are augmented observations, not independent biological replicates.

Table S1. Reference input files and expected dimensions

Table S1. Deposited files used by the reference analysis. Use the listed dimensions to follow the worked example; for a new dataset, record the corresponding study-specific dimensions.

File group	Repository path	Expected dimensions	Expected property
Raw RNA	Dataset/Collaboration_data.csv	27,179 rows × 54 RNA samples before filtering	Gene identifiers in first column
Processed RNA	Dataset/Data_cpm/Data_control.csv	21,190 genes × 6 features	Finite; ordered gene IDs
Processed RNA	Dataset/Data_cpm/Data_0_5preg.csv	21,190 × 6	Finite; ordered gene IDs
Processed RNA	Dataset/Data_cpm/Data_1_5preg.csv	21,190 × 6	Finite; ordered gene IDs
Processed RNA	Dataset/Data_cpm/Data_2_5preg.csv	21,190 × 6	Finite; ordered gene IDs
Raw proteomics	Dataset/Labels_orig.csv	3,469 rows; 3,060 retained proteins	Stable protein accessions
Processed proteomics	Dataset/Labels_proc_log10_minmax/Labels_*.csv	3,060 proteins × 1 abundance	Finite; range [0,1]
TF catalog	Dataset/Mouse_TFs1	1911 entries	1,911 observed TFs; generate flags in RNA gene-row order (do not iterate a set)
Mapping	Dataset/genetoprotein.csv (canonical loader path)	2,789 gene-protein records	2,406 usable pairs

Table S2. Reference pairing and data allocation

Table S2. Condition-level allocation in the reported MintOmics analysis. Generated views inherit the role of their source condition and are not counted as biological samples.

Paired condition	Biological role	Training views	Validation views	Masked inference views	Unmasked inference
Control/estrus	Development	4,000	1,000	0	No
0.5 dpc	Development	4,000	1,000	0	No
1.5 dpc	Development	4,000	1,000	0	No
2.5 dpc	Held-out condition	0	0	2,000	One complete view
Total	Four paired conditions	12,000	3,000	2,000	One

Table S2.1. Pairing-table schema and representative control-condition entry

Column name	Data type	Example: Control condition
paired_unit_id	String	control
rna_file	String (file path)	Dataset/Collaboration_data.csv
protein_file	String (file path)	Dataset/Labels_orig.csv
biological_group	String	control (estrus)
biological_source	String	Mouse oviduct
replicate_structure	String	Six regional RNA measurements; one condition-level protein vector
batch	String	batch_01*
role	Categorical string	development

Note: The role field accepts development or held_out. All observations belonging to the same biological group must be assigned to the same analysis partition to prevent information leakage. The RNA and protein file paths identify the combined

source assay tables; condition-specific measurement columns are selected using the corresponding biological-group labels. The batch identifier is illustrative and should be replaced with the actual recorded batch information.

Table S3. Reference identifier, mapping, and masking values

Table S3. Expected mapping values for commit 0119209. Recalculate these values whenever an input dictionary or mapping resource changes.
Figure S3 provides the representative held-out threshold 0.6 prediction diagnostic, including the probability distributions and mapped-pair confusion matrix (TN = 860, FP = 18, FN = 205, TP = 1,323). Table S5 shows the rounded reference performance values reported in the main manuscript.

Quantity	Expected value	Calculation
Genes per processed condition	21,190	Rows in each Data_cpm file
Protein outputs	3,060	Rows in each normalized protein file
TF catalogue	2,133	Unique uppercase entries in Mouse_TFs1
Observed TF genes	1,911	Ordered RNA gene list intersect TF catalog
Mapping records	2,789	Rows in genetoprotein.csv
Mapped genes in RNA	2,671	Mapping genes present in RNA dictionary
Mapped proteins in proteomics	2,515	Mapping accessions present in protein dictionary
Jointly usable pairs	2,406	Mapping rows present in both dictionaries
Masked genes per reference view	400	Floor (0.15 × 2,671) sampled from usable mapped indices
TF flag assignment	Ordered 21,190-gene list	Catalog membership preserves the original expression-matrix row positions
Mask generation	Fresh source-matrix copy for every view	np.array (dat, copy=True) before zeroing selected rows
Input scaling	Six expression features, 99.9th-percentile clip-and-rescale	Keep the seventh TF feature binary (0/1) in every loader
Target thresholds	0.6 and 0.8, in separate model runs	Set --abundance_threshold explicitly during training and inference

Table S4. Reference model and training parameters

Table S4. Parameters required to reproduce the reported MintOmics configuration. A new application may change dimensions but must report the substituted values.

Component	Reference setting
Input tensor	B × 21,190 genes × 7 features
Input features	Six regional RNA measurements plus one TF indicator
Transformer	One active encoder layer; model dimension 7; one head; feed-forward dimension 32
Regularization	Dropout 0.5; residual connections; layer normalization
Pooling	Mean across genes
Classification head	7 -> 896 -> 224 -> 3,060 logits; ReLU between linear layers
Masked supervision	400 matched protein positions per training/validation view
Loss	BCEWithLogitsLoss
Optimizer	Adam; learning rate 0.0001
Scheduler	ReduceLROnPlateau; factor 0.1; patience 20; epsilon 1×10^{-10}
Training control	Maximum 100 epochs; early-stopping patience 30; save top 10 by validation loss
Data loading	Batch size 1; one worker; shuffled training; ordered validation/test
Reproducibility	Seed 42; deterministic cuDNN on; cuDNN benchmark off
Batch-one input/attention	(1, 21190, 7) / (1, 21190, 21190)
Batch-one pooled / output	(1, 7) / (1, 3060)
Mapped mask index	2 × 400 integer indices; first row genes, second row paired proteins
TF flag and scaling	Ordered gene-list membership; six RNA expression columns clipped/scaled at 99.9th percentile; TF flag remains 0/1
Masking independence	Create a new np.array(dat, copy=True) before zeroing each sampled row set
Target-abundance threshold	--abundance_threshold 0.6 and 0.8 run separately; sigmoid probability cutoff 0.5
Pinned environment	Python 3.11; torch 2.6.0+cu124; Lightning 2.6.6; torchmetrics 1.9.0; seed 42

Table S5. Representative held-out prediction results

Table S5. Reported prediction metrics for the unmasked 2.5 dpc condition. The target-abundance threshold changes the target vector; the model probability cutoff is 0.5.

Target threshold	Positive targets	Accuracy	F1	Precision	Recall
0.6	1,887 in deposited 2.5 dpc vector	0.907	0.922	0.986	0.866
0.8	128 in deposited 2.5 dpc vector	0.983	0.778	0.972	0.648

Table S6. Required machine-readable output fields

Table S6. Minimum fields for prediction and interpretation outputs. Store these as long-form TSV or CSV files so every value remains traceable to a biological unit, model, and identifier dictionary.

Output file	Required fields
protein_predictions.tsv	paired_unit_id, protein_accession, mapped_gene_id, observed_state, logit, probability, predicted_state, target_threshold, probability_cutoff, seed, checkpoint
attention_rankings.tsv	paired_unit_id, protein_accession, tf_gene_id, raw_attention, rank, top_k, diagonal_removed, seed, checkpoint
differential_transcripts.tsv	contrast, gene_id, log2_fold_change, statistic, p_value, adjusted_p_value, direction
differential_proteins.tsv	contrast, protein_accession, effect_size, statistic, p_value, adjusted_p_value, direction
integrated_evidence.tsv	contrast, protein_accession, tf_gene_id, attention_rank, transcript_evidence, protein_evidence, evidence_status
enrichment_results.tsv	contrast, input_set, library, library_version, term_id, term_name, overlap, p_value, adjusted_p_value, genes, analysis_date
Tfs_allprot_{control,0_5,1_5,2_5}.csv	Condition-specific TF-protein attention intermediate; retain gene/protein identifiers and attention values as expected by Result_analysis.py
Siggenebasedprotlist_TCT{0.5,1.5,2.5}.csv	Contrast-specific significant-gene/protein intermediate consumed by Result_analysis.py
AllTfs_diffcodinggene_<condition>.csv	Integrated contrast-specific TF/influenced-protein output for enrichment; use the corresponding condition labels

Supplementary figures

Figures S1–S5 show representative masking, training, held-out prediction, attention-based prioritization, and pathway-enrichment outputs. Each figure is described in its legend; the corresponding source files are maintained with the reproducibility materials.

Figure S1. Independent transcript masking and mapped protein indices

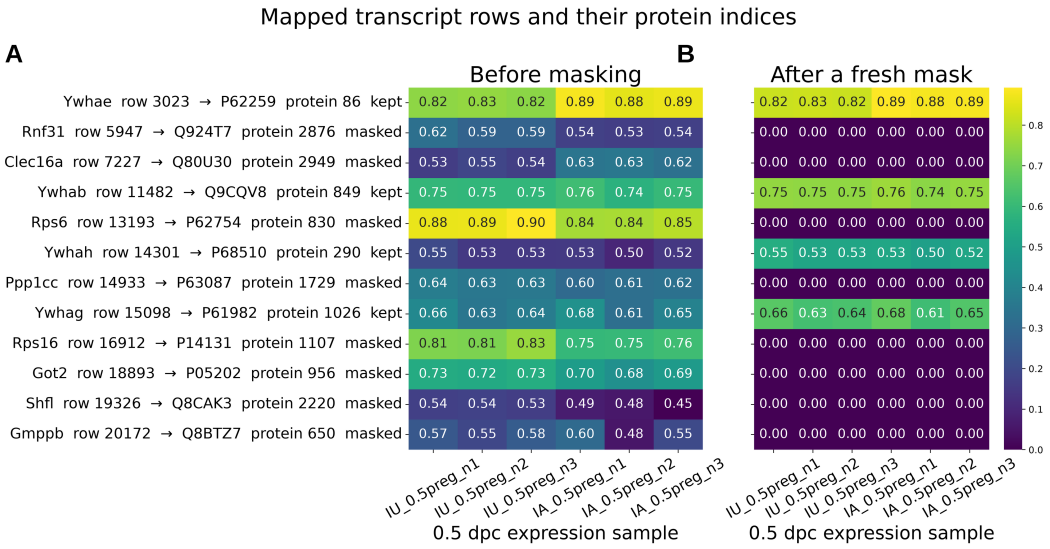


Figure S1. Representative independent masking of mapped transcriptomic features. The heatmaps show RNA-expression values for selected mapped transcript rows before masking (left) and after applying a newly sampled mask to a fresh copy of the original condition matrix (right). For each masked gene, all six displayed RNA measurements and the transcription-factor indicator in the model input are set to zero. The paired protein identifiers and indices are retained to select the supervised protein-abundance targets. The small subset shown illustrates the procedure; 400 mapped gene rows are masked per reference training view.

Figure S2. Training and validation performance of MintOmics

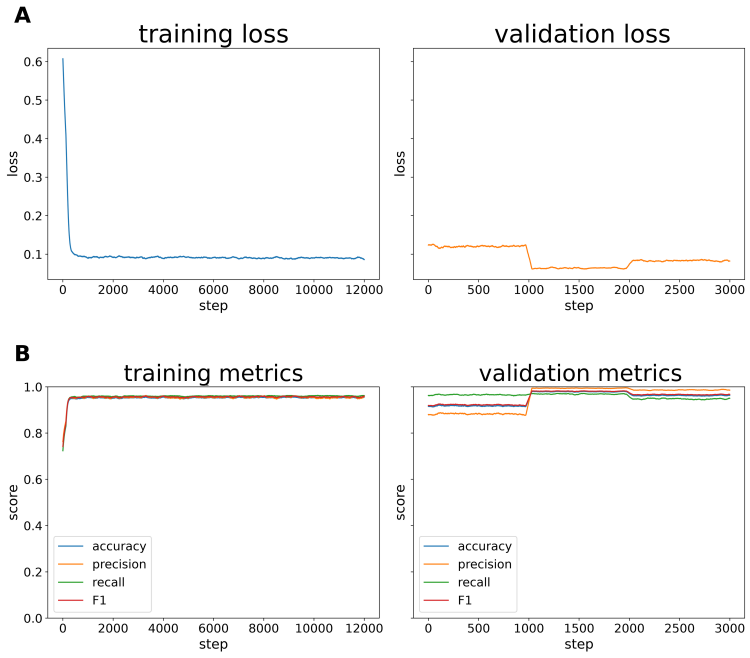


Figure S2. Training and validation performance of MintOmics at the protein target-abundance threshold of 0.6. (A) Training and validation loss plotted against optimization step. (B) Accuracy, precision, recall, and F1 trajectories for the training and validation views. The curves provide representative model-fitting and convergence diagnostics; computational masking views are not independent biological replicates.

Figure S3. Held-out protein-abundance prediction performance

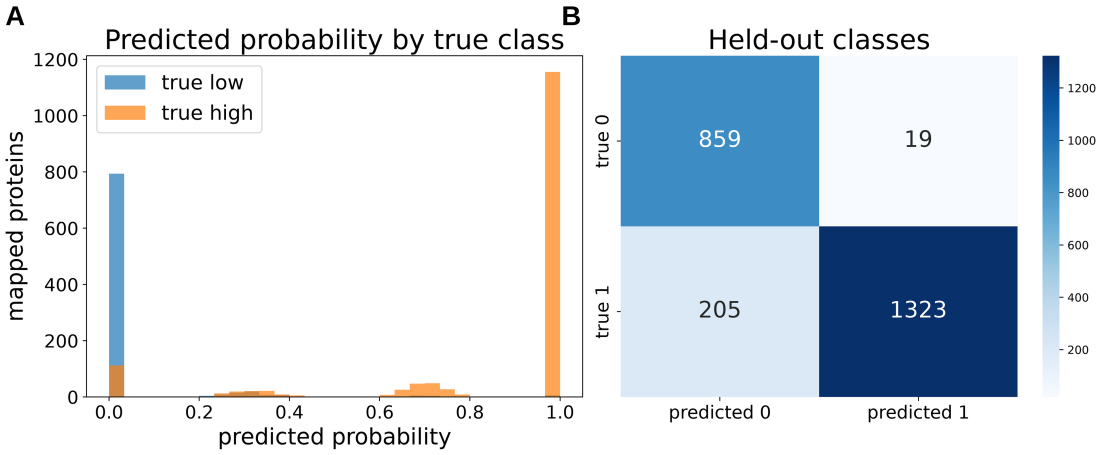


Figure S3. Protein-abundance classification in the unmasked held-out 2.5 days post coitus (dpc) condition. (A) Distribution of model-predicted probabilities for observed low-abundance (class 0) and high-abundance (class 1) proteins. (B) Confusion matrix for the 2,406 jointly usable gene–protein pairs at target-abundance threshold 0.6 and prediction-probability cutoff 0.5 (true negatives = 860, false positives = 18, false negatives = 205, true positives = 1,323). The protein target-abundance threshold determines the reference class labels and is distinct from the probability cutoff applied to model predictions.

Figure S4. Attention-based transcription factor and protein prioritization

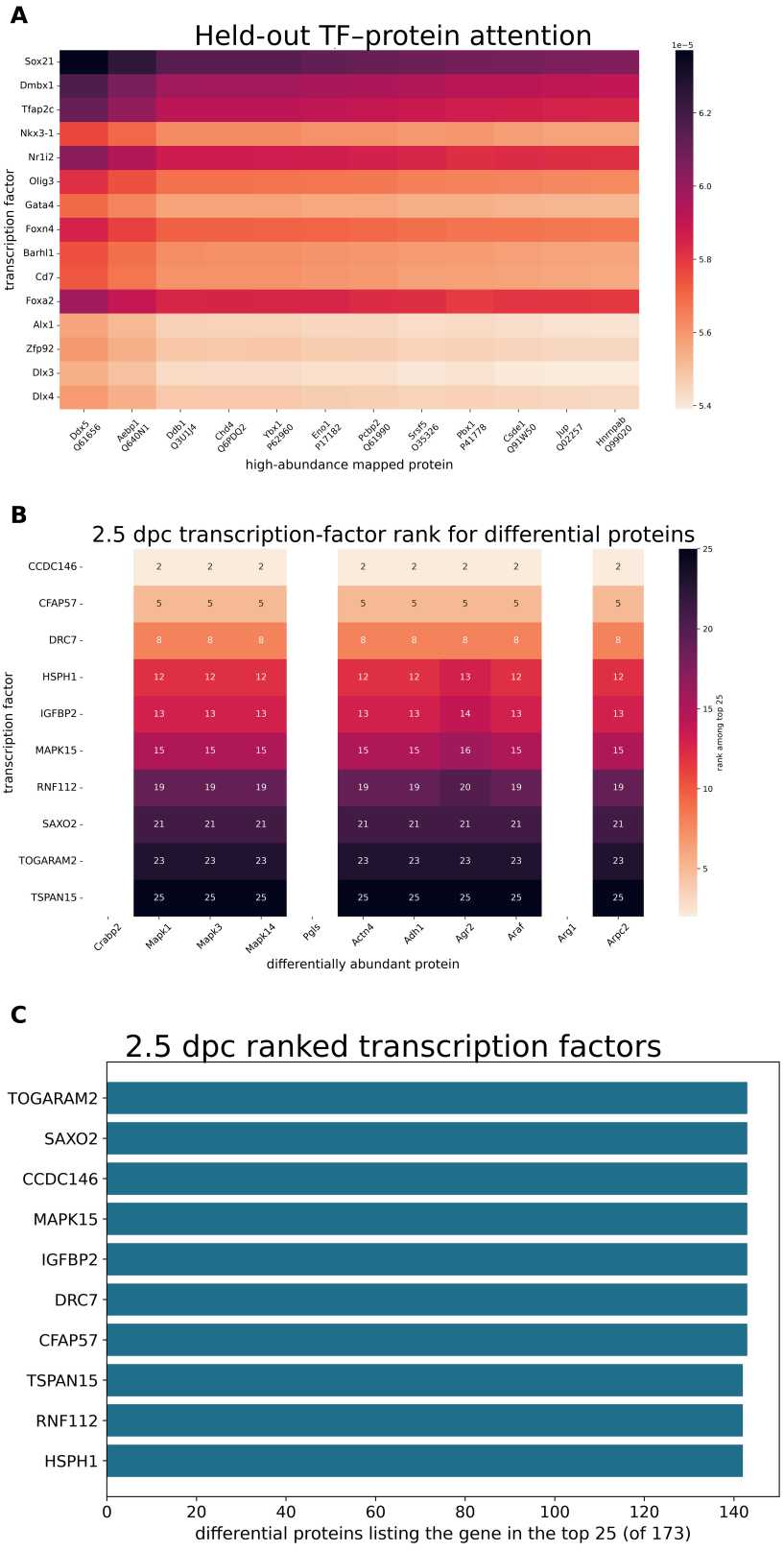


Figure S4. Attention-based prioritization of transcription factors (TFs) and mapped proteins. (A) Model-derived attention weights for selected TF rows and high-abundance mapped protein columns in the held-out condition; color indicates raw attention magnitude. (B) Rank positions of TFs across selected differentially abundant proteins in the 2.5 dpc condition; a smaller rank indicates greater prioritization. (C) Number of protein targets for which each TF appears among

the top 25 prioritized candidates, summarized across 173 targets in the 2.5 dpc analysis. Attention is interpreted as a model association for hypothesis generation, not direct evidence of biological regulation.

Figure S5. Functional enrichment of prioritized gene and protein sets

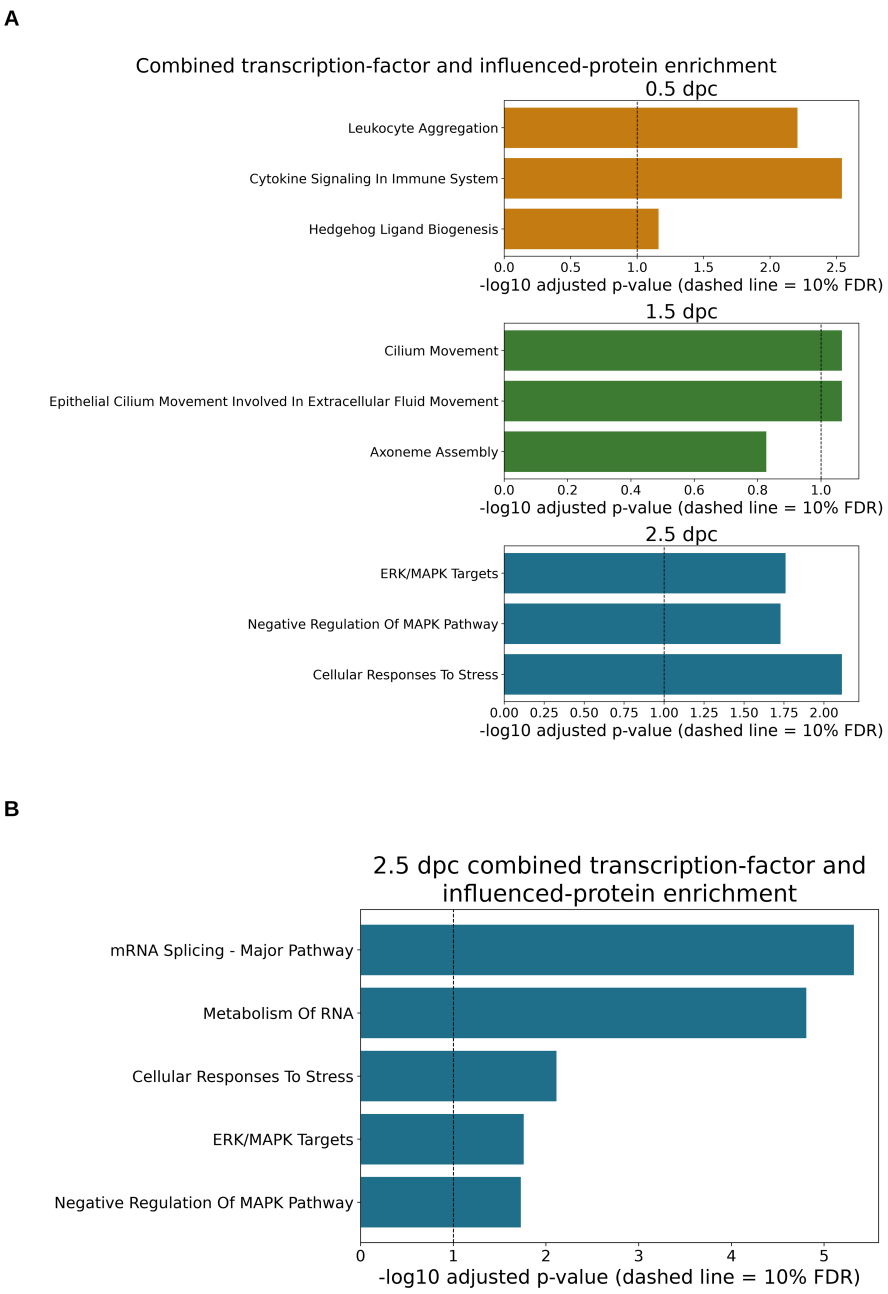


Figure S5. Functional enrichment of combined TF and influenced-protein gene sets. (A) Representative enrichment terms for the control/estrus comparisons with 0.5, 1.5, and 2.5 dpc, respectively. Bars represent the negative base-10 logarithm of the Enrichr adjusted P value for selected terms. (B) Expanded display of representative enrichment terms for the 2.5 dpc comparison. The dashed vertical line represents a 10% false-discovery-rate threshold (adjusted P = 0.10). The displays summarize supplied enrichment results; enrichment calculations were not repeated solely for figure preparation. Data provenance: The enrichment visualizations use the supplied Enrichr workbooks for the combined transcription-factor and influenced-protein sets. Supporting machine-readable outputs and required field descriptions are provided separately in Table S6 and the accompanying data materials.