Software Article

# panomiX: Investigating mechanisms of trait emergence through multi-omics data integration

Ankur Sahu [a], Dennis Psaroudakis [a,b], Hardy Rolletschek [a], Kerstin Neumann [a], Ljudmilla Borisjuk [a], Axel Himmelbach [a], Kalyan Pinninti [a], Dominic Knoch [a], Nadine Töpfer [c], Jędrzej Szymański [a,b,*]

[a] *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Seeland, Germany*
[b] *Forschungszentrum Jülich, Institute of Bio- and Geosciences (IBG-4 Bioinformatics), CEPLAS, BioSC, Jülich, Germany*
[c] *Institute for Plant Sciences, Cluster of Excellence on Plant Sciences (CEPLAS), University of Cologne, Zülpicher Str. 47b, 50674, Cologne, Germany*

## ARTICLE INFO

## ABSTRACT

Complex omics approaches and high-throughput phenotyping generate large, heterogeneous datasets that make linking molecular signatures to plant traits challenging. To address this challenge, here we introduce panomiX, a user-friendly toolbox for multi-omics integration, designed to enable non-experts to apply advanced computational methods with ease. PanomiX automates data preprocessing, variance analysis, multi-omics prediction, and interaction modeling through machine learning, revealing meaningful molecular interactions and synergies. We applied panomiX to a tomato heat-stress experiment combining image-based phenotyping, transcriptomics, and Fourier-transform infrared spectroscopy data, with the aim of identification of condition-specific, cross-domain relationships between gene expression, metabolite levels, and phenotypic traits. Our approach identified a network of such connections, with those linking photosynthesis traits with stress-responsive kinases in elevated temperatures among most significant ones. By simplifying complex analyses and improving interpretability, panomiX offers a platform to accelerate the discovery of trait emergence in plants and select specific candidate genes based on multi-omics analyses.

## 1. Introduction

Recent advances in high-throughput plant phenomics have greatly expanded both the scale and depth of acquired data [1,2]. Phenomics data are often combined with molecular profiling, including transcriptomics and metabolomics [3–5], for instance in GWAS experiments [6], time series [7], or genotype-contrast studies [8]. These combined approaches provide valuable insights into the genetic and molecular basis of trait emergence [9,10]. A prominent example is plant stress resilience, which is shaped at multiple levels: from genetic variation, through gene regulation, to metabolism and biophysical processes [11]. Several studies have shown that profiling these molecular layers (i.e., endophenotypes) can help explain the mechanisms of plant stress responses (summarized in Table 1). Nevertheless, integrating and interpreting such diverse datasets remains a major challenge.

Multi-omics deals with high-dimensional data [18] (Fig. 1A) that vary in coverage, variance scales, and exhibit batch effects [19]

(Fig. 1B). Although many specialized tools exist for analyzing specific omics data types (spanning network-based [20–22], correlation-based [23], similarity-based [24–26], Bayesian [27–29], fusion-based [30], and multivariate methods [31–33]; Fig. 1C), these often fail to integrate different omics layers due to usability constraints, complex workflows, or limited interfaces. Consequently, there is an urgent need for a flexible, comprehensive, and user-friendly solution that can seamlessly integrate diverse omics data types using advanced machine learning techniques.

To address this need, we introduce panomiX, a user-friendly, open-source platform designed for implementing data integration using eXtreme Gradient Boosting (XGBoost) for multi-omic data of varying scales and quality [34]. PanomiX employs a harmonization and scaling pipeline prior to data integration, and accommodates both quantitative and categorical data utilizing cross-predictions between omics layers. XGBoost's inherent ability to handle missing values eliminates the requirement for explicit data imputation. PanomiX provides model interpretability using "SHapley Additive exPlanations" (SHAP) [35],

---

**Table 1**
Examples of recent multi-omic studies related to plant stress phenotypes.

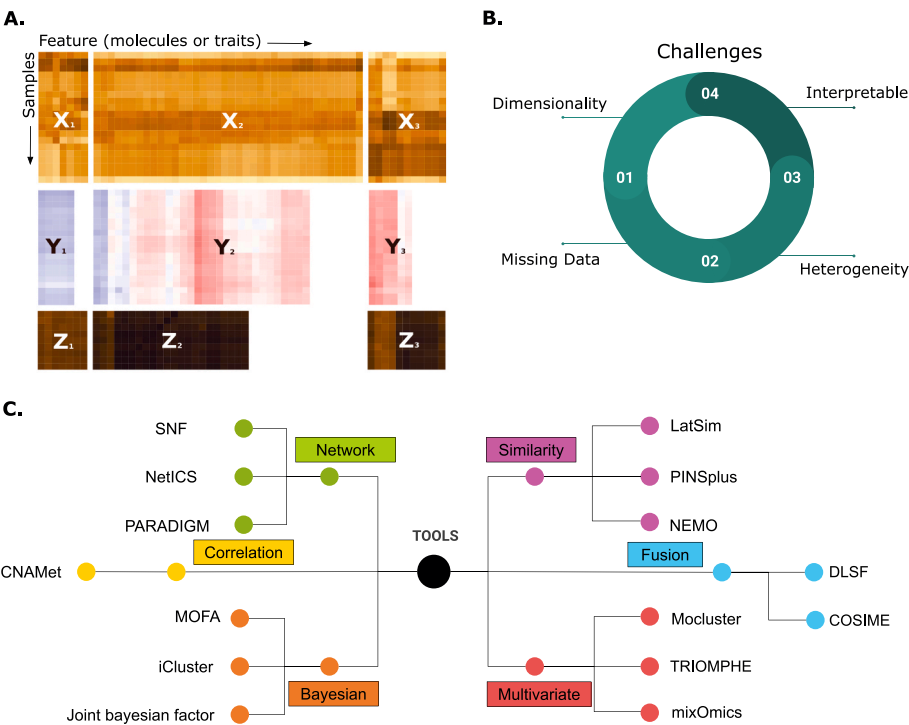| Plant | Omics | Stress | Trait predicted | References |
|---|---|---|---|---|
| *Zea mays* L. | Transcriptomics | *Ustilago maydis* | Control and Biotic stress | Nazari et al. (2023) [12] |
| *Solanum lycopersicum plants (cv Micro-Tom)* | Phenomics, Ionomics, Transcriptomics, and Metabolomics | Salinity, Heat, and Salinity + Heat | Stress treatment group | Pardo-Hernández et al. (2024) [13] |
| *Solanum tuberosum* | Transcriptomic, Metabolomics, and Proteomics | Heat, Drought, and Waterlogging | Control and Abiotic stress | Zagorščak et al. (2025) [14] |
| *Arabidopsis thaliana* | Transcriptomics | *Botrytis cinerea, Sclerotiorum sclerotiorum, and Pseudomonas syringae* | Disease severity | Sia et al. (2025) [15] |
| *Zea mays* ssp. *mays* | SNPs, Transcriptomics, and Metabolomics | Salt | Salt-tolerance-related traits | Liu et al. (2025) [16] |
| *Triticum aestivum* | Metaboolomics | Drought | High-throughput phenotyping | Wonneberger et al. (2025) [17] |



**Fig. 1.** Multi-omic integration - challenges and methods. (A) Omic data sets usually differ in sample size, number of features and variance even if collected in the frame of the same experiment. (B) Different challenges in multi-omics data integration (C) List of available multi-omics data integrative tools.

facilitating transparent feature selection, ranking, and visualization. A key feature of panomiX is its capacity to apply user-defined constraints. This enables researchers to test hypotheses on specific biological interactions of interest. This capability extends its utility beyond data integration, allowing for detailed exploration of potential molecular mechanisms between pre-selected sets of phenotype, genotype, and molecular features.

Importantly, panomiX extends beyond classical statistical inference methods, such as ANOVA or correlation analyses, by leveraging XGBoost's strengths in prediction, feature interaction, and ranking. This enables the identification of robust, high-value features across omics layers by capturing nonlinear relationships, which serve as starting points for conventional approaches to address mechanistic questions, such as identifying stress-responsive pathways. In this way, panomiX bridges predictive machine learning with hypothesis-driven inference.

To demonstrate the power of panomiX, we applied it to study the response of tomato seedlings (*Solanum lycopersicum*) to elevated temperatures - an agriculturally relevant process affecting tomato yield both in the greenhouse and field cultivation. Tomato plants are highly sensitive to heat during early growth and later during pollen development,

leading to poor germination, inhibited growth, and reduced fruit set [36–38]. At the molecular level, heat stress induces transient gene reprogramming, triggering heat shock proteins (HSPs) and heat shock factors (HSFs), often regulated by MAP kinases and Cysteine-rich Receptor-like kinases (CRKs) [39–43]. Multi-omic approaches capturing transcriptomic, metabolomic, and protein data were crucial to unraveling some elements of these complex regulatory networks [13].

In this study, we analyzed the heat stress response and subsequent recovery of young tomato seedlings under controlled conditions, combining daily high-throughput phenomic profiling with transcriptomics using RNA-seq analysis and biochemical fingerprinting using Fourier-transform infrared spectroscopy (FTIR) at selected time points. FTIR is used here as a proxy for metabolic profiling, a rapid, non-destructive technique that measures the infrared absorbance spectrum of a sample, which reflects the abundance of major functional groups such as lipids, proteins, and carbohydrates. While it does not identify individual metabolites, FTIR captures system-wide biochemical shifts that, when integrated with other omics layers via panomiX, can be linked to specific biological processes. This provided a rapid, global view of metabolic changes that complemented our transcriptomic and

phenomic datasets. We describe both common response patterns across all data types and those unique to specific datasets and highlight key interactions among transcripts, metabolic features, and phenotypes uncovered by panomiX. Together, these demonstrate the functionality and versatility of panomiX for multi-omics data integration and analysis.

## 2. Materials and methods

### 2.1. Data preprocessing and analysis

The panomiX platform integrates several specialized R libraries for data harmonization, variance analysis, multi-omics prediction, interaction, and visualization. We provide default data preprocessing methods for transcriptomics and FTIR data. For transcriptomics, we provide a 'DESeq2-based' [44] approach, adjusting for factors like sequencing depth and sample composition. For FTIR spectral data, the 'baseline' R library [45] is employed to remove background noise, and the 'signal' library [46] for Savitzky-Golay smoothing [47] of spectral features without loss of peak sharpness and spectral integrity. Variance analysis is conducted using the 'irlba' R library and prcomp function [48], including dimensionality reduction, clustering, and identification of sources of variance. Multi-omics prediction workflows are facilitated by 'caret' R library [49], with the built-in framework for hyperparameter tuning and model evaluation. XGBoost is used for predictive modeling, and SHAP and Boruta-SHAP are employed for model interpretability and feature selection [50].

### 2.2. Visualizing insights with interactive plots

The panomiX user interface (UI) uses 'shinyjs' [51] to create dynamic interactions, allowing users to control UI element visibility based on their actions, for a streamlined experience. Interactive plots are created with 'plotly' [52], enabling users to zoom, pan, and hover over data points for deeper exploration. The 'DT' package [53] enables interactive data tables, facilitating sorting and exploration of large datasets. 'shinycssloaders' [54] provides visual feedback on application progress during data loading. Additionally, 'bslib' [55] offers advanced plot formatting tools, and 'bsicons' [56] incorporates intuitive icons, enhancing UI usability and aesthetics.

### 2.3. Source code availability and community collaboration

The panomiX toolbox is developed fully in R [57] and Shiny [58] and deployed on Shinyapps. io https://www.shinyapps.io/. The source code is managed with a GitHub repository connected to the Shinyapps. io via 'rsconnect' [59]: https://szymanskilab.shinyapps.io/panomiX/. The source code for the platform is available on GitHub: https://github.com/NAMlab/panomiX-tool. The repository contains all the necessary R scripts for data processing, visualization, and machine learning prediction.

### 2.4. Experimental data and design

Seeds of genotype *Moneymaker* were planted in well-watered soil trays and kept under a plastic cover in a walk-in phytochamber with a day/night cycle of 16/8h and temperatures of 24 °C and 20 °C respectively. The seeds were kept without light for one day, then illumination of 320 μmol per square meter per second (from Whitelux Plus metal halide lamps, Venture Lighting Europe Ltd., Rickmansworth, Hertfordshire, England) was added during the day phase. After a total of 7 days after sowing, evenly germinated and healthy seedlings were transplanted into individual pots (10 cm diameter, 8 cm height) without plastic covers at 60–70 % relative air humidity. After 9 further days of establishment in the pot, the plants were exposed to heat stress at 37 °C/28 °C (day/night) for 6 days followed by a recovery phase at 24 °C/

20 °C. Plants were watered once or twice daily by an automated system replenishing evapotranspirated water by weight to ensure the plants do not experience drought stress. All plants were phenotyped daily using an imaging-based high-throughput system and the youngest fully formed leaf pair was harvested (directly put in liquid nitrogen and then stored at −80 °C) for molecular measurements mid-day (between 13:00 and 14:00, light phase was from 06:00–22:00) from individual plants directly before the temperature switch, each day during the heat stress phase and after 1, 2, and 4 days of recovery. Control plants were grown exactly the same way except that they stayed at 24 °C/20 °C throughout the whole experiment. They were sampled at the same respective time points as the heat-treated plants.

### 2.5. Data acquirement

During daily phenotyping, plants were photographed from the top and three side angles in visible and in fluorescent light (excitation: 400–500 nm, emission: 520–750 nm). The images were then analyzed using the Integrated Analysis Platform (IAP) software [60] to yield 138 phenotypic traits for each measurement, as described in Ref. [61]. Next we reduced the 138 phenotypes into a smaller set of non-correlated variables. Using hierarchical clustering, we grouped related phenotypes together. The number of clusters was estimated using PCA, determining the minimal set of clusters, for all of which the first PC described at least 50 % of their variance (representing the cluster co-linearity). Through this process, we narrowed the phenotypes down to 27 key representatives, which we used for model training.

For the molecular assays, the frozen leaves were ground and aliquots subjected to specific measurements: Total RNA was isolated from 70 mg of the ground material using the RNAeasy Plant Mini Kit (QIAGEN) according to the manufacturer's protocol. The construction of sequencing libraries involved the Illumina stranded mRNA Prep Ligation Kit (standard Illumina protocol; Illumina, San Diego, California, USA) and 1 μg DNAse I digested total RNA. Sequencing of equimolar library pools (average size: 332 bp) was performed on an Illumina NovaSeq 6000 device (IPK-Gatersleben), using the XP-workflow and a S2 flowcell (Illumina, San Diego, California, USA). On average, 41 M reads (single reads, 118 cycles) were generated per sample. Reads were then mapped to the ITAG4.1 tomato reference genome (https://solgenomics.net/ftp/tomato_genome/annotation/ITAG4.1_release/) using our rnaseq-mapper pipeline built around kallisto [62], available at https://github.com/NAMlab/rnaseq-mapper; [63].

FTIR analysis of freeze-dried leaf material was performed as in Ref. [64]. Briefly, approximately 2 mg of freeze-dried material was used per sample. Spectra were produced from ATR-FTIR measurements using the INVENIO-S FTIR spectrometer (Bruker Optics, Ettlingen, Germany) with a Globar light source under continuous purging with dry air. ATR absorbance spectra were recorded in the spectral range of 4000-400 $cm^{-1}$ at a spectral resolution of 4 $cm^{-1}$. Each spectrum consisted of 32 co-added scans. As a background reading, the spectrum of the empty ATR crystal was collected prior to measurement and subtracted automatically from each recorded spectrum using the OPUS software (Bruker Optics).

## 3. Results

### 3.1. Data input

Our experiment provided 138 phenotypic variables reduced to 27 uncorrelated phenotype clusters (Fig. S1), 34,688 gene expression values, and 2526 FTIR data points across 10 time points and 2 to 3 biological replicates in two experimental conditions: control growth, and treatment with 6 days of heat stress treatment followed by 3 days recovery phase (Fig. 2).
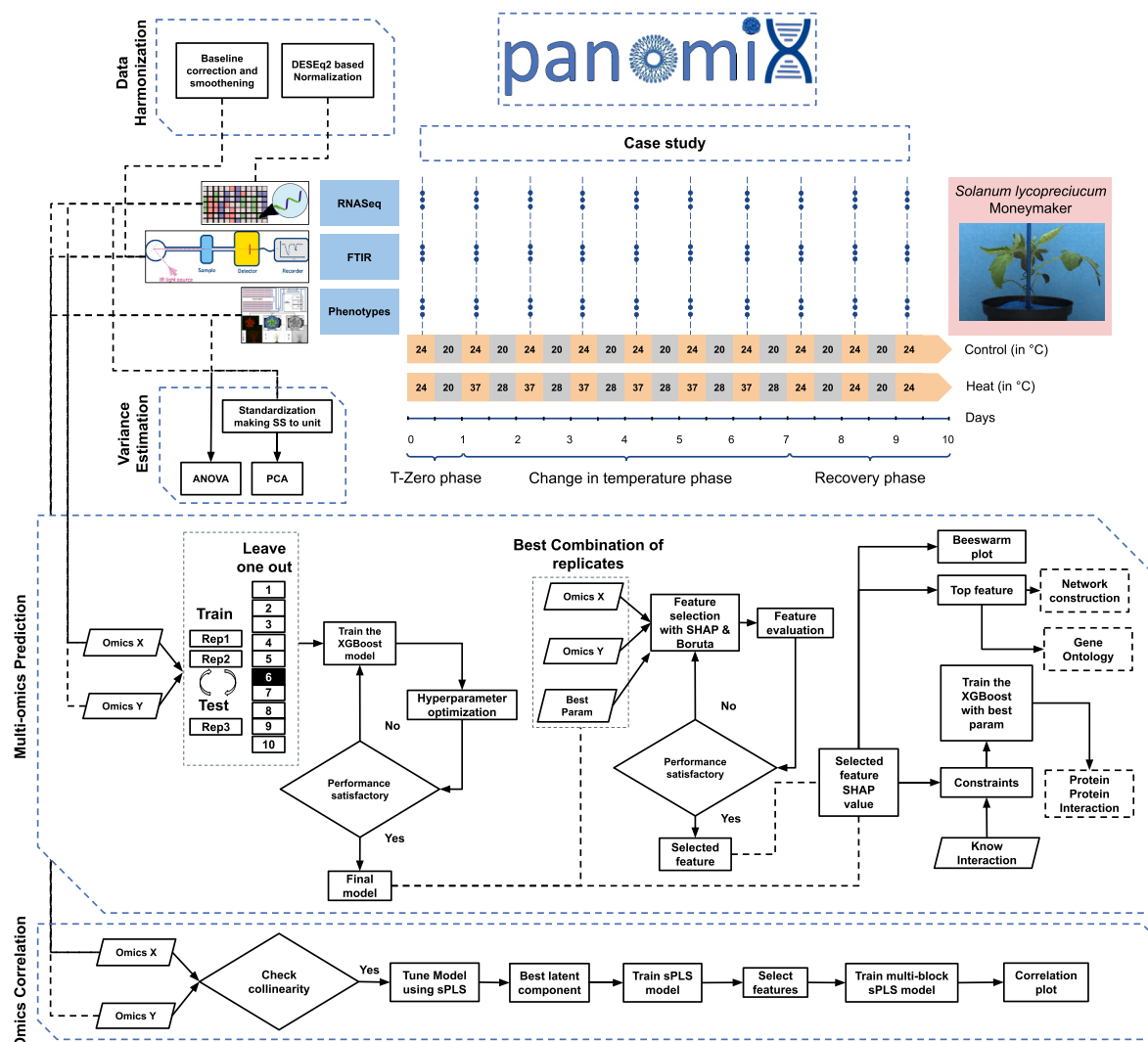
**Fig. 2.** Flow diagram of the panomiX toolbox and the experimental design for the tomato heat stress case study.

### 3.1.1. Tool settings and recommendations

PanomiX works with continuous molecular data, such as normalized RNA-seq counts, protein abundances, metabolite concentrations, or FTIR spectra. Your data should be in a feature matrix format, where: Columns represent biological samples (e.g., individuals or time points). Rows represent molecular features (e.g., genes, proteins, metabolites, or spectral variables). A continuous outcome variable (y) is needed for regression tasks. PanomiX can handle large datasets with tens of thousands of features, but pre-filtering your data is recommended for better performance and computational efficiency. Here's how you can optimize your dataset: filter low-variability features, remove features that show little variation across samples, exclude low-count features for sequencing data (like RNA-seq), and drop features with consistently low counts, remove near-zero variance features which contribute little to the model and can be safely excluded. These steps help reduce the size of high-dimensional datasets, minimize overfitting, and improve the speed of machine learning algorithms like XGBoost during model training and hyperparameter tuning. Before using panomiX, ensure that your datasets are pre-processed and normalized according to the requirements of the specific 'omics platforms.

While methods such as XGBoost can accommodate $p \gg n$ settings, omics datasets with fewer than ~50 observations should generally be regarded as exploratory. The effective number of required samples for panomiX depends on factors such as the signal-to-noise ratio, the correlation structure among predictors, and the stability of cross-validation

performance. We recommend that readers first examine the variance structure of their data using multivariate approaches (e.g., Principal Component Analysis, PCA) to estimate how many principal components explain the majority of variance. For example, if three PCs capture ~90 % of the biological variance, a relatively simple model with only a few effective predictors may be sufficient, regardless of the initial dimensionality. When sample size is limited, model robustness could additionally be assessed using learning curves or resampling strategies. Importantly, the number of predictive features retained by the model is constrained by sample size and should remain considerably smaller than $n$.

### 3.2. Data processing and filtering

The RNA-seq data processing with rnaseq-mapper [https://github.com/NAMlab/rnaseq-mapper] provided a data matrix of estimated TPM values (transcript per million, kallisto estimates; [62]) for 49 samples (both control and treatment), with columns representing samples and rows representing genes-aggregated transcript levels. We used TPM counts without applying the data harmonization component for further analysis. To filter out transcripts with consistently low TPM counts, we retained only those transcripts that have at least one sample with a count of 50 or higher, ensuring that transcripts with TPM counts below 50 across all samples were removed. Ultimately, we used 5479 transcripts for further analysis. Similarly, for FTIR spectral data, we

created a data matrix of absorbance values, containing 49 samples (both control and treatment). We selected the absorbance values in the range from 4000 to 400 cm$^{-1}$. We then applied FTIR normalization using the data harmonization component in the panomiX toolbox. For illustration Fig. 3A presents the raw FTIR spectra before preprocessing. After baseline correction and Savitzky-Golay smoothing, the normalized spectra are displayed in Fig. 3B, illustrating the improved signal quality.

### 3.2.1. Tool settings and recommendations

PanomiX provides standard methods for normalization of transcriptomic data [44]. First, a natural log transformation is applied to ($x_{ij}$) a raw expression count for gene $i$ in sample $j$ to stabilize variance and improve comparability (Eq. (1)). Then, the transformed gene expression data is used to determine the average estimation for each gene. Next, non- and sparsely expressed genes are filtered out, allowing us to focus only on stably expressed genes ($G$). The previously log-transformed value of each gene is then subtracted from the respective average, helping identify genes within each sample that have higher or lower expression levels than the average. Subsequently, the average-subtracted values are used to determine the median across genes, which reduces the influence of high-value outliers. Finally, the median values are transformed back from log to normal scale to obtain the final scaling factor (Eq. (1)). The raw read counts are then divided by these scaling factors (Eq. (2)). This log and median-based scaling method effectively corrects batch effects and extreme variations. For

consistency, we recommend performing all steps of data normalization from the raw counts within panomiX. However, the platform also accepts normalized transcriptomics data in TPM or CPM counts. For the spectral data, such as those obtained by FTIR, panomiX utilizes the baseline correction on the raw spectrum to remove artifacts and the background noise. Following this, Savitzky-Golay smoothing can be applied to each baseline-corrected spectrum. This step applies a polynomial smoothing filter to reduce high-frequency noise while maintaining essential spectral features.

The scaling factor for sample j, is calculated as:

$$s_j = exp\left( median_{i \in G} \left[ log\left(x_{ij}\right) - \frac{1}{n} \sum_{k=1}^{n} log\left(x_{ik}\right) \right] \right) \tag{1}$$

where $x_{ij}$ is the raw expression count for gene $i$ in sample $j$, $n$ is the number of samples, and $G$ is the set of genes with finite mean log-expression.

The normalized expression values are then given by:

$$x_{ij}^{norm} = \frac{x_{ij}}{s_j} \tag{2}$$

### 3.3. Variance analysis

The exploratory analysis of the multi-omic data variance highlighted both data type-specific, and shared patterns of changes driven by plant
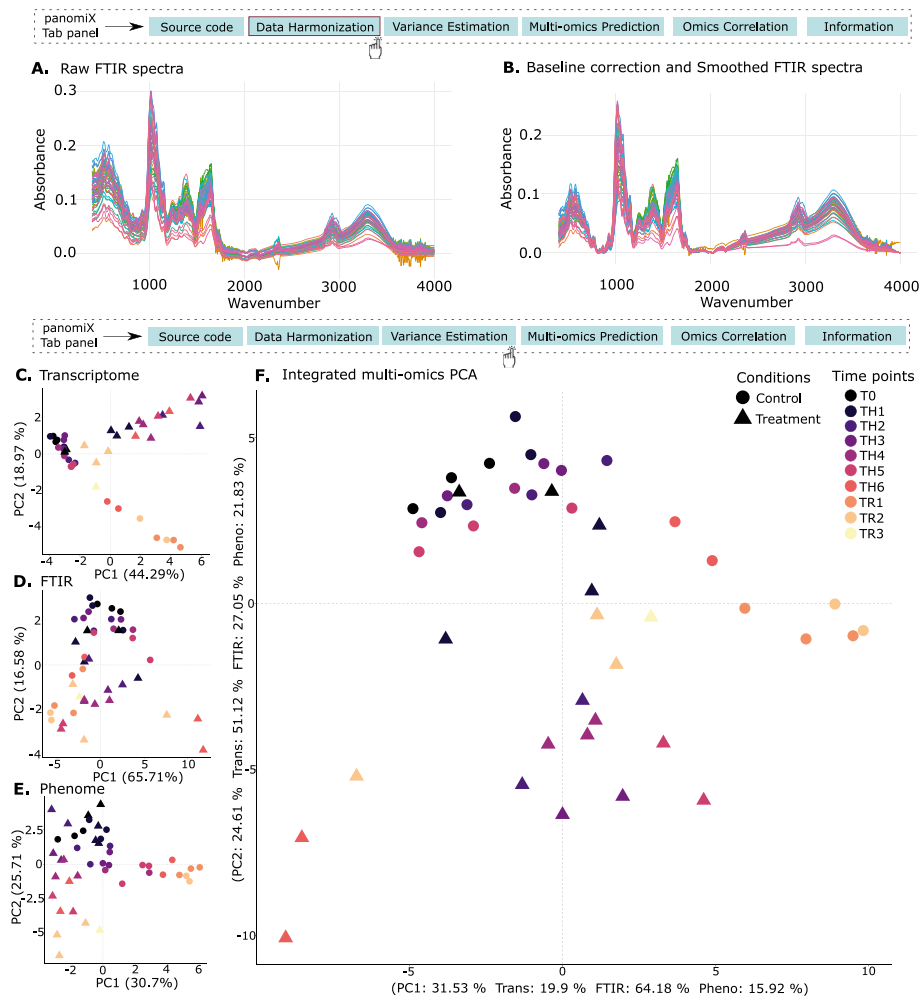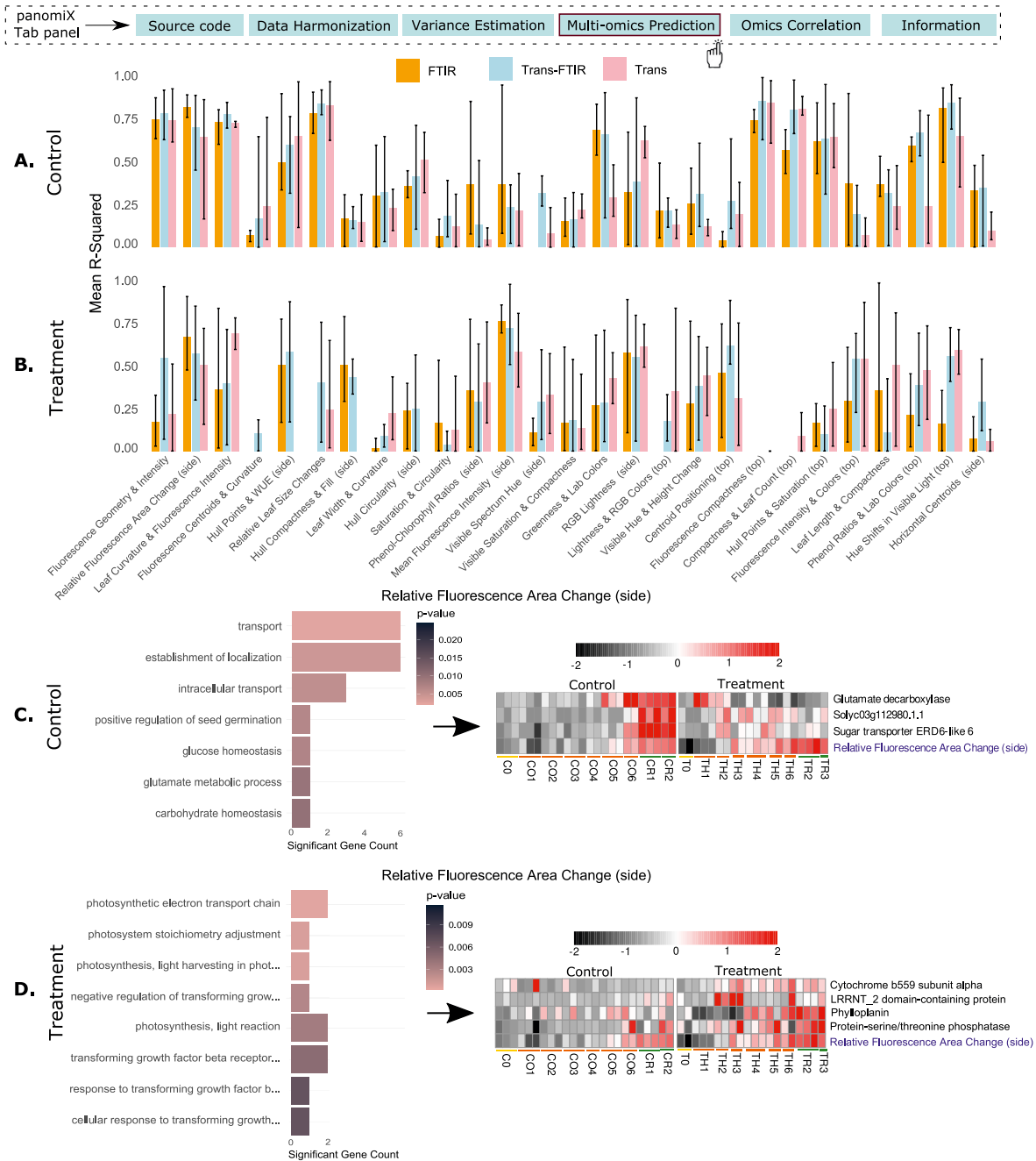


**Fig. 3.** Data harmonization and variance estimation component (A) Raw FTIR spectra (B) Baseline corrected and smoothed FTIR spectra. Individual PCA for each omics data (C) Transcriptome, (D) FTIR, and (E) Phenome. (F) An integrative PCA, where the multi-omic data matrix was combined to identify shared variance components.

development and heat-stress treatment. In the transcriptome PCA (Fig. 3C), PC1 (44 % variance) largely reflects time-dependent changes in gene expression, while PC2 (18 % variance) highlights heat-stress effects. Early-stage controls and T0 heat samples cluster together, indicating minimal transcriptional differences at the start, whereas later-stage heat treatments diverge, revealing progressive stress adaptation. In FTIR absorbance data (Fig. 3D), PC1 (65 % variance) again captures temporal shifts, and PC2 (16 % variance) distinguishes stress from control conditions; notably, mid-stage heat-treated seedlings cluster with recovery-stage controls, indicating partially overlapping metabolic responses. Phenomic data (Fig. 3E) show stronger treatment separation on PC1 (30 % variance) and time progression on PC2 (25 % variance), where early-stage controls resemble T0 heat samples, and later stages

separate more distinctly. Integrating all omics layers (Fig. 3F) confirms the interplay of time and treatment: PC1 (31 % variance) is dominated by FTIR and transcriptome signals, clearly separating recovery-stage heat-treated seedlings, while PC2 (24 % variance), driven by transcriptional variation, further distinguishes controls from heat-stressed plants at various stages. These results accentuate the distinct yet complementary biological insights gained by analyzing temporal and stress effects across transcriptomic, metabolic, and phenotypic data.

### 3.3.1. Tool settings and recommendations

PanomiX supports exploratory data analysis with PCA or ANOVA to identify dominant sources of variance. Before PCA, all datasets are standardized to a unit sum of squares, ensuring comparable total



**Fig. 4.** (A) Performance of control samples for transcripts predicting phenotypes, FTIR predicting phenotypes and combined transcripts + FTIR predicting phenotypes (B) Performance of treatment samples for transcripts predicting phenotypes, FTIR predicting phenotypes and combined transcripts and FTIR predicting phenotypes (C) Gene ontology and expression profile for the features selected for Relative Fluorescence Area Change (side) from the control model (D) Gene ontology and expression profile for the features selected for Relative Fluorescence Area Change (side) from the treatment model.

deviations across different data types, but all the relevant dataset-specific normalization steps should be performed beforehand. First, PCA is performed separately on each omics dataset to capture dataset-specific variance patterns. Second, the integrated PCA is conducted on a combined dataset to reveal shared variance and global patterns. In this case each block of omic data is first scaled to a unit sum of squares and then concatenated into a single input. For PCA, the tool automatically handles data centering and scaling. For the ANOVA and plotting the PCA scores, the user needs to upload metadata with at least an ID column (matching the omics data) and respective condition columns if experimental factors are present. PanomiX often exposes differences in resolution and variability among different omic datasets, demonstrating how multi-omics integration can also provide insights into each individual dataset. Notably, these differences should be carefully evaluated, as differences in normalization, coverage and data filtering procedures can confound their biological interpretation.

### 3.4. Multi-omics prediction components

Linking the high-throughput profiling of plant traits with molecular assays such as RNA-seq or metabolomics, can yield valuable biological insights into molecular mechanisms of trait emergence. To explore these connections, we conducted cross-prediction on phenotypes using transcriptomics alone, FTIR alone, and combined transcriptomics + FTIR data, each trained separately on control and treatment samples. This approach enabled identification of condition-specific molecular markers. Model performance, assessed via mean R-squared values across three replicate splits, showed generally higher accuracy in control samples; however, some treatment-based models (e.g., Phenol–Chlorophyll Ratios (side), Mean Fluorescence Intensity (side)) outperformed their control counterparts (Fig. 4A and B). Next, we performed gene ontology (GO) enrichment using 'topGO' R package [65] on transcripts with high SHAP values (i.e., highly predictive of phenotype), focusing on models with $R^2 > 0.5$ and replicate-model $R^2 > 0.7$. This criterion yielded 10 control and 7 treatment models (Fig. S2A and S2B). Certain phenotypes had unique GO enrichments, while others overlapped. For instance, a control model predicting Relative Fluorescence Area Change (side) highlighted Glutamate decarboxylase and Sugar transporter ERD6-like 6 both linked to enhanced photosynthetic capacity [66] (Fig. 4C). Meanwhile, the corresponding treatment model identified cytochrome *b*559 subunit alpha, part of the photosynthetic electron transport chain (Photosystem II), and phylloplanin, which contributes to stress defense via type VI glandular trichomes [67,68] (Fig. 4D). Next, we extended the analysis to FTIR and combined transcriptomics + FTIR models, applying the same performance thresholds ($R^2 > 0.5$ overall and $R^2 > 0.7$ for individual replicate models). Within this subset, combined transcriptomics + FTIR models improved predictive ability by an average of 10 % (up to 43 %) compared to transcriptomics alone and average of 5 % (up to 24 %) compared to FTIR alone under control conditions. Under treatment conditions, integration yielded an average improvement of 12 % (up to 33 %) over transcriptomics alone and 12 % (up to 40 %) over FTIR alone (Fig. 4A and B). This consistent performance gain demonstrates that multi-omic fusion provides complementary information not captured by single-omic models. Notably, several transcripts and FTIR features were consistently identified as top predictors across all three omics-specific models (Table S2). For example, in the Relative Fluorescence Area Change (side) phenotype (Fig. S3), the transcript-only model pinpointed cytochrome *b*559 subunit alpha and phylloplanin as key contributors (Fig. S3A). In the FTIR-only model, spectra 2002 and 2003 showed strong predictive power (Fig. S3B). The integrated model (transcriptomics + FTIR) retained these same predictors, while also incorporating additional predictors (Fig. S3C), confirming their importance. Further analysis of the transcript-to-FTIR prediction revealed that phylloplanin significantly predicted the 2002 FTIR spectrum (Fig. S3D).

### 3.4.1. Tool settings and recommendations

Training the XGBoost model requires splitting the data into training and test samples. Users can split datasets either randomly (using a training size slider) or by replicate to ensure consistent grouping of the same replicates, and hyperparameter tuning is automated via the caret method. In the cloud version of panomiX, two main hyperparameters (number of rounds, tree depth) can be configured, with others (learning rate, gamma, subsample, etc.) preset to optimal ranges. The desktop version provides greater flexibility, letting users adjust a wider range of parameters (see the full documentation; https://github.com/NAMlab/panomiX-tool). Model generalization is validated through cross-validation (CV) or leave-one-out cross-validation (LOOCV), and performance is reported using $R^2$ and RMSE. A model performance plot ($R^2$) is displayed, alongside a feature-importance table that ranks predictors. SHAP values further reveal how individual features positively or negatively influence phenotype predictions, visualized via beeswarm plots for intuitive interpretation. The Boruta-SHAP algorithm is also available for alternative feature selection, with its results shown in a separate beeswarm plot. For biological replicates, we recommend replicate-based splitting (with metadata specifying sample names and replicate identifiers) to improve model consistency and accuracy. Random splitting with an adjustable training size can be used for general evaluation. In the cloud version, tuning the number of rounds and tree depth is crucial for better performance; the desktop version allows deeper parameter control. CV is advisable for robust model generalization, while LOOCV suits smaller datasets. For deeper insights, users can leverage SHAP values and feature-importance rankings to interpret and prioritize key predictors.

### 3.5. Multi-omics prediction with interaction constraints

PanomiX can analyze interactions of predictive features identified by the model as well as "known" features provided by users - such as genes of interest - by incorporating both sets into a single prediction framework. It uses SHAP values to show how these features (positively or negatively) influence a given outcome (e.g., a phenotype) and to set constraints for the model (Fig. 5A). In our experiment, these constraints relate to a predetermined list of literature- and annotation-based candidate transcripts. We tested two scenarios: a) evaluation of data-derived candidate transcript set; and b) evaluation of a predetermined list of candidates.

In the first scenario, we tested 57 heat-stress-related transcripts (46 up-regulated, 11 down-regulated; Psaroudakis et al. [63]) alongside the features previously identified by panomiX. Notably, several transcripts showed positive interactions with both the predicted features and the target phenotypes (Table S3). For instance, in a heat-treatment model predicting Mean Fluorescence Intensity (side), GTP-binding nuclear protein (down-regulated under heat stress) maintained its down-regulation relative to controls, aligning with its predictive role. Similarly, in a model predicting Hue Shifts in Visible Light (top), thioredoxin domain–containing protein interacted with splicing factor U2af small subunit B-like, suggesting a joint effect on the phenotype under heat stress (Fig. 5B and C).

In the second scenario, we explored whether stress responses extend beyond heat shock proteins to include kinases and other signaling elements. Transcriptional regulators together with protein kinases constitute key classes of regulatory proteins that govern plant growth, development, and responses to both biotic and abiotic stimuli [69]. In addition, enzymes such as guanylate cyclases, which catalyze the conversion of GTP to cyclic GMP, act as important upstream regulators that activate protein kinase–mediated signaling pathways during abiotic stress responses [70–73]. For that purpose, we compiled 1531 candidate genes (99 guanylate cyclases, 1038 other kinases, 394 heat-related genes) from literature and databases [69,74,75]. We then trained models using no constraints, random constraints, or monotonic constraints (Fig. 5D). Most cases showed improved performance with
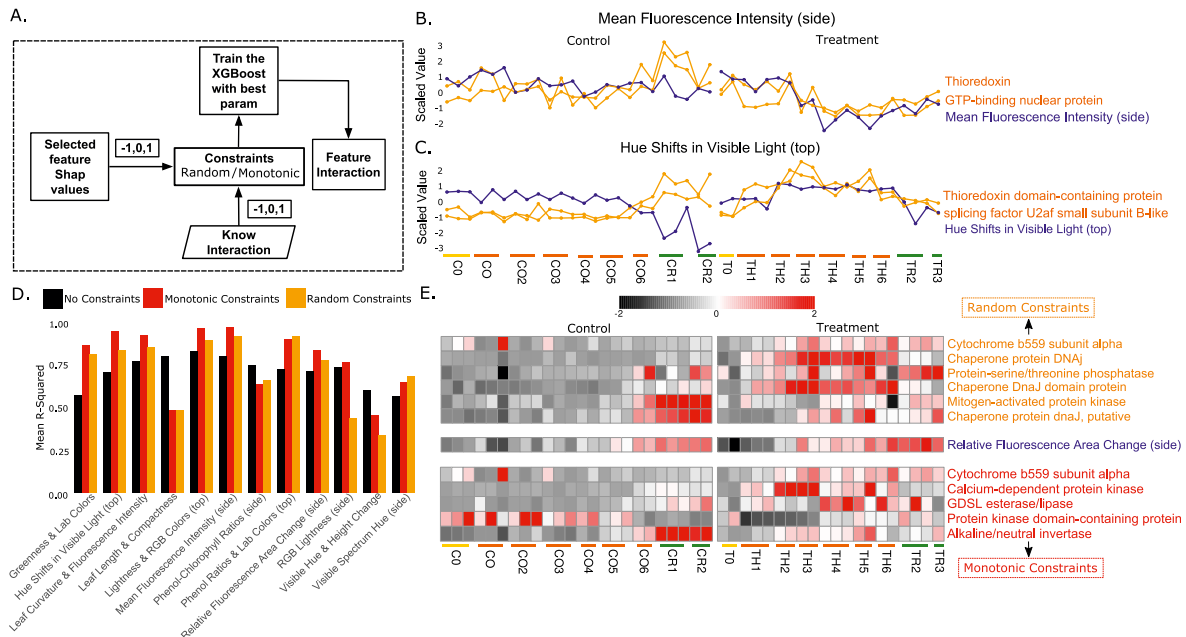
**Fig. 5.** Feature interaction. (A) Illustration of feature interaction as a flow diagram (B) Interaction between the heat stress related (scenario a) transcripts and predictive features from trained model Mean Fluorescence Intensity (side) (C) Interaction between the heat stress related (scenario a) transcripts and predictive features from trained model Hue shifts in Visible Light (top) (D) Performance of treatment samples for transcripts predicting phenotypes (scenario b). Black color bars represent models trained without constraints, red color bars models with monotonic constraints, and orange color bars models with random constraints, respectively. (E) Expression profile for the features selected from the constraints model of Relative Fluorescence Area Change (side) (scenario b). Orange colored features are selected from the random constraints models and red colored features are selected from the monotonic constraints models.

random constraints over no constraints, and an additional ~5 % improvement under monotonic constraints. A representative example is a model predicting Relative Fluorescence Area Change (side), where random constraints boosted performance by 5 % compared to no constraints. Key transcripts (e.g., cytochrome *b*559 subunit alpha, protein-serine/threonine phosphatase) interacted with MAP kinase and DnaJ domain proteins (Table S4). Under monotonic constraints, performance rose another 5 %, and cytochrome *b*559 subunit alpha was linked to GDSL esterase/lipase and calcium-dependent protein kinase, both implicated in stress response [76–78].

Overall, these findings highlight panomiX's capacity to reveal significant interactions among user-specified and model-predicted features, offering mechanistic insights into phenotype determination and guiding further experimental validation.

### 3.5.1. Tool settings and recommendations

By including a user-provided list of relevant (or potentially important) known features, panomiX can determine whether these inputs significantly contribute to the model's outcome or interact with features already identified by the model. The "already predicted features" are those uncovered after training, together with their SHAP values, which may be positive or negative in relation to the outcome. PanomiX uses these SHAP values to set constraints, illustrating each feature's relationship with the predicted outcome. In 'random constraints' setting, the direction (positive or negative) of feature-outcome relationships is not predefined. In this case, users provide a table with two columns—"feature" and "final_association"—where "final_association" contains only 0 (meaning the relationship is unknown). PanomiX then applies random constraints using this information together with the "predicted outcome." In the 'constrained' setting, monotonic constraints require a table with the same columns, but users assign 1 for a known positive relationship, −1 for a known negative relationship, or 0 if it is unknown. PanomiX merges this information with the "predicted outcome" from the trained model, ensuring predictions remain consistent with domain knowledge.

### 3.6. Comparing panomiX with other tools of multi-omics data integration

To evaluate the efficiency, explainability, and stability of the panomiX multi-omics cross-prediction functionality, we tested it against mixOmics [32], a comparable state-of-art tool offering similar functionalities. Specifically, we compared the statistical methods offered by both tools as a main technique of data integration, namely the sparse Projection to Latent Structures (sPLS) of mixOmics with the XGBoost implemented in panomiX. We benchmarked both methods against standard implementations of Random Forest (RF) and Support Vector Machine (SVM) [49]. For that purpose, we conducted two cross-prediction analyses: a) transcriptomics data was used to predict phenotypic traits; b) the same phenotypic traits were predicted from the FTIR data. For each phenotype prediction, models were trained using four independent train-test splits (Fig. 6A and B).

To further investigate robustness, we compared panomiX and mixOmics (sPLS) on transcriptomics subsets with varying correlation coefficients (0.2–0.9), using one phenotype cluster (Leaf Curvature & Fluorescence Intensity) as a test case. Despite sPLS's known strength with highly collinear data, panomiX consistently outperformed it (Fig. 6C). We also evaluated feature importance stability across the correlation gradients by comparing the top 10 ranked features (Jaccard index). PanomiX consistently identified common features (Fig. 6D), indicating it effectively captures both linear and non-linear relationships in transcript–phenotype data while maintaining high predictive performance.

It is important to note that several tools exist for the integration of multi-omic data, each with different primary focus: *e.g.* DIABLO (which employs sPLS-DA from the mixOmics framework) [79] and MOGONET are primarily designed for classification [80], while MOFA can be applied to both classification and regression tasks [27]. While panomiX primarily excels in regression problems, it might be also deployed for classification. Thus to compare these methods, we performed a simple benchmarking for a two class classification task (control and heat treatment), using the default model parameters for both integration and
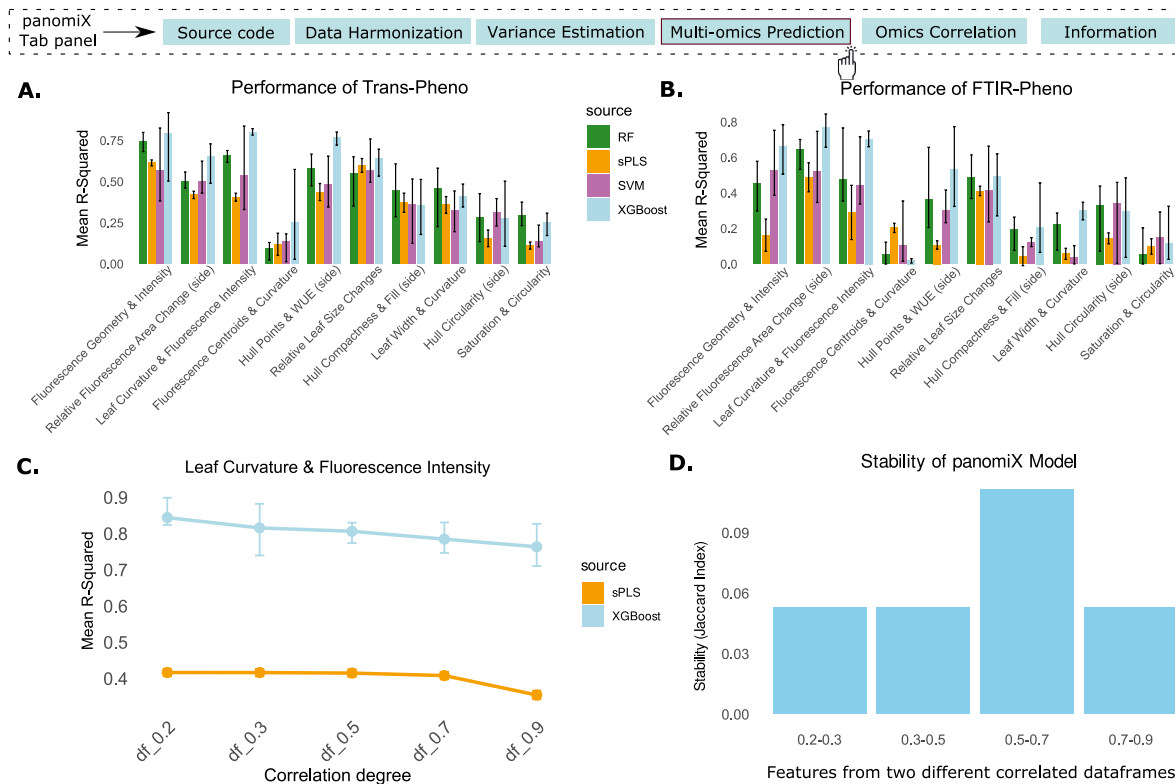
**Fig. 6.** Multi-omics integration in panomiX and other methods. (A) Performance for transcripts-phenotype prediction. (B) Performance for FTIR-phenotype prediction. (C) Effect of multicollinearity on model performance (linear vs. non-linear method). (D) Robustness of feature selection for different thresholds of the feature collinearity in the panomiX XGBoost implementation.

prediction. For the analysis, transcriptomics and FTIR data were jointly used to predict treatment classes. Samples were grouped according to biological time points: no-stress (days T0), early heat treatment (days T1-T3), and recovery phase (days T8-T10). In each iteration, one time point was held out for testing, while the remaining time points were used for training. This leave-one-out procedure was repeated across all grouped time points to assess temporal generalizability. Statistical testing revealed no significant differences between panomiX and the other methods. Thus, the results demonstrate that panomiX performs comparably to existing approaches (Fig. S4).

### 3.7. Evaluating performance of panomiX across species and traits

To further validate and benchmark the performance of panomiX in other experimental setups other than time series and treatment contrasts, we utilized two publicly available multi-omic datasets for genetic diversity panels. The first dataset was obtained from Gemmer et al. [81] comprised eight agronomically relevant traits obtained from multi-year field trials and 128 metabolites profiled from the HEB-25 barley (*Hordeum vulgare*) nested association mapping (NAM) population, with 1307 individual lines. Using this metabolic dataset, we applied panomiX to predict each trait, without accounting for population structure during model training. panomiX outperformed the BayesB method for predicting the time to shooting (SHO) trait and showed comparable performance to BayesB for time to heading (HEA), grain yield (YLD), grain number per ear (GNE), and plant height (HEI). In contrast, BayesB had better performance for time to maturity (MAT), ears per $m^2$ (EAR), and thousand grain weight (TGW) (Fig. S5A). These differences suggest that while panomiX effectively captures polygenic signals, accurate prediction for some traits in stratified populations (subgroups of related individuals with shared ancestry) may require explicit modeling of population structure, as done in the BayesB method. Performance

differences may also arise from the underlying genome–trait relationship. For example, Bayesian regression is expected to perform well when the association between predictors and a trait is primarily linear or additive, while XGBoost will capture non-linearities, thresholds and interactions.

The second dataset Knoch et al. [10] features a large comprehensive multi-omics panel of spring-type oilseed rape (*Brassica napus*) on a diverse population with 477 lines. This dataset integrates high-throughput phenotyping, mRNA-Seq transcriptomics, and GC-MS-based metabolomics. We selected six phenotypic traits for validation: the manually determined plant biomass (fresh weight) and four image-derived traits — digital biovolume, plant height, compactness, projected leaf area and a color-related trait — measured at representative time points. For all six traits, panomiX consistently outperformed the RF-based regression models reported in the original study (Fig. S5B), achieving an average improvement of 12 % in the predictive ability ($R^2$). Notably, several of the top-ranked transcripts identified in the original study were again detected by panomiX among the features with the highest SHAP values (Table S5), further supporting their relevance across different modeling approaches. These include C07p48260.1_BnaDAR (PRL1), one of the prime candidate genes reported to be associated with early vegetative biomass and growth-related traits, but also A03p39940.1_BnaDAR, annotated as an ethylene-responsive element binding protein, C07p57790.1_BnaDAR, a homolog of the Arabidopsis xyloglucan endotransglucosylase/hydrolase 17, and the two homeologs A05p28550.1_BnaDAR and C05p43970.1_BnaDAR, both encoding TRAF-like family proteins. Furthermore, several additional promising candidate genes were reported such as A10p28380.1_BnaDAR, showing homology to the Arabidopsis Succinyl-CoA ligase alpha subunit, a key catalytic enzyme of the citric acid cycle, as well as further putative xyloglucan endotransglucosylase-s/hydrolases such as A01p07860.1_BnaDAR and C01p08760.1_BnaDAR

with putative function in the cutting and rejoining xyloglucan chains, which remodels the cell wall and allows cell expansion and growth.

## 4. Discussion

In this study we introduced panomiX, a user-friendly toolbox for multi-omics data integration. We demonstrated the versatility and robustness of panomiX on a unique multi-omic data set designed to identify transcriptomic and metabolic features associated with phenotypic traits. By addressing key challenges in normalization, variance analysis, and predictive modeling, we showed how panomiX enables extracting biologically meaningful insights from such complex datasets.

Our tool provided a robust workflow for transcriptomics and FTIR data, including tailored normalization and scaling of individual datasets. As shown in Fig. 3, stepwise transformations - here, from raw FTIR spectra to normalized outputs - improve data quality and interpretability through baseline correction and smoothing. This mitigates biases in raw datasets and ensures compatibility with downstream analyses such as PCA. PanomiX's PCA and ANOVA functions further standardize and visualize variance across separate and combined datasets, revealing consistent clustering patterns driven by time points and treatment conditions. These insights highlight the complementary nature of multiple omics layers, setting the stage for more targeted multi-omics analyses.

PanomiX implementation of XGBoost consistently outperformed sPLS, RF and SVM for the integration of our tomato heat stress datasets, showing higher accuracy, robustness, and feature stability. Its XGBoost-based framework effectively handles varying collinearity and identifies non-linear relationships, a key limitation in linear models. High Jaccard indices confirm feature stability across correlation gradients, while GO enrichment validates the biological relevance of identified predictors. Notably, the differential performance between control and treatment models underscores condition-specific mechanisms, exemplified by transcripts such as cytochrome *b*559 subunit alpha and phylloplanin in our experiment. These results highlight panomiX's capacity to reveal critical molecular interactions and support hypothesis-driven research.

By integrating transcriptomics, FTIR, and phenomics data, panomiX uncovers shared predictors - such as cytochrome *b*559 subunit alpha and specific FTIR spectra - indicating convergent biological mechanisms. These cross-layer insights reinforce panomiX's ability to reveal broadly relevant molecular features. Moreover, the option to add user-defined elements (e.g., heat-stress genes) illustrates its versatility for hypothesis-driven analyses. Identified interactions among kinases and stress-response transcripts emphasize the tool's value for mapping complex biological networks, and point to the need for evaluating both positive and negative feature associations when predicting phenotypic outcomes.

Compared to other multi-omics integration frameworks, panomiX is distinct in its focus on regression-based prediction of continuous phenotypes, whereas many established tools such as DIABLO or MOGONET are primarily for classification tasks. This feature expands its applicability to quantitative traits such as growth rates or stress indices, which are often more biologically informative than binary outcomes. It should be emphasized, however, that while individual models trained with panomiX are not directly transferable across experiments, the framework can be readily retrained on new datasets. Additional benchmarking on publicly available barley and oilseed rape population datasets further demonstrates this adaptability and supports the broader applicability of panomiX across different species, omics layers, and trait types. Beyond the prediction of agronomic/phenotypic traits, the use of SHAP values allows for the interpretation of model predictions by quantifying the contribution of each feature, thereby allowing to identify key genetic drivers linked to biological processes. This approach was exemplified in the oilseed rape dataset through the identification of promising candidate genes, such as PLEIOTROPIC REGULATORY LOCUS1 (PRL1), which has recently associated with actin microfilament

integrity and cell morphogenesis [82]. Furthermore, additional transcripts with high SHAP values (Table S5) suggest biological relevance to the trait under study. These include several xyloglucan endo-transglucosylases/hydrolases, which are cell wall remodeling enzymes involved in xyloglucan metabolism, and have found to be correlated with growth-related traits, including the projected leaf area, compactness, and early plant height [10].

As other statistical integration methods, generalization of the panomiX models is limited to datasets with the same variables measured. However, the applicability of the panomiX pipeline has been successfully demonstrated for various sources of variance; here complex time-series/treatment experiment, as well as a large-scale genetic diversity panel and a breeding population. While panomiX does not currently adjust for population structure or relatedness, its predictions are comparable with *e.g.* BayesB regression that includes this information explicitly in the model. In future implementations, incorporating a kinship matrix or population structure could further improve predictive performance in stratified populations. Additionally, environmental information could potentially be included as a monotonic constraint to guide predictions.

While panomiX offers significant advancements, several future challenges merit consideration. For example, in extending analyses to proteomics and metabolomics, peptide/protein ambiguities and metabolite identification uncertainties may affect the interpretation of features. Potential platform-specific biases and incomplete or uncertain annotations could influence pathway-level insights. Addressing these challenges through improved data curation steps, incorporation of databases, robust handling of missing values, and careful interpretation of predictive features will be important for broadening panomiX's applicability across diverse omics datasets.

### Code and data availability

The code for panomiX is freely available at https://github.com/NAMlab/panomiX-tool under the terms of the MIT license (also archived at Zenodo at time of publication: https://doi.org/10.5281/zenodo.15193421). The sequence data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB85881 (https://www.ebi.ac.uk/ena/browser/view/PRJEB85881). Phenotyping and FTIR data as well as preprocessed inputs for reproducing the results of this article with panomiX are available at https://doi.org/10.5447/ipk/2025/3.

### Author contributions

A.S., J.S., D.P., K.P., D.K., and N.T. wrote the manuscript with input from all the authors. D.P. planned, conducted, and coordinated the experimental work for transcriptomics analysis, FTIR analysis, and phenotyping under J.S. and K.N. supervision. A.H. performed RNA sequencing, H.R., and L.B. performed FTIR analysis. A.S. implemented SHAP-based monotonic constraints for machine learning models. A.S. developed the toolbox under J.S.'s supervision. A.S., and D.P. performed the data analysis under J.S.'s supervision. A.S., and D.K. performed the comparative analysis of oilseed rape data under J.S.'s supervision.

### Competing interests

The authors declare that they have no competing interests.

## Acknowledgements

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.plaphe.2025.100131.

## References

[1] D. Houle, D.R. Govindaraju, S. Omholt, Phenomics: the next challenge, Nat. Rev. Genet. 11 (12) (Dec. 2010) 855–866, https://doi.org/10.1038/nrg2897.

[2] R. Pieruschka, U. Schurr, Plant phenotyping: past, present, and future, Plant Phenomics 2019 (Mar. 2019) 7507131, https://doi.org/10.34133/2019/7507131.

[3] S. Ma, et al., WheatOmics: a platform combining multiple omics data to accelerate functional genomics studies in wheat, Mol. Plant 14 (12) (Dec. 2021) 1965–1968, https://doi.org/10.1016/j.molp.2021.10.006.

[4] W. Kong, et al., Genomic analysis of 1,325 Camellia accessions sheds light on agronomic and metabolic traits for tea plant improvement, Nat. Genet. (Mar. 2025) 1–11, https://doi.org/10.1038/s41588-025-02135-z.

[5] J. Szymański, et al., Analysis of wild tomato introgression lines elucidates the genetic basis of transcriptome and metabolome variation underlying fruit traits and pathogen response, Nat. Genet. 52 (10) (Oct. 2020) 1111–1121, https://doi.org/10.1038/s41588-020-0690-6.

[6] J. Chen, Q. Li, D. Jiang, From images to loci: applying 3D deep learning to enable multivariate and multitemporal digital phenotyping and mapping the genetics underlying nitrogen use efficiency in wheat, Plant Phenomics 6 (Dec. 2024) 270, https://doi.org/10.34133/plantphenomics.0270.

[7] Q. Lou, et al., Phenomics-assisted genetic dissection and molecular design of drought resistance in rice, Plant Commun. 6 (3) (Mar. 2025), https://doi.org/10.1016/j.xplc.2024.101218.

[8] C. Fu, et al., Transcriptomic and methylomic analyses provide insights into the molecular mechanism and prediction of heterosis in rice, Plant J. 115 (1) (2023) 139–154, https://doi.org/10.1111/tpj.16217.

[9] T.B. Mersha, From Mendel to multi-omics: shifting paradigms, Eur. J. Hum. Genet. 32 (2) (Feb. 2024) 139–142, https://doi.org/10.1038/s41431-023-01420-x.

[10] D. Knoch, et al., Integrated multi-omics analyses and genome-wide association studies reveal prime candidate genes of metabolic and vegetative growth variation in canola, Plant J. (2024) 713–728, https://doi.org/10.1111/tpj.16524, 117, no. 3.

[11] M.Y. Hirai, et al., Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*, Proc. Natl. Acad. Sci. (July 2004) 10205–10210, https://doi.org/10.1073/pnas.0403218101, 101, no. 27.

[12] L. Nazari, M.F. Aslan, K. Sabanci, E. Ropelewska, Integrated transcriptomic meta-analysis and comparative artificial intelligence models in maize under biotic stress, Sci. Rep. 13 (1) (2023) 15899, https://doi.org/10.1038/s41598-023-42984-4. Sept.

[13] M. Pardo-Hernández, P. García-Pérez, L. Lucini, R.M. Rivero, Multi-omics exploration of the involvement of ABA in identifying unique molecular markers for single and combined stresses in tomato plants, J. Exp. Bot. (2024) erae372, https://doi.org/10.1093/jxb/erae372. Sept.

[14] M. Zagorščak, et al., Integration of multi-omics data and deep phenotyping provides insights into responses to single and combined abiotic stress in potato, Plant Physiol. 197 (4) (Mar. 2025) kiaf126, https://doi.org/10.1093/plphys/kiaf126.

[15] J. Sia, W. Zhang, M. Cheng, P. Bogdan, D.E. Cook, Machine learning-based identification of general transcriptional predictors for plant disease, New Phytol. 245 (2) (Jan. 2025) 785–806, https://doi.org/10.1111/nph.20264.

[16] S. Liu, et al., Identification of salt-responsive genetic variants using cross-condition multi-omics association analysis in maize, Plant Commun. 6 (3) (Mar. 2025) 101219, https://doi.org/10.1016/j.xplc.2024.101219.

[17] R. Wonneberger, et al., Integrating metabolomics and high-throughput phenotyping to elucidate metabolic and phenotypic responses to early-season drought stress in Nordic spring wheat, BMC Plant Biol. 25 (1) (July 2025) 987, https://doi.org/10.1186/s12870-025-06914-y.

[18] D. Feldner-Busztin, et al., Dealing with dimensionality: the application of machine learning to multi-omics data, Bioinformatics 39 (2) (2023), https://doi.org/10.1093/bioinformatics/btad021.

[19] Y. Luo, C. Zhao, F. Chen, Multiomics research: principles and challenges in integrated analysis, BioDesign Res. 6 (Jan. 2024) 59, https://doi.org/10.34133/bdr.0059.

[20] B. Wang, et al., Similarity network fusion for aggregating data types on a genomic scale, Nat. Methods 11 (3) (2014) 333–337, https://doi.org/10.1038/nmeth.2810.

[21] C. Dimitrakopoulos, et al., Network-based integration of multi-omics data for prioritizing cancer genes, Bioinformatics 34 (14) (2018) 2441–2448, https://doi.org/10.1093/bioinformatics/bty148.

[22] C.J. Vaske, et al., Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM, Bioinformatics 26 (12) (2010) i237–i245, https://doi.org/10.1093/bioinformatics/btq182.

[23] R. Louhimo, S. Hautaniemi, CNAmet: an R package for integrating copy number, methylation and expression data, Bioinformatics 27 (6) (2011) 887–888, https://doi.org/10.1093/bioinformatics/btr019.

[24] H. Nguyen, S. Shrestha, S. Draghici, T. Nguyen, PINSPlus: a tool for tumor subtype discovery in integrated genomic data, Bioinformatics 35 (16) (2019) 2843–2846, https://doi.org/10.1093/bioinformatics/bty1049.

[25] N. Rappoport, R. Shamir, NEMO: cancer subtyping by integration of partial multi-omic data, Bioinformatics 35 (18) (2019) 3348–3356, https://doi.org/10.1093/bioinformatics/btz058.

[26] A. Orlichenko, et al., Latent similarity identifies important functional connections for phenotype prediction, IEEE Trans. Biomed. Eng. 70 (6) (June 2023) 1979–1989, https://doi.org/10.1109/TBME.2022.3232964.

[27] R. Argelaguet, et al., Multi-omics factor Analysis—A framework for unsupervised integration of multi-omics data sets, Mol. Syst. Biol. 14 (6) (June 2018) e8124, https://doi.org/10.15252/msb.20178124.

[28] R. Shen, et al., Integrative subtype discovery in glioblastoma using iCluster, PLoS One 7 (4) (2012) e35236, doi: 10.1371/journal.pone.0035236.

[29] P. Ray, L. Zheng, J. Lucas, L. Carin, Bayesian joint analysis of heterogeneous genomics data, Bioinformatics 30 (10) (2014) 1370–1376, https://doi.org/10.1093/bioinformatics/btu064.

[30] C. Zhang, Y. Chen, T. Zeng, C. Zhang, L. Chen, Deep latent space fusion for adaptive representation of heterogeneous multi-omics data, Brief. Bioinform. 23 (2) (Mar. 2022), https://doi.org/10.1093/bib/bbab600 bbab600.

[31] C. Meng, D. Helm, M. Frejno, B. Kuster, moCluster: identifying joint patterns across multiple omics data sets, J. Proteome Res. 15 (3) (Mar. 2016) 755–765, https://doi.org/10.1021/acs.jproteome.5b00824.

[32] F. Rohart, B. Gautier, A. Singh, K.-A. Lê Cao, mixOmics: an R package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (11) (2017) e1005752, https://doi.org/10.1371/journal.pcbi.1005752.

[33] K. Kaitoh, Y. Yamanishi, TRIOMPHE: Transcriptome-Based inference and generation of molecules with desired phenotypes by machine learning, J. Chem. Inf. Model. 61 (9) (2021) 4303–4320, https://doi.org/10.1021/acs.jcim.1c00967. Sept.

[34] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, Aug. 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[35] H. Chen, I.C. Covert, S.M. Lundberg, S.-I. Lee, Algorithms to estimate shapley value feature attributions, Nat. Mach. Intell. 5 (6) (2023) 590–601, https://doi.org/10.1038/s42256-023-00657-x.

[36] D. Camejo, P. Rodríguez, M. Angeles Morales, J. Miguel Dell'Amico, A. Torrecillas, J.J. Alarcón, High temperature effects on photosynthetic activity of two tomato cultivars with different heat susceptibility, J. Plant Physiol. 162 (3) (Mar. 2005) 281–289, https://doi.org/10.1016/j.jplph.2004.07.014.

[37] C. Bita, T. Gerats, Plant tolerance to high temperature in a changing environment: scientific fundamentals and production of heat stress-tolerant crops, Front. Plant Sci. 4 (July 2013), https://doi.org/10.3389/fpls.2013.00273.

[38] M.M. Raja, et al., Pollen development and function under heat stress: from effects to responses, Acta Physiol. Plant. 41 (4) (Mar. 2019) 47, https://doi.org/10.1007/s11738-019-2835-8.

[39] C. Jonak, S. Kiegerl, W. Ligterink, P.J. Barker, N.S. Huskisson, H. Hirt, Stress signaling in plants: a mitogen-activated protein kinase pathway is activated by cold and drought, Proc. Natl. Acad. Sci. 93 (20) (Oct. 1996) 11274–11279, https://doi.org/10.1073/pnas.93.20.11274.

[40] Y.-X. Lin, H.-Y. Jiang, Z.-X. Chu, X.-L. Tang, S.-W. Zhu, B.-J. Cheng, Genome-wide identification, classification and analysis of heat shock transcription factor family in maize, BMC Genom. 12 (1) (Jan. 2011) 76, https://doi.org/10.1186/1471-2164-12-76.

[41] H. Tanaka, et al., Abiotic stress-inducible receptor-like kinases negatively control ABA signaling in arabidopsis, Plant J. 70 (4) (2012) 599–613, https://doi.org/10.1111/j.1365-313X.2012.04901.x.

[42] Y. Liu, Z. Feng, W. Zhu, J. Liu, Y. Zhang, Genome-wide identification and characterization of cysteine-rich receptor-like protein kinase genes in tomato and their expression profile in response to heat stress, Diversity (June 2021), https://doi.org/10.3390/d13060258, 13, no. 6, Art. no. 6.

[43] S. Mo, et al., Mitogen-activated protein kinase action in plant response to high-temperature stress: a mini review, Protoplasma 258 (3) (May 2021) 477–482, https://doi.org/10.1007/s00709-020-01603-z.

[44] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15 (12) (Dec. 2014) 550, https://doi.org/10.1186/s13059-014-0550-8.

[45] K.H. Liland, B.-H. Mevik, Baseline: Baseline Correction of Spectra, 2011, https://doi.org/10.32614/CRAN.package.baseline, 1.3-5, Jan. 10.

[46] U. Ligges, T. Short, P. Kienzle, Signal: Signal Processing 1.vols. 8–1, 2006, https://doi.org/10.32614/CRAN.package.signal. Dec. 10.

[47] M.U.A. Bromba, Horst Ziegler, Application hints for Savitzky-Golay digital smoothing filters, Anal. Chem. 53 (11) (Sept. 1981) 1583–1586, https://doi.org/10.1021/ac00234a011.

[48] J. Baglama, L. Reichel, B.W. Lewis, Irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices, May 27, 2011, https://doi.org/10.32614/CRAN.package.irlba, 2.3.5.1.

[49] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of Statistical Software 28 (5) (2008) 1–26. https://doi.org/10.18637/jss.v028.i05.

[50] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (2010) 1–13, https://doi.org/10.18637/jss.v036.i11.

[51] D. Attali, Shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds, Apr. 22, 2015, https://doi.org/10.32614/CRAN.package.shinyjs, 2.1.0.

[52] C. Sievert, et al., Plotly: Create Interactive Web Graphics via 'Plotly.Js, Nov. 17, 2015, https://doi.org/10.32614/CRAN.package.plotly, 4.10.4,.

[53] Y. Xie, J. Cheng, X. Tan, DT: a Wrapper of the Javascript Library 'DataTables, June 09, 2015, https://doi.org/10.32614/CRAN.package.DT, 0.33.

[54] D. Attali, A. Sali, Shinycssloaders: Add Loading Animations to a 'Shiny' Output While It's Recalculating, May 10, 2017, pp. 1–1.0, https://doi.org/10.32614/CRAN.package.shinycssloaders.

[55] C. Sievert, J. Cheng, G. Aden-Buie, Bslib: Custom 'Bootstrap' 'Sass' Themes for 'Shiny' and 'Rmarkdown', 2021, https://doi.org/10.32614/CRAN.package.bslib, 0.9.0, Jan. 25.

[56] C. Sievert, Bsicons: Easily Work with 'Bootstrap' Icons, 2022, https://doi.org/10.32614/CRAN.package.bsicons, 0.1.2, Nov. 22.

[57] R. Ihaka, R. Gentleman, "R: a Language for Data Analysis and Graphics,", J. Comput. Graph Stat. 5 (3) (1996) 299–314, https://doi.org/10.1080/10618600.1996.10474713. Sept.

[58] W. Chang, et al., Shiny: Web Application Framework for R, 2012, https://doi.org/10.32614/CRAN.package.shiny.

[59] A. Atkins, T. Allen, H. Wickham, J. McPherson, J. Allaire, "Rsconnect: Deploy Docs, 2016, https://doi.org/10.32614/CRAN.package.rsconnect. Apps, and APIs to 'Posit Connect', 'shinyapps.io', and 'RPubs.'" p. 1.3.4, Mar. 20.

[60] C. Klukas, D. Chen, J.-M. Pape, Integrated analysis platform: an open-source information System for high-throughput plant phenotyping, Plant Physiol. 165 (2) (June 2014) 506–518, https://doi.org/10.1104/pp.113.233932.

[61] D. Arend, et al., Quantitative monitoring of Arabidopsis thaliana growth and development using high-throughput plant phenotyping, Sci. Data 3 (1) (Aug. 2016) 160055, https://doi.org/10.1038/sdata.2016.55.

[62] N.L. Bray, H. Pimentel, P. Melsted, L. Pachter, Near-optimal probabilistic RNA-seq quantification, Nat. Biotechnol. 34 (5) (May 2016) 525–527, https://doi.org/10.1038/nbt.3519.

[63] D. Psaroudakis, et al., Identification of core, conditional and crosstalk components of tomato heat stress response using integrative transcriptomics and orthology, Research Square (2024), https://doi.org/10.21203/rs.3.rs-4337825/v1.

[64] A. Guendel, A. Hilo, H. Rolletschek, L. Borisjuk, "Probing the Metabolic Landscape of Plant Vascular Bundles by Infrared Fingerprint Analysis, Imaging and Mass Spectrometry,", Biomolecules (Nov. 2021) https://doi.org/10.3390/biom11111717, 11, no. 11.

[65] J.R. Adrian Alexa, topGO. Bioconductor, 2017, https://doi.org/10.18129/B9.BIOC.TOPGO.

[66] S. Zarbakhsh, A.R. Shahsavar, Exogenous γ-aminobutyric acid improves the photosynthesis efficiency, soluble sugar contents, and mineral nutrients in pomegranate plants exposed to drought, salinity, and drought-salinity stresses, BMC Plant Biol. 23 (1) (Nov. 2023) 543, https://doi.org/10.1186/s12870-023-04568-2.

[67] S. Singh, et al., Identification of key genes and molecular pathways regulating heat stress tolerance in pearl millet to sustain productivity in challenging ecologies, Front. Plant Sci. 15 (2024) 1443681, https://doi.org/10.3389/fpls.2024.1443681.

[68] M. Koul, L. Thomas, K. Karmakar, Functional aspects of solanaceae trichomes in heavy metal detoxification, Nord. J. Bot. 39 (5) (May 2021) 03171, https://doi.org/10.1111/njb.03171 njb.

[69] Y. Zheng, et al., iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases, Mol. Plant 9 (12) (2016) 1667–1670, https://doi.org/10.1016/j.molp.2016.09.014.

[70] S. Pasqualini, et al., Ozone and nitric oxide induce cGMP-dependent and -independent transcription of defence genes in tobacco, New Phytol. 181 (4) (Mar. 2009) 860–870, https://doi.org/10.1111/j.1469-8137.2008.02711.x.

[71] J. Li, H. Jia, J. Wang, cGMP and ethylene are involved in maintaining ion homeostasis under salt stress in arabidopsis roots, Plant Cell Rep. 33 (3) (Mar. 2014) 447–459, https://doi.org/10.1007/s00299-013-1545-8.

[72] D. Qian, M. Wang, Y. Niu, Y. Yang, Y. Xiang, Sexual reproduction in plants under high temperature and drought stress, Cell Rep. 44 (3) (Mar. 2025) 115390, https://doi.org/10.1016/j.celrep.2025.115390.

[73] Y. Feng, et al., Preventive regulation of cellular heat stress injury by mung bean polyphenols revealed by multi-omics analysis, Npj Sci. Food 9 (1) (July 2025) 163, https://doi.org/10.1038/s41538-025-00522-8.

[74] H. Rahman, X.-Y. Wang, Y.-P. Xu, Y.-H. He, X.-Z. Cai, Characterization of tomato protein kinases embedding guanylate cyclase catalytic center motif, Sci. Rep. 10 (1) (2020) 4078, https://doi.org/10.1038/s41598-020-61000-7.

[75] S. Graci, A. Barone, Tomato plant response to heat stress: a focus on candidate genes for yield-related traits, Front. Plant Sci. 14 (2024) 1245661, https://doi.org/10.3389/fpls.2023.1245661.

[76] V. Berková, et al., The fungus *Acremonium alternatum* enhances salt stress tolerance by regulating host redox homeostasis and phytohormone signaling, Physiol. Plant. 176 (3) (May 2024) e14328, https://doi.org/10.1111/ppl.14328.

[77] Y. Zhao, et al., The calcium-dependent protein kinase ZmCDPK7 functions in heat-stress tolerance in maize, J. Integr. Plant Biol. 63 (3) (Mar. 2021) 510–527, https://doi.org/10.1111/jipb.13056.

[78] J. Zhang, et al., Releasing a sugar brake generates sweeter tomato without yield penalty, Nature (2024) 1–10, https://doi.org/10.1038/s41586-024-08186-2.

[79] A. Singh, et al., DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays, Bioinformatics 35 (17) (Sept. 2019) 3055–3062, https://doi.org/10.1093/bioinformatics/bty1054.

[80] T. Wang, et al., MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification, Nat. Commun. 12 (1) (June 2021) 3445, https://doi.org/10.1038/s41467-021-23774-w.

[81] M.R. Gemmer, et al., Can metabolic prediction be an alternative to genomic prediction in barley? PLoS One 15 (6) (June 2020) e0234052 https://doi.org/10.1371/journal.pone.0234052.

[82] X. Gao, et al., PLEIOTROPIC REGULATORY LOCUS1 maintains actin microfilament integrity to regulate pavement cell morphogenesis, Plant Physiol. 195 (1) (Apr. 2024) 356–369, https://doi.org/10.1093/plphys/kiae018.