Computational and Structural Biotechnology Journal 23 (2024) 3454–3466

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Review Article

Confronting the data deluge: How artificial intelligence can be used in the study of plant stress

Eugene Koh, Rohan Shawn Sunil, Hilbert Yuen In Lam, Marek Mutwil *

School of Biological Scie nces, Nanyang Technological University, Singapore, Singapore

ARTICLE INFO

Keywords:
Plant stress resilience
Large-scale data
Artificial intelligence
Large language models

ABSTRACT

The advent of the genomics era enabled the generation of high-throughput data and computational methods that serve as powerful hypothesis-generating tools to understand the genomic and gene functional basis of plant stress resilience. The proliferation of experimental and analytical methods used in biology has resulted in a situation where plentiful data exists, but the volume and heterogeneity of this data has made analysis a significant challenge. Current advanced deep-learning models have displayed an unprecedented level of comprehension and problem-solving ability, and have been used to predict gene structure, function and expression based on DNA or protein sequence, and prominently also their use in high-throughput phenomics in agriculture. However, the application of deep-learning models to understand gene regulatory and signalling behaviour is still in its infancy. We discuss in this review the availability of data resources and bioinformatic tools, and several applications of these advanced ML/AI models in the context of plant stress response, and demonstrate the use of a publicly available LLM (ChatGPT) to derive a knowledge graph of various experimental and computational methods used in the study of plant stress. We hope this will stimulate further interest in collaboration between computer scientists, computational biologists and plant scientists to distil the deluge of genomic, transcriptomic, proteomic, metabolomic and phenomic data into meaningful knowledge that can be used for the benefit of humanity.

1. Introduction

The intricate relationship between plants and their environment forms the bedrock of ecological balance and sustenance. In the face of evolving climatic patterns, burgeoning populations and environmental degradation, understanding how plants respond to stress has become a matter of paramount significance [1]. Environmental stress, encompassing both abiotic and biotic factors, poses formidable challenges to plant life, affecting growth, productivity and overall ecological resilience. The repercussions of these stressors extend beyond individual plants, influencing entire ecosystems and, consequently, the global environment [2]. As we navigate an era marked by climate change and global ecological shifts, the ability to discern how various plant species adapt and thrive—or falter—under diverse stress conditions is pivotal. This understanding not only informs sustainable agricultural practices but also holds the key to preserving biodiversity, ensuring ecosystem resilience, and ultimately, securing our planet's environmental well-being.

While stress is intensively studied by molecular biologists who have already uncovered multiple genes and biological processes underpinning stress resilience mechanisms [3–5], computational biologists can also use advances in large-scale data generation to form powerful predictive models about gene function [6]. Various machine learning (ML) models were developed to aid in this process with varying amounts of success, but still required significant amounts of expert manual curation, especially for heterogeneous data. The development and proliferation of powerful deep-learning models which are able to process more complex and heterogeneous datasets have provided scientists with new tools to interrogate and interpret the vast reams of data produced by high-throughput omics methods. In this review, we will cover how large-scale data can be used to fuel powerful AI approaches able to resolve the complex biological mechanisms underlying plant-environment interactions and advance our understanding of plant stress resilience.

* Correspondence to: School of Biological Sciences Nanyang Technological University, 60 Nanyang Drive, 637551, Singapore.
E-mail address: mutwil@ntu.edu.sg (M. Mutwil).

https://doi.org/10.1016/j.csbj.2024.09.010
Received 31 July 2024; Received in revised form 14 September 2024; Accepted 16 September 2024
Available online 17 September 2024

2. Kingdom-level overview of genomic and transcriptomic plant stress studies

Genomic and transcriptomic data are invaluable to understanding gene functions and how organisms respond to the changing environment. Databases such as the Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) play pivotal roles in conducting meta-analyses of environmental stress studies across the plant kingdom [7,8]. These repositories serve as invaluable reservoirs of high-throughput sequencing data, transcriptomic profiles, and other omics datasets, providing researchers access to a wealth of information derived from diverse experimental setups [9]. The GEO database was launched by NCBI in 2000 as a repository for high-throughput gene expression data [7], which were dominated initially by data from DNA microarray experiments. Today, abundant omics data from protein or tissue arrays, genome methylation, genome binding/occupancy, protein profiling, chromosome conformation, and genome variation/copy number studies can also be found in GEO [10].

The SRA was established as a public repository for next-generation sequence data as a part of the International Nucleotide Sequence Database Collaboration (INSDC) [11], that acts as a nexus collating raw next-generation genomic and transcriptomic sequencing data, and is closely integrated with other large databases such as the ArrayExpress at EBI (http://www.ebi.ac.uk/arrayexpress) [12] and DDBJ Omics Archive (http://trace.ddbj.nig.ac.jp/dor) [13] hosted by the European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ). Both

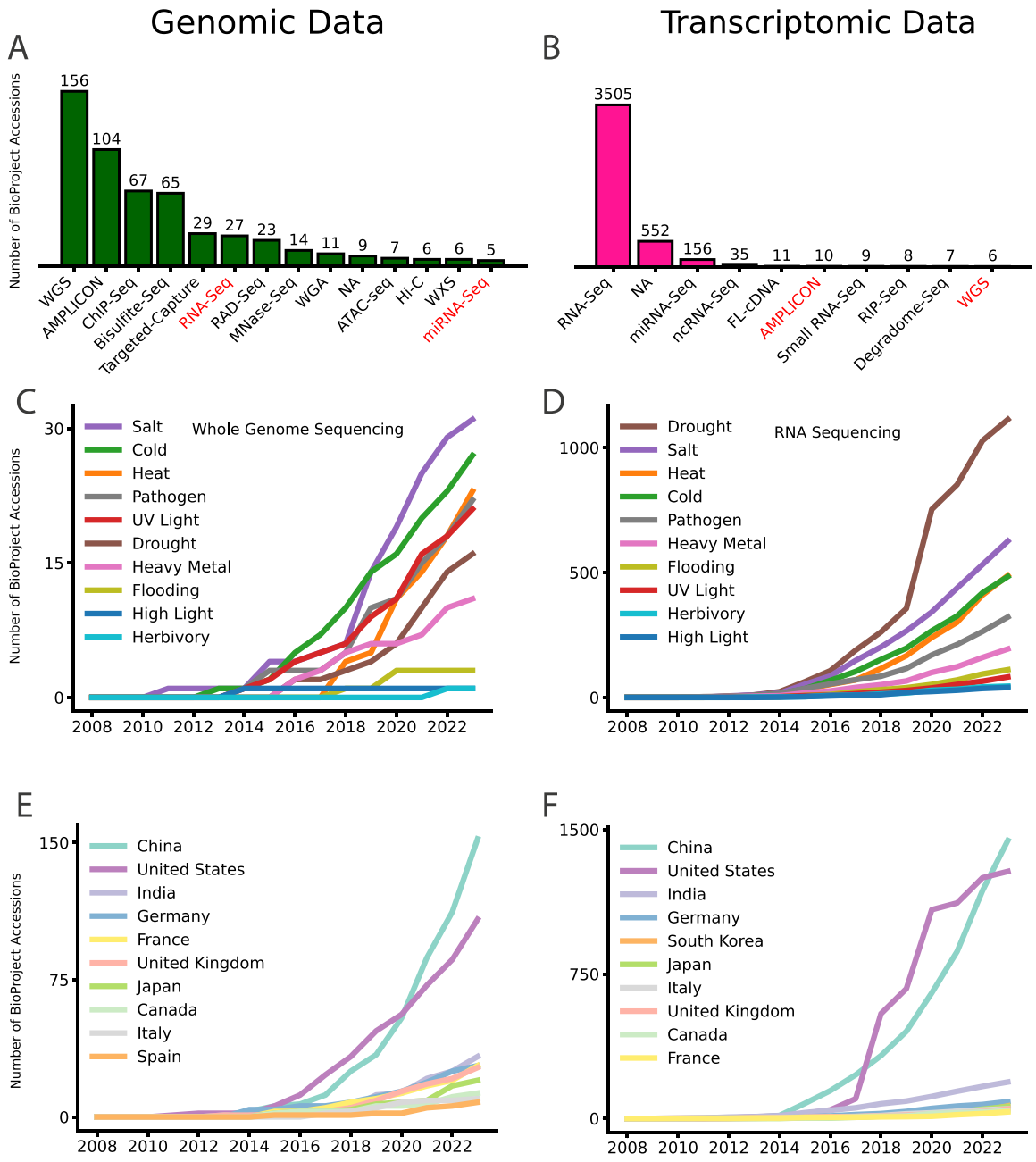

Fig. 1. Overview of stress-related BioProject IDs present in the SRA database for genomic and transcriptomic studies. (A, B) After filtering for stress-related BioProject IDs based on our keywords, BioProjects were classified based on the sequencing type used, only displaying those with > 5 BioProjects. Misannotated sequencing types are labelled in red. (C, D) The cumulative number of BioProjects in SRA over the years were also classified by stress type, and (E, F) country of origin. Figure legends show the respective ranks of the cumulative total number of BioProjects.

databases share significant overlaps, with new accessions in GEO automatically and simultaneously uploaded to SRA as well, but not vice versa. Besides the biological data present in these databases, plentiful metadata also exists in these databases which provide contextual information on the purpose of the studies, or detailed information on specific treatment conditions, plant organ type, species / genotype / cultivar variations, plant age / developmental stage, sample geographical origin and others. These metadata can be further analysed to identify specific research trends of scientific interest, or which may be useful for funding bodies and policymakers.

In order to derive a global overview of the data available from stress studies across the plant kingdom, we began by querying for the term 'Viridiplantae' within the SRA database, which encompasses two major clades: Streptophyta, which contains streptophyte algae and land plants (embryophytes), and Chlorophyta, which includes all remaining green algae. We observed that the SRA houses a total of 558,292 transcriptomic and 1,387,717 genomic accessions for Viridiplantae at the time of writing (May 2024), and we performed a preliminary grouping of this data into their respective unique BioProject accessions. The BioProject database is a comprehensive and centralised repository managed by the National Center for Biotechnology Information (NCBI) that serves as a platform for the organisation and dissemination of information related to biological research projects [14]. Each BioProject entry includes detailed information about the project's objectives, experimental design, and data sources. The metadata available at this step allowed us to filter for experiments conducted in stress studies encompassing 10 specific stress types we had defined: high light, heat, cold, UV light, salt, drought, heavy metal, pathogen, flooding and herbivory. This was achieved by parsing the metadata with keywords specific to each of the 10 stress types. In total, we retrieved 4323 and 592 stress-related BioProjects for transcriptomic and genomic sets, respectively. By inspecting the metadata, we could determine the types of experimental procedures performed to obtain the respective genomic and transcriptomic data. As expected, whole genome sequencing (WGS) and RNA-Seq were the most common methods used to obtain genomic and transcriptomic data (Fig. 1). For genomic data, we observed a significant number of BioProjects which were tagged with epigenomic type studies performed with ChIP-Seq, BiSulfite-Seq, ATAC-Seq etc, which would highlight the studies characterising the various stress-induced epigenomic changes. The WGS accessions obtained are possibly part of larger studies for previously unsequenced species. For transcriptomic data, we observed that microRNA (miRNA) sequencing, noncoding RNA (ncRNA) sequencing, full-length cDNA (FL-cDNA) sequencing and small RNA sequencing were the next most popular transcriptomic methods. We noticed that there was a significant number of unannotated or incorrectly annotated entries, for instance, RNA-Seq or miRNA-Seq (coloured in red) being tagged in genomic data or unclassified samples (NA), WGS and Amplicon sequencing being tagged in transcriptomic data.

2.1. Drought, salt, heat, cold and pathogen stress dominate the stress study landscape

Next, to understand which stresses are most studied, we plotted the number of unique BioProject IDs annotated with specific stress-related keywords. We observed that drought, salt, heat, cold and pathogen stress were the top five most common stress types for genomic (Fig. 1C) and transcriptomic (Fig. 1D) data. Of note is the observation that the accumulation of genomic and transcriptomic data present in the SRA began around 2010–2012, when next-generation sequencing technologies such as RNA-Seq came to the fore, supplanting the previous generation of transcriptomic technologies such as microarrays. Transcriptomic data derived from microarrays were previously stored in repositories such as ArrayExpress and the Gene Expression Omnibus (GEO) databases but are not represented in our chart.

2.2. Two nations produce > 80 % of all available data

The metadata present in our dataset also allows us to plot the progress of scientific investment into plant stress research in a country-specific manner, using the submitting institutions as a proxy (Fig. 1). From the charts, we observed that the United States, China and India were the top three drivers of plant stress research, with the United States having led for close to a decade but has been recently overtaken by China. The interest shown by both superpowers in understanding the mechanisms involved in plant stress signalling is an encouraging sign for the field and is extremely timely considering the looming effects of climate change affecting global food security. Other members include developed countries in Europe and Asia, such as Germany, France, the United Kingdom, Italy, Spain, Japan and South Korea. Of note are the obvious absences of developing countries and those from the global South, which hints at a potential widening divide between the rich and poor nations of the world in preparation for the incoming climate shifts [15].

2.3. Species- and stress-specific research trends

Next, we investigated how the various species were used for stress research. To that end, we identified the top 10 species per stress type that were studied by genomic (Fig. 2) and transcriptomic (Fig. 3) approaches and divided the species based on their uses (e.g., crop, model, ornamental). Overall, food crops were most comprehensively represented (27 out of 82 species), followed by model plants (18 out of 82) and timber crops (10 out of 82). For transcriptomic data, heat, cold, salt, drought and pathogen stress were the most abundant (>100 BioProject accessions), followed by UV, heavy metal and flooding stress (50–100 BioProject accessions), and high light and herbivory stress (<50 BioProject accessions) (Fig. 3). It was unsurprising to note that the model plant *Arabidopsis thaliana* dominated for most stress types, with the exception of drought stress, where *Brachypodium distachyon* and *Brassica napus* led the field; and herbivory, where *Zea mays* came in slightly ahead of Arabidopsis (Fig. 3). Since Arabidopsis contained the highest number of samples, we decided to identify species coming in second to fifth places to reveal current research efforts. We observed that the crop plants *Triticum aestivum*, *Oryza sativa*, *Solanum lycopersicum*, *Glycine max* and *Nicotiana tabacum* were strongly represented in the majority of the stress types for genomic (Fig. 2) and transcriptomic (Fig. 3) data. Of interest was the presence of *Kalanchoe fedtschenkoi* in second place in heat/cold stress and *Panicum virgatum* in fifth place in drought stress, but with still more than 50 BioProject accessions. Kalanchoe is a common ornamental houseplant [16], while Panicum is used as cattle feedstock [17] or also as ground cover to prevent soil erosion [18]. The position of Kalanchoe in second place for heat/cold stress was unexpected, which prompted us to examine the source data more closely. We found that the observed 'overrepresentation' of Kalanchoe was due to individual accessions from the same submitting institution being assigned with unique BioProject IDs in SRA. This is a distinct example of the difficulty of standardising and curation of metadata due to inadvertent human error and other inconsistencies. Nevertheless, the rise in atypical species such as these is an encouraging sign that scientific knowledge, interests and resources are percolating through the system and leading to more translational discoveries.

High light research was conducted to a large extent on Arabidopsis (Figs. 2 and 3), but this stress was surprisingly not studied in crop plants, and only a few datasets were present in other model organisms. On the other hand, UV research was also dominated by Arabidopsis, but with several common crop plants like *Prunus persica*, *Malus domestica*, *Camellia sinensis* and even *Ginkgo biloba*, known for its medicinal properties, also in the mix. Climate change can come with changing rainfall patterns and consequently decreased cloud cover, while geographical locations at risk of desertification are also exposed to such stresses [19]. High intensities or prolonged duration of visible or UV light can cause

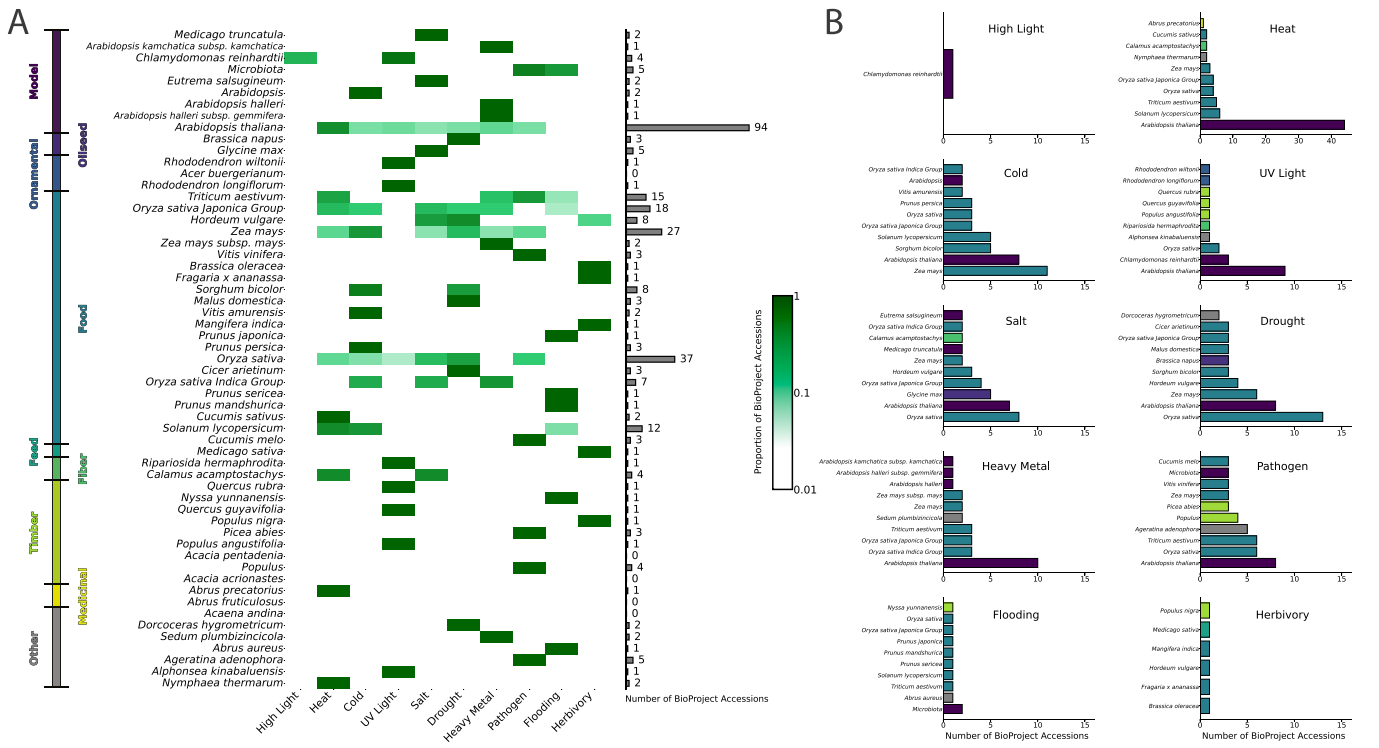

Fig. 2. Distribution of top 10 species per stress studied across stress-related BioProject IDs for genomic data. The counts of different species studied across the BioProject IDs for each stress type are represented in a (A) log-scale heat map normalised within each species, and in (B) bar charts showing the top few species. Some stresses contained < 10 species studied.

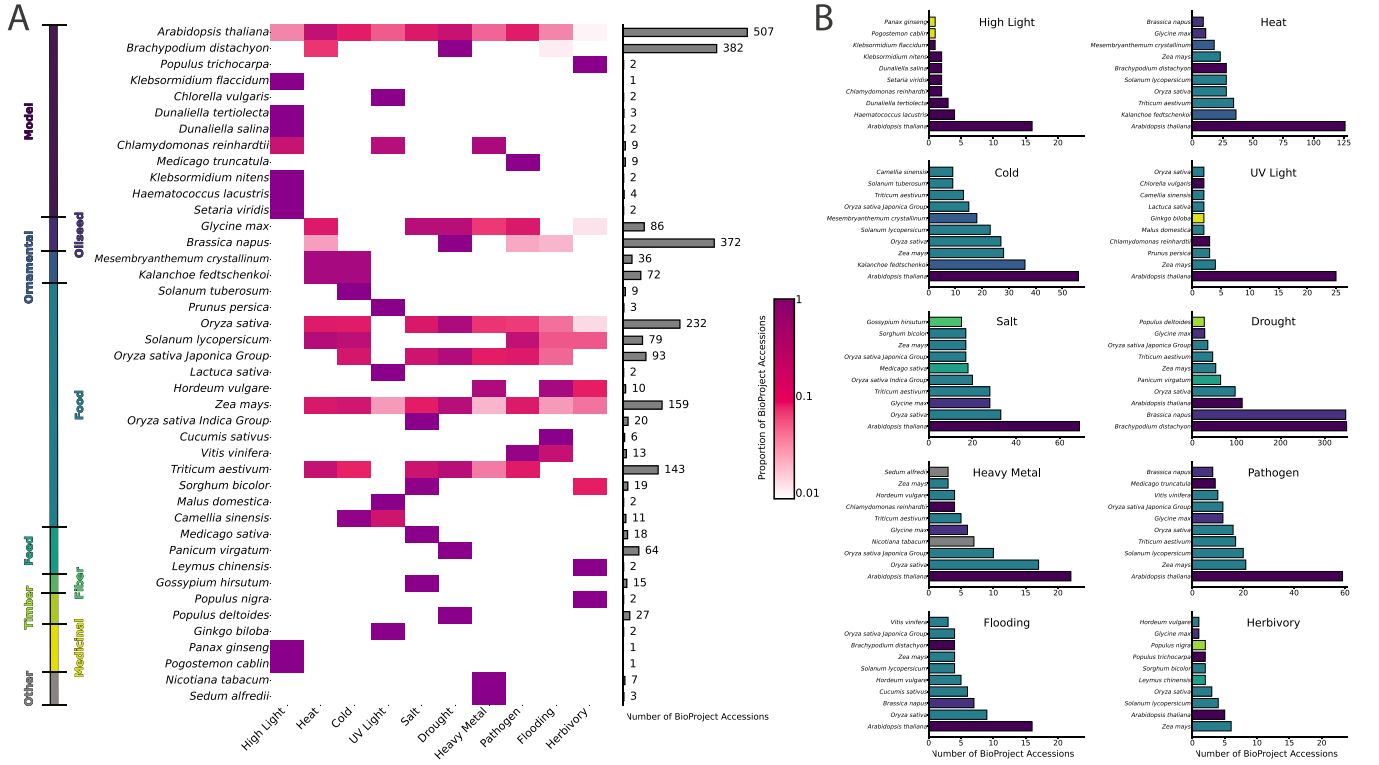

Fig. 3. Distribution of top 10 species per stress studied across stress-related BioProject IDs for transcriptomic data. The counts of different species studied across the BioProject IDs for each stress type are represented in a (A) log-scale heat map normalised within each species, and in (B) bar charts showing the top few species.

significant Reactive Oxygen Species (ROS) production, resulting in damage to various biomolecules such as DNA/RNA, proteins and lipids, resulting in cellular dysfunction and the accumulation of mutations [20]. We believe that increasing focus on these relatively neglected areas of research would be pivotal in understanding the chronic biological effects of these stresses [21].

For genomic data, other than heat stress, which had > 50 BioProject accessions, we found < 30 BioProject accessions for the remaining stress types, with there being even fewer (<5 BioProject accessions) for high light, flooding and herbivory stress in particular (Fig. 2). Overall, this showcases considerably fewer stress-related genomic studies as compared with those utilising transcriptomic data. From a fundamental standpoint, this does make sense, as changes in gene expression would be one of the main focuses for studies investigating the effect of various stresses on plants. Again, *Arabidopsis thaliana* was a key contributor in seven of the stress types in terms of genomic data, with *Zea mays* being the most studied for cold stress and *Oryza sativa* for salt and drought stress.

2.4. Online resources providing analyses of large-scale data

A vast variety of bioinformatic tools were developed that utilise the genomic, epigenomic, transcriptomic, proteomic, and other '-omic' data [22]. For genomics, the 1001 Genomes Project capturing natural variation data allows the exploration of 462 phenotypes across 1496 accessions in *Arabidopsis thaliana* [23]. One of the most widely used portals, TAIR (http://www.arabidopsis.org, [24]), provides detailed information about gene functions and maintains a 'super-portal' to keep track of and categorise various Arabidopsis tools (https://conf.arabidopsis.org/display/COM/Resources). Databases such as Ensembl Plants [25], PLAZA [26] and PANTHER [27], allow the exploration of gene families and gene trees. Epigenomic DNA modifications from numerous sequencing experiments can be viewed in the EPIC-CoGe Browser [28] and the 1001 Epigenomes Browser [29], giving the researchers unprecedented means to study the epigenetic regulation of genes.

Transcriptomic tools allow the analysis of expression profiles across publicly available transcriptomic data comprised of RNA-sequencing and microarrays [9], and are accessible through sites such as the eFP browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi, [30]), Arabidopsis RNA-Seq database (http://ipf.sustech.edu.cn/pub/athrna, [31]), Plant Single Cell RNA-Sequencing Database (https://www.zmbp-resources.uni-tuebingen.de/timmermans/plant-single-cell-browser/; [32]), CoNekT (https://conekt.sbs.ntu.edu.sg/, [33]) and Plant Expression Omnibus (https://expression.plant.tools/, [34]). Since genes with similar expression patterns (co-expression) can be functionally related, several tools that explore co-expression networks for Arabidopsis and other species are available. The tools include ATTED-II (https://atted.jp/, [35]), CoNekT (https://conekt.sbs.ntu.edu.sg/, [33]), Expression Angler (https://bar.utoronto.ca/ExpressionAngler/, [36]), and others.

Proteomic tools range from methods that study protein subcellular localisation (https://version4legacy.suba.live/, [37]), protein modifications, such as phosphorylation, acetylation, methylation, nitrosylation, ubiquitination and glycosylation (http://p3db.org, http://www.psb.ugent.be/PlantPTMViewer, [38,39]). Since interacting proteins are also likely functionally related, tools to visualise protein-protein interactions, such as BAR's Arabidopsis Interactions Viewer 2 (https://bar.utoronto.ca/interactions2/, [40]) and AtPID (http://www.megabionet.org/atpid/, [41]). Finally, with the advances in AI, accurate protein structures can be predicted (https://alphafold.ebi.ac.uk, [42]), opening new possibilities for structure-function studies.

There are a number of plant stress specific databases which have been published, which are based on extensive manual curation, and cater to a variety of species and stress types. A majority are transcriptomics-based (Plant stress RNA-seq Nexus (PSRN) [43], Plant Environmental stress transcript database [44], Plant Resistance Genes database (PRGdb 4.0) [45], DroughtDB [46], Stress Responsive Transcription Factor Database (STIFDB V2.0) [47], CerealESTDb [48]) or genomic databases (HEATSTER [49], Rice Stress-Resistant SNP Database [50]), but more specialised databases focusing on specific non-coding RNAs such as (PncStress database [51], CropCircDB [52]) or alternative splicing events (PlaASDB [53]) also exist in the literature. Of special mention is the Stress Knowledge Map, which provides information of specific protein-protein interactions, protein-DNA interactions, small RNA-transcript interactions and enzymatic transformations of metabolites in the form of knowledge graphs [54]. These databases provide a useful resource for the community and are summarised in Table 1.

2.5. Large language model-assisted literature mining and data analysis

While genomic and transcriptomic approaches are invaluable to studying plant stress resilience mechanisms, these approaches are typically used in combination with other methods. To better understand how the different methods are combined in research papers, we performed a large language model (LLM)-based analysis of methods appearing in the papers that generated the stress-related genomic and transcriptomic data. The advent of AI tools provides a unique opportunity to rapidly mine the burgeoning scientific literature, which was previously performed through arduous manual curation. Recently, we deployed the text mining capacities of Generative Pre-trained Transformer (GPT) to process over 100,000 plant biology abstracts, revealing nearly 400,000 functional relationships between a wide array of biological entities - genes, metabolites, tissues, and others, with remarkable accuracy of over 85 % [55], and have further applied this tool across multiple species, such as yeast [56]. Here, we used GPT to perform a relational analysis of the experimental tools and methods commonly used in the study of plant stress in a bid to create a knowledge network for the training of plant biology specific models.

For our LLM-based analysis, we collated ~2000 articles associated with the total set of BioProject accessions discussed in the previous section for transcriptomic data, representing all plant stress studies recorded in the SRA (Fig. 4A). To this end, we queried the PubMed Central (PMC) and GEO databases where these BioProject accessions were referenced and obtained PubMed IDs for downloading the relevant research articles. Not all BioProject accessions have an associated PubMed ID, so in total, 2041 articles (1769 full-text, 272 abstract only) were downloaded from a starting list of 4323 unique stress BioProject IDs. Using the OpenAI API, we performed four LLM queries to help us parse and categorise the various research methods used in the plant stress literature (Fig. 4B). In our workflow, LLM1 was tasked to recognise and list the various experimental and bioinformatic methods performed in the 'Methods' section of our list of full-text articles. As LLMs are prone to hallucinating, we manually checked the accuracy of the LLM output for 50 random papers against their respective texts to ensure that no superfluous terms were listed. This quality control check showed that LLM1 had an accuracy value of 99.3 % at its task (Fig. 4C). LLM2 and LLM3 were designed to take the output lists and iteratively categorise them into groups and supergroups, respectively. Finally, LLM4 was tasked to group the original full-text derived lists into the supergroups obtained by LLM3. Within the OpenAI API, the "temperature" parameter dictates the variability of the output. Of the 4 LLMs, we adjusted the temperature of LLM1 and LLM4 to 0 to reduce hallucination but provided LLM2 and LLM3 (responsible for supergroup assignment) with a temperature of 0.5 to allow greater flexibility and creativity, as higher temperature values introduce more variation. In total, LLM1 identified 50,560 experimental approaches (e.g., alignment of protein sequences using MAFFT, Screening of differentially expressed genes using DESeq2), LLM2 combined every 500 approaches into 15–50 groups (e.g., microRNA (miRNA) Analysis, Read Quality Control - FASTX), and LLM3 further condensed these groups into 106 supergroups. Finally, LLM4 reclassified the 50,560 approaches into the 106 supergroups.

Finally, in order to plot a knowledge graph of the analyses used in the plant stress literature, we plotted counts of supergroups as nodes, where the size of the node denotes the number of articles each supergroup was found in, and the edge width represents the pointwise mutual information metric (Fig. 5C) between the two supergroups [57], indicating the strength of association based on how often they appear together in

Table 1
Summary of databases used in plant stress research. Each database is annotated with the type of data it contains, its area of focus, the type of species from which the data was obtained, and its associated reference(s).

Database	Data type	Focus area	Species type	References
Stress Knowledge Map	Knowledge graphs. Protein - protein interactions Protein - DNA interactions, smallRNA - transcript interactions Enzymatic transformations of metabolites	Heat, Drought, Flooding, Extra- and intracellular pathogens, Herbivory	*Arabidopsis thaliana*, *Solanum tuberosum*	[54]
Plant stress RNA-seq Nexus (PSRN)	Transcriptomic, long noncoding RNAs (lncRNAs)	ABA, Darkness, Cold, Pathogen, Ozone, Nutrition, Dehydration, Heat, Light, NaOH, PEG−8000	*Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Glycine max*, *Manihot esculenta*, *Oryza sativa* indica, *Oryza sativa* Japonica, *Panicum virgatum*, *Populus tremuloides*, *Solanum lycopersicum*, *Sorghum bicolor*, *Triticum aestivum*, *Vitis vinifera*	[43]
Plant Environmental stress transcript database	Transcriptomic	Heat, Cold, Dehydration, Salt	Wheat, Maize, Rice, Barley, Sorghum, Pearl millet, Rye, Arabidopsis, Common bean, Tomato, Soybean, Cowpea, Groundnut, Potato, Chickpea, Medicago	[44]
Plant Resistance Genes database (PRGdb 4.0)	Transcriptomic	Plant disease resistance	*Solanum lycopersicum*, *Oryza sativa*, *Triticum aestivum*, *Vitis vinifera* and *Arabidopsis thaliana*	[45]
DroughtDB	Transcriptomic	Drought stress response	Arabidopsis, rice, sorghum, maize, Brachypodium, tomato, barley, rye and *Aegilops tauschii*	[46]
Stress Responsive Transcription Factor Database (STIFDB V2.0)	Transcriptomic, Transcription Factor Binding Site	Bacteria, ABA, drought, cold, salinity, osmotic, wounding, dehydration, UV-B, high light, heat, heavy metals	*Arabidopsis thaliana* and *Oryza sativa* L.	[47]
CerealESTDb	Transcriptomic	ABA, cold, drought, heat and salt	Maize, Rice, Sorghum, Wheat	[48]
HEATSTER	Genomic	Heat stress transcription factors	65 species including 1 Phaeophyta, 3 Rhodophyta, and 61 species of Viridiplantae (5 Chlorophyta, 1 Bryophyta, 1 Lycopodiidae, 1 basal Magnoliophyta, 3 Gymnosperms, 15 Monocotyledons, and 35 Eudicotyledons)	[49]
Rice Stress-Resistant SNP database	Genomic SNPs specific to biotic and abiotic stress-resistant	Cold, Salt, Rice blast fungus, Heat, Bacterial leaf blight, Flood, Alkali, Bacterial stripe virus, Brown planthopper, White-backed planthopper, Bacterial sheath blight, Gall midge pest, Bacterial planthopper	Rice	[50]
PncStress	ncRNAs microRNAs, long non-coding RNAs, circular RNAs	48 biotic and 91 abiotic stresses	114 species	[51]
CropCircDB	circular RNAs	Drought, Cold, Salt	Maize and Rice	[52]
PlaASDB	Alternative splicing events	Salt, Heat, Hormone, Wounding, Drought, Osmotic, Dark, Chemical, Nutrient, Pathogen	Arabidopsis and Rice	[53]

articles (Fig. 5A). From the annotated heatmap in Fig. 5B, we observe patterns like how DEG detection is more strongly associated with other gene expression analyses, indicating that these analyses are more frequently performed together. Similarly, the close relationship between read quality control and sequence read mapping is expected as both are crucial steps in post-processing of sequencing data. Conversely, sequence read mapping and read quality control have significantly less association with more experimental methods such as phenotypic data analysis and RNA extraction. A caveat of the PMI metric is that it enriches for terms which have lower occurrence, resulting in more specialised methods being assigned a higher PMI score [58].

This information provides an overview of the various methods commonly and uncommonly used by researchers around the world. Indeed, while scientific methods and protocols are relatively well-described in the literature, the connections between them were previously only known intuitively and passed on from teacher to student by word of mouth. These connections have now been visualised in our graph, which will be valuable in improving plant biology specific models. Such endeavours have been rapidly gaining steam in diverse fields such as agriculture [59], finance [60], biomedical research [61] and materials chemistry [62], and have proven effective at rapidly and accurately distilling the scientific literature to its salient points.

3. The application of deep learning and large language models in biology

In general, ML methods used to analyse plant stress data rely on approaches that are able to learn from a priori ("supervised"), inductive ("unsupervised"), and reward-based ("reinforcement") experience [63]. "Supervised" algorithms are based on the training of models containing previously labelled data, include the likes of Multilayer Perceptron, Naive Bayes, Random Forest and Support Vector Machine models, while 'unsupervised' approaches such as k-means clustering, hierarchical clustering and Principal Component Analysis are based on raw or unlabelled data. (reviewed in [64]). Reinforcement learning (RL) is a branch of machine learning where an agent learns to make decisions by performing actions in an environment to maximise some notion of cumulative reward [65]. This paradigm is distinct from supervised and unsupervised learning, as it focuses on learning from interactions with the environment rather than from static datasets. These techniques have been applied to several use cases, such as gene function prediction [66], plant phenomics [67], and identification of specific stress responsive gene modules [68], with some algorithms more suited for certain tasks (reviewed in [69]).

Large Language Models (LLMs) are a newcomer to the application of AI in biology and are a hybrid amalgamation of these different strategies, which show great promise in being able to infer new insights from biological sequence data [70]. The initial phase of training LLMs, such as

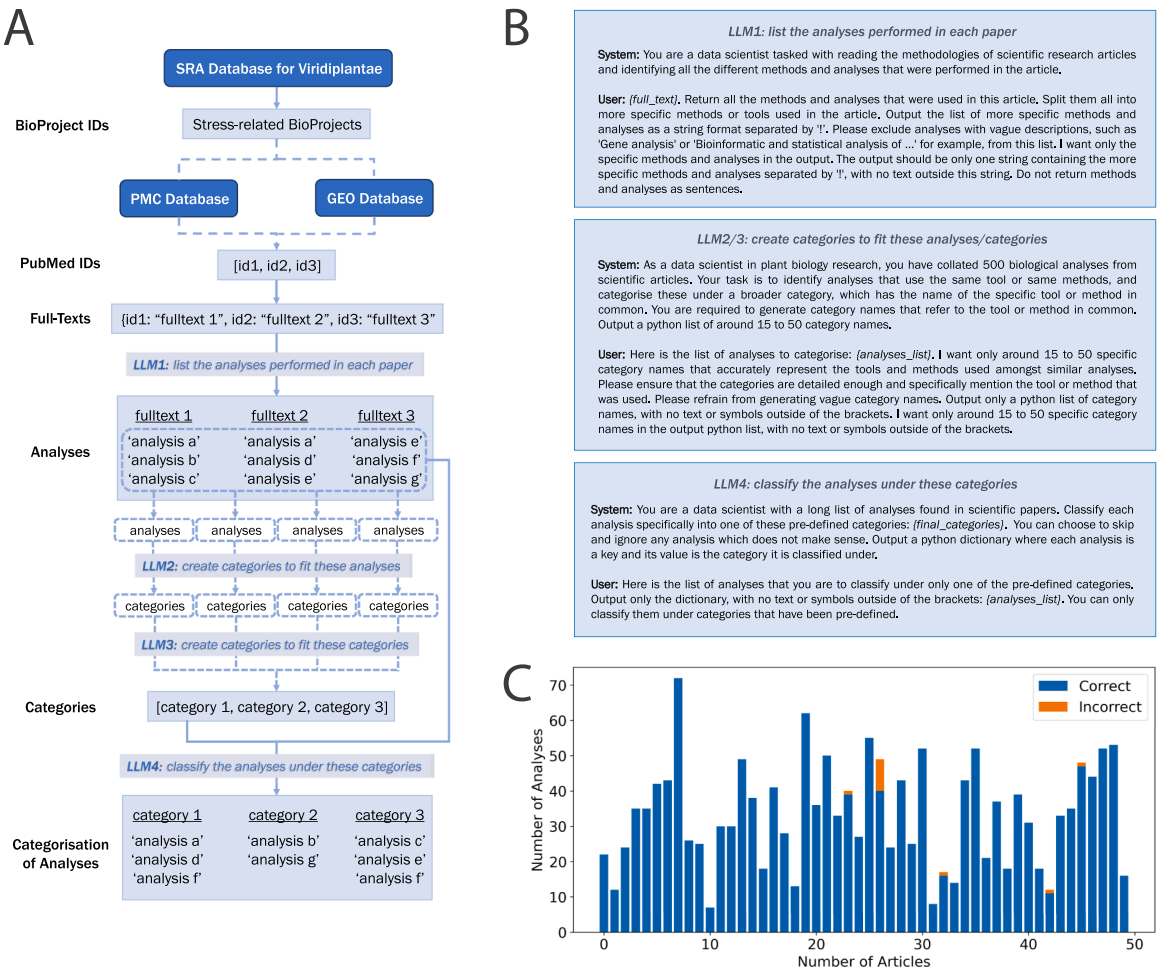

Fig. 4. Categorisation of analyses performed across transcriptomic studies. (A) Stress-related BioProject accessions with transcriptomic data were used to obtain their associated research articles. All the performed analyses that were stated in the articles were classified into broader supergroups using a series of prompts with the OpenAI API, as further detailed in (B). The accuracy of LLM1 was validated with 50 random articles by manually cross-checking whether each analysis generated was mentioned in the article. The percentage of correct analyses generated by LLM1 can be seen in (C), giving a mean accuracy value of 99.3 %.

GPT (Generative Pre-trained Transformer), is largely unsupervised. In this phase, the model learns from a large corpus of text without explicit labels. After the unsupervised pre-training, LLMs often undergo a supervised fine-tuning phase. In this phase, the model is further trained on a smaller, task-specific with explicit labels. LLMs can be viewed as engaging in self-supervised learning, which is a subset of unsupervised learning. In self-supervised learning, the model creates its own supervisory signals from the input data. For instance, in language modelling, the model uses the context surrounding a word as the supervisory signal to predict the word itself [71]. This approach allows the model to leverage the vast amounts of available text data without the need for manually labelled datasets. Some LLMs, including advanced versions like GPT-4, are also fine-tuned using reinforcement learning from human feedback (RLHF) techniques [72]. This involves using human feedback to improve the quality and accuracy of the generated responses. Typically, the 'classical' algorithms require significant data curation to limit noise, while the more advanced deep-learning models such as LLMs stand out from traditional ML algorithms due to their ability to process more complex and heterogeneous datasets as found in multi-omic or metadata (reviewed in [73]).

3.1. *Text mining for hidden connections from available data using LLMs*

The burgeoning volume of plant stress research can make it difficult to summarise how the different entities (e.g., genes, metabolites, organs, cell types, organisms) respond to stress and how the different entities are interrelated. While we have demonstrated here only the use of LLMs for the purpose of categorisation in a relatively small subset of the scientific literature (Fig. 5), the detection of patterns and trends across these diverse datasets brings out the real strength of using LLMs in such analyses. Previously, we had also demonstrated the text mining capacities of Generative Pre-trained Transformer (GPT) to identify functional relationships between various biological entities such as genes, metabolites and tissues in Arabidopsis [55] and yeast [56]. Similar approaches were also performed to extract protein-protein or protein-chemical associations from text [74]. The ability of LLMs to recognise categories of interest when defined in the prompts has also allowed for rapid selection of articles of interest pertaining to the effects of immunotherapy in glioma patients for downstream meta-analyses [61], and even to distil out experimental conditions used for metal-organic framework synthesis to build an AI chemistry assistant [62]. These approaches harness the inherent strengths of LLMs in their ability to understand natural language to perform parsing and categorisation tasks.

It was shown that ChatGPT has sufficient insight to the scientific corpus to generate a series of pertinent questions in the plant sciences, albeit relatively general in nature [75]. It was also noted that ChatGPT suffered from 'plant blindness' in which its answers to plant-specific questions posed by the user were scientifically inaccurate [76]. However, these issues could be addressed by coupling an LLM with a topic specific knowledge-base, that contains distilled information on a topic of

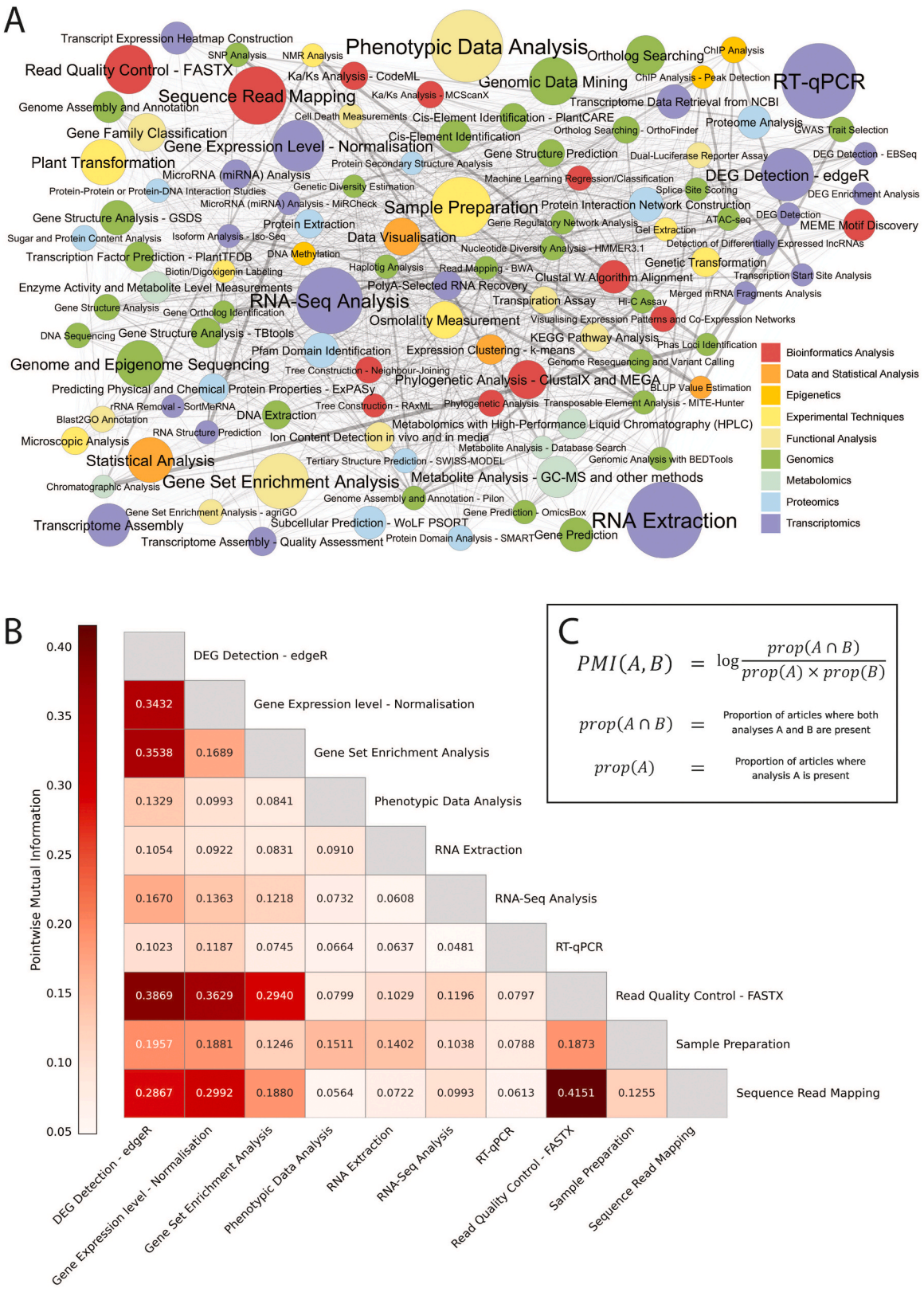

Fig. 5. LLM-assisted meta-analysis of methods present in the plant stress literature for transcriptomic studies. (A) All methods/analyses performed across all full-text articles were categorised under 106 supergroups. These supergroups were connected with one another based on the number of articles in which they appeared together. (B) The edge widths between the ten most abundant supergroups are displayed in an annotated heatmap. (C) Edge widths were calculated using the pointwise mutual information (PMI) metric, taking into account the proportion of articles in which each pair of supergroups co-occur in.

interest. For example, Retrieval Augmented-Generation (RAG) comprises a retrieval component that first identifies relevant information from the knowledge base based on the input query [77]. The generative component of the LLM then utilises this retrieved information to produce responses that are more accurate and contextually relevant. This approach is particularly beneficial in providing precise and informative answers, especially when the model's inherent knowledge might be outdated or insufficient. A recent development, GeneGPT, is a LLM-based approach which was trained to use the Web APIs of the National Center for Biotechnology Information (NCBI) for answering genomics questions [78]. While such approaches have not yet been applied to plant stress, an LLM that provides factual information about plant stress would be invaluable.

3.2. Predicting protein structure and function from sequence

The rise of self-supervised deep learning models for DNA and protein sequences, such as ESM-2 [79], DNABERT-2 [80] and MSA transformer [81], has made it possible to extract previously unknown insights on structure, function and evolution of proteins and DNA sequences (reviewed in [82]). In these deep learning models, labelled data is not required, and the model self-learns from raw sequences alone the fundamental properties of the input sequence without any further human guidance. ESM-2 itself is trained on the UniRef dataset, and is primarily trained by masking out certain regions of the protein sequence and training the model to "fill in the gaps" for said masked regions [79]. This allows the model to fundamentally understand the evolutionary "language" of proteins by inferring amino acid sequences from neighbours, and the model can then be finetuned into specific tasks [79]. Advances in this field have allowed the accurate gene ontology prediction from protein sequences alone, allowing for elucidation of the three main groupings in gene ontology: biomolecular processes, molecular functions, and cellular components. An example of such a tool is NetGO3, which uses the ESM model to provide state-of-the-art predictions into gene ontology [83]. The accurate prediction of GO terms can, in turn, provide leads that can guide further laboratory validation in a typical gene function prediction approach [6].

Furthermore, tools such as AlphaFold2 and conformational subsampling of such can provide structural insights into how proteins related to plant stress bind and interact with each other [84]. In essence, by sampling the multiple sequence alignment inputs into AlphaFold2, it is possible to find out what are the possible conformations of a certain protein and better understand the differing functions in different conformations [84]. The latest work has also enabled the ab initio prediction of the three-dimensional structure of ligands embedded in proteins, allowing for *de novo* design of binding molecules and cofactors [85]. The recent AlphaFold3 can computationally predict the folding and binding conformations of DNA, RNA and common biological ligands as well, which would greatly accelerate the elucidation of molecular interactions and their biological functions [86].

AI in biology is growing at a rapid pace, as exemplified by methods that can design proteins with desired properties by diffusion models [87] and predict protein-protein interactions [88]. These advances are made possible by the increased ease of generating high-throughput data, more powerful and affordable computer hardware and better AI/ML methods (reviewed in [82]). While the above examples provide general predictions pertaining to gene functions, we envision that they can be repurposed to provide insights into stress resilience-related processes.

3.3. Predicting stress resilience and responses from omics data

Predicting stress resilience that integrates genomic (genotype-to-phenotype (G2P)), transcriptomic, metabolomic and phenomic data could help us understand the molecular basis for stress resilience and engineer more resilient plants (reviewed in [64]). As the variation of complex traits is subject to complex regulatory circuits acting on the level of DNA, RNA, proteins and metabolites [73], genomic signatures that can underpin stress resilience are often difficult to pinpoint, highlighting a need to improve current G2P methods [89,90]. Indeed, integration of various types of omics data is known to improve the performance of models. For example, by integrating transcriptomics and metabolomics, yield and phenotypes of maize could be predicted with an increased accuracy [91]. Abiotic stress-tolerant crop phenotypes were similarly identified by integrating genomics, transcriptomics and proteomics [92]. A recent study showed that gene expression in combined stresses (e.g., heat and high light) could be predicted with high accuracy by regression approaches [93], showing that stress responses can be predicted even in more complex environments.

While integrating various types of data can improve the performance of predictive models, generating and analysing such data is still difficult and out of reach for most researchers. Fortunately, recent advances in computer science and AI have made it possible to perform new types of analyses that can predict gene expression from DNA sequences. Enformer (a portmanteau of enhancer and transformer) was built to predict gene expression and chromatin states in humans and mice from DNA sequences only [94]. While this approach has not been applied to plants yet, such an AI model could be used to predict the expression of stress resilience-conferring genes from DNA sequence, thus enabling new means to identify resilient crops. This goal could also be achieved with AI models that can predict the outcomes of genetic perturbations on gene expression. For example, the Geneformer model trained on DNA and the scGPT model trained on single-cell RNA-sequencing data were able to predict the consequences of genetic perturbations [95]. Such an analysis could be repurposed to plant stress responses, where a model trained on gene expression data from stress experiments could 'learn' to predict stress responses in new experimental settings.

Plant-specific models such as the Genomic Pre-trained Network (GPN) were trained using reference genomes from the Brassicales (including *A. thaliana*) based on a convolutional neural network and were able to identify DNA motifs, various types of genomic regions (intergenic, CDS, introns) and predict the effect of single-nucleotide polymorphisms [96]. FloraBERT is a transformer-based model that was trained on the promoter sequences of several hundred species of plants obtained from the Ensembl Plants, RefSeq, and MaizeGD databases and was used to predict and verify against the expression levels of a test set of *Zea mays* promoter sequences [97]. AgroNT is a foundational large language model trained on 48 plant reference genomes with a predominant focus on crop plants and was shown to achieve state-of-the-art results for the prediction of genomic elements and gene expression [98]. We envision that such approaches will be increasingly used to perform *in silico* mutational experiments to identify stress-resilient plants. While the above approaches have not yet been used to study stress responses, these AI models can potentially be used to identify emerging trends and repeated patterns within the data, providing a richer understanding of how different plant species react to environmental stresses. Such insights are crucial for evolving strategies to boost plant resistance to environmental challenges and for designing new studies in this domain.

3.4. High-throughput phenotyping

High-throughput phenotyping (HTP) in the field allows the breeders to rapidly investigate the potential yield of different cultivars and identification of desired traits such as biotic and abiotic stress resilience [67]. For example, HTP allowed the rapid measurement of plant maturity, seed size, and yield at early stages in 2551 genotypes of soybean (*Glycine max*) [99], and identification of desired traits in wheat and barley [100,101]. AI technologies, particularly ML and computer vision, have been pivotal in automating the process of phenotyping [102–105]. High-throughput phenotyping platforms that utilise drone and satellite imagery now employ AI to rapidly analyse plant traits such as growth patterns, stress responses, and disease resistance across large numbers of

plants [106]. ML's greatest success involves inferring trends from the collected data and generalising the results by training models that can e.g., predict consequences of non-ideal environmental conditions and crop yield [107]. However, traditional ML approaches require manual definition of relevant features (e.g., definition of specific spectral ranges), which is a significant effort that requires expertise in computation and image analysis [108]. Fortunately, deep learning (DL) incorporates benefits of both advanced computing power and massive datasets and allows for hierarchical data learning, which allows the models to infer the most relevant features from the data independently [109,110]. These approaches will be instrumental in monitoring plant responses and rapidly selecting plants with desired traits [111].

3.5. The democratisation of AI and its challenges

The application of artificial intelligence (AI) in biology holds immense potential, offering revolutionary ways to understand complex biological systems and accelerate research. AI democratises access to sophisticated analytical tools and computational power, enabling researchers from smaller institutions and developing countries to perform advanced analyses without needing extensive resources (reviewed in [112]). Many AI tools and models are available as open-source software, which encourages innovation and allows researchers to build on existing work, driving progress in the field. While AI tools are becoming more accessible, there is still a disparity in computational resources and expertise, which can hinder truly equal access and usage. The widespread availability of AI tools also raises ethical issues related to data privacy, misuse of AI, and the potential for unequal benefits across different socioeconomic groups (reviewed in [113–115]).

In biology, data can be noisy, incomplete, or inconsistent, which can impair the performance of AI models. The AI approaches often require copious amounts of data that need to be organised and machine-readable [116,117]. The inconsistency of proper annotation within databases like the Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO) pose significant challenges for researchers seeking to exploit the abundant wealth of data stored therein [118]. In many instances, essential information such as experimental conditions, plant species and stress types are either inadequately documented or entirely absent, which hampers the ability to categorise and compare studies due to incomplete contextual information. Thus, it is imperative for these databases to prioritise and enforce standardised annotation practices, with efforts such as establishment of the Minimum Information About a Microarray Experiment (MIAME) standards and enforcement of the Findability, Accessibility, Interoperability, and Reusability (FAIR) guiding principles by various stakeholders in the scientific community to improve the accessibility and interpretability of publicly available data [119,120]. As it has been shown that improved instructional detail significantly improves annotation quality [121], a possible implementation is for databases like SRA and GEO to provide clear case examples to decrease the uncertainty for authors when uploading and annotating their data. The quality of metadata affects both human and AI-driven research, and the onus is on the scientific community to enforce these rules and good practices.

Although tools have been developed to improve the searchability and visualisation of the inherent metadata [122,123], the underlying issue of missing or incorrect annotation is still not dealt with. We have shown previously that this lack of consistency in annotation prevents effective computational parsing of the sample metadata, leading to decreased accuracy and coverage of available samples for analysis [124]. Improved annotation would enhance the usability of these valuable resources and expedite scientific progress by enabling researchers to extract meaningful patterns and trends from the available data. Encouragingly, LLM-mediated parsing as performed here (Fig. 4C) and previously, showed significant improvement in automated sample annotation and categorisation [55,56]. It was capable enough to extract distinct experimental parameters pertaining to the synthesis of metal-organic frameworks from peer-reviewed research articles, and predict experimental outcomes [62]. Indeed, ML/AI models such as LLMs scale with model size, dataset size, and amount of compute used for training [125], so that as the amount of available data increases, the greater the improvement in model performance [126].

AI systems, particularly generative models, can produce erroneous or "hallucinated" results that do not reflect reality, which can mislead researchers if not carefully validated. These hallucinations may arise from several sources, ranging from inaccuracies in the original training dataset to biases arising during the pre-training of the LLM model (reviewed in [127]). The means used to address the issue of hallucinations depends on one's level of control over the LLM model. For instance, if one is only accessing publicly available LLMs such as ChatGPT, there are two main means by which one can decrease the frequency of hallucinations. The first is to decrease the 'Temperature' setting of the LLM (a value of zero means the replies are completely deterministic), which decreases the amount of randomness in its replies to queries [128]. The second is to specify clearly one's queries and prompts, the use of clear and concise language and provision of detail greatly diminishes LLM hallucination [129]. On the other hand, if one is creating their own LLM model, several parameters can be fine-tuned to decrease the frequency of LLM hallucination. In the pre-training phase, the number of training tokens, heterogeneity of training materials and the representation frequency of specific knowledge topics all play an important role in influencing model bias [130]. In the operational phase, Retrieval Augmented-Generation (RAG), is a promising strategy in which the LLM is provided with additional relevant documents retrieved from an external knowledge source, that helps mitigate some of the issues caused by hallucinations [77,131]. AI models may also suffer from "catastrophic forgetting", where they lose previously learned information when new data is introduced [132]. Keeping models updated with the latest data without losing previous knowledge is a challenge. Recent advances in addressing this issue have drawn inspiration from biology where physical modifications to excitatory synapses in the brain were shown to correlate with the acquisition of new knowledge or skills, and protect against forgetting [133]. Computer scientists have mimicked this concept by developing an algorithm termed elastic weight consolidation (EWC), which slows down learning on certain weights based on how important they are to previously seen tasks [134].

Many AI models, especially deep learning networks, function as "black boxes," providing little insight into how decisions are made. This lack of interpretability can be problematic in understanding biological mechanisms and ensuring trust in AI-driven conclusions [135]. Reproducing results generated by AI models can be challenging due to variability in data, model parameters, and computational environments (reviewed in [112], which complicates the validation and verification of scientific findings. This challenge is being tackled head on by the field of eXplainable Artificial Intelligence (XAI), which has made significant strides in developing methods or alternative models to provide greater interpretability and hence confidence in the model output for a large number of real-world applications (reviewed in [136]). 'White-box' models, which as the term suggests, allows for the interpretation of the logic and processes occurring between input and output. These models are typically designed using linear regression, decision tree, and rule-based models, which excel at performing single tasks, but fall behind current deep-learning models in the context of more general artificial intelligence. Most interpretability methods for explaining deep learning models apply to image classification and feature the concept of saliency maps, which highlights the relative 'importance' of the different image regions in producing the output (reviewed in [137]). Other approaches to understand the logic behind the relative weights and connections generated within the model involve the use of adversarial examples to invoke errors such as misclassifying an image by applying a single minute perturbation in the sample input [138].

4. Summary and outlook

Plant stress research is essential for us to develop crops that can thrive in more challenging environments. Biological systems are robust and reproducible, despite comprising millions of components interacting in ways that have evolved over billions of years of selection. The resulting systems are marvellously complex and are beyond current human comprehension [139]. To be able to model biological systems, we have to resort to simplistic rules that might result in digestible but incomplete narratives. However, the rapidly accumulating biological data and improving AI methods could be used to connect the different types of data and identify yet unknown and possibly unimaginable patterns present in biological systems.

The democratisation of AI-assisted tools heralds a revolution in biology and medicine. Previously the domain of specialised professions, LLM-based AI agents such as ChatGPT, Mistral, Claude, Gemini and others, have virtually taken over the world in diverse industries from agriculture [73] to zoology [140]. From decoding the secrets of the plant genome to understanding animal language, it seems that the development of AI is opening up completely new vistas in previously intractable fields, and the only limits are of our imagination and creativity in finding new problems to solve. The harnessing of this new technological wave promises significant acceleration of our efforts to understand more about plants and how they respond to diverse stresses across the plant kingdom. These developments have arrived at a critical juncture, where humanity is faced with the looming threat of climate change, and a rapidly growing population.

Addressing the impending challenges of climate change and global food security demands a concerted and collaborative effort from the global scientific community. Scientists play a crucial role in developing innovative, sustainable and resilient solutions to mitigate the impacts of climate change on agriculture and enhance global food security. The widening gap between richer and poorer countries in their ability to address climate change and food security requires a collective commitment to global equity. International collaborations, technology transfer and fair distribution of resources can contribute to narrowing this gap, ensuring that vulnerable populations have the means to adapt to changing climates and secure their food supply. As scientists contribute their expertise and advocate for inclusive policies, they play a pivotal role in fostering a more sustainable and equitable future for global food security.

CRediT authorship contribution statement

Marek Mutwil: Writing – review & editing, Supervision, Conceptualization. **Rohan Sunil:** Data curation. **Hilbert Lam:** Data curation. **Eugene Koh:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

MM thanks the Singaporean Ministry of Education for funding MOE-T2EP30123-0009 ‘Predicting the stress resilience mechanisms in the plant kingdom’. EK and RSS thank the Singaporean Ministry of Education for funding MOE-MOET32022-0002 ‘From Tough Pollen to Soft Matter’.

References

[1] Verslues PE, Bailey-Serres J, Brodersen C, Buckley TN, Conti L, Christmann A, et al. Burning questions for a warming and changing world: 15 unknowns in plant abiotic stress. Plant Cell 2023;35:67–108. https://doi.org/10.1093/plcell/koac263.

[2] Intergovernmental Panel on Climate Change (IPCC), editor. Technical Summary. Clim. Change 2022 – Impacts Adapt. Vulnerability Work. Group II Contrib. Sixth Assess. Rep. Intergov. Panel Clim. Change, Cambridge: Cambridge University Press; 2023, p. 37–118. https://doi.org/10.1017/9781009325844.002.

[3] Saijo Y, Loo EP. Plant immunity in signal integration between biotic and abiotic stress responses. N Phytol 2020;225:87–104.

[4] Zhang H, Zhu J, Gong Z, Zhu JK. Abiotic stress responses in plants. Nat Rev Genet 2022;23:104–19. https://doi.org/10.1038/s41576-021-00413-0.

[5] Zhu J.K. Abiotic Stress Signaling and Responses in Plants. vol. 167. 2016. https://doi.org/⟨10.1016/j.cell.2016.08.029⟩.

[6] Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. Trends Plant Sci 2014;19:212–21. https://doi.org/10.1016/j.tplants.2013.10.006.

[7] Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207–10. https://doi.org/10.1093/nar/30.1.207.

[8] Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res 2011;39:D19–21. https://doi.org/10.1093/nar/gkq1019.

[9] Julca I, Tan QW, Mutwil M. Toward kingdom-wide analyses of gene expression. Trends Plant Sci 2023;28:235–49. https://doi.org/10.1016/j.tplants.2022.09.007.

[10] Clough E, Barrett T. The gene expression omnibus database. Stat Genom Methods Protoc 2016:93–110.

[11] Karsch-Mizrachi I, Nakamura Y, Cochrane G. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res 2012;40:D33–7. https://doi.org/10.1093/nar/gkr1006.

[12] Parkinson H, Sarkans U, Kolesnikov N, Abeygunawardena N, Burdett T, Dylag M, et al. ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. Nucleic Acids Res 2011;39:D1002–4. https://doi.org/10.1093/nar/gkq1040.

[13] Kodama Y, Mashima J, Kaminuma E, Gojobori T, Ogasawara O, Takagi T, et al. The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. Nucleic Acids Res 2012;40:D38–42. https://doi.org/10.1093/nar/gkr994.

[14] Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res 2012;40:D57–63. https://doi.org/10.1093/nar/gkr1163.

[15] Auge G, Sunil RS, Ingle RA, Rahul PV, Mutwil M, Estevez JM. Current challenges for plant biology research in the Global South. New Phytologist 2024.

[16] Smith GF, Figueiredo E, Van Wyk AE. Kalanchoe (Crassulaceae) in Southern Africa: classification, biology, and cultivation. Academic Press; 2019.

[17] Griffin J, Jung G. Yield and forage quality of Panicum virgatum. CRC Press; 2019. p. 491–4.

[18] Wang E, Cruse RM, Sharma-Acharya B, Herzmann DE, Gelder BK, James DE, et al. Strategic switchgrass (Panicum virgatum) production within row cropping systems: Regional-scale assessment of soil erosion loss and water runoff impacts. GCB Bioenergy 2020;12:955–67. https://doi.org/10.1111/gcbb.12749.

[19] Barnes PW, Robson TM, Zepp RG, Bornman JF, Jansen MAK, Ossola R, et al. Interactive effects of changes in UV radiation and climate on terrestrial ecosystems, biogeochemical cycles, and feedbacks to the climate system. Photochem Photobio Sci 2023;22:1049–91. https://doi.org/10.1007/s43630-023-00376-7.

[20] Quiroz D, Lensink M, Kliebenstein DJ, Monroe JG. Causes of mutation rate variability in plant genomes. Annu Rev Plant Biol 2023;74:751–75. https://doi.org/10.1146/annurev-arplant-070522-054109.

[21] Levin G, Schuster G. Light tolerance in light-tolerant photosynthetic organisms: a knowledge gap. J Exp Bot 2024:erae338. https://doi.org/10.1093/jxb/erae338.

[22] Cantó-Pastor A, Mason GA, Brady SM, Provart NJ. Arabidopsis bioinformatics: tools and strategies. Plant J Cell Mol Biol 2021;108:1585–96. https://doi.org/10.1111/tpj.15547.

[23] Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 2016;166:481–91.

[24] Reiser L, Subramaniam S, Li D, Huala E. Using the Arabidopsis Information Resource (TAIR) to Find Information About Arabidopsis Genes. Curr Protoc Bioinforma 2017;60. https://doi.org/10.1002/cpbi.36.

[25] Kersey PJ, Allen JE, Allot A. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 2018;46. https://doi.org/10.1093/nar/gkx1011.

[26] Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van De Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res 2018;46:D1190–6. https://doi.org/10.1093/nar/gkx1002.

[27] Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res 2010;38. https://doi.org/10.1093/nar/gkp1019.

[28] Nelson ADL, Haug-Baltzell AK, Davey S, Gregory BD, Lyons E. EPIC-CoGe: managing and analyzing genomic data. Bioinforma Oxf Engl 2018;34:2651–3. https://doi.org/10.1093/bioinformatics/bty106.

[29] Kawakatsu T, Huang SSC, Jupe F. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell 2016;166:492–505. https://doi.org/10.1016/j.cell.2016.06.044.

[30] Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An “electronic fluorescent pictograph” Browser for exploring and analyzing large-scale

biological data sets. PLoS ONE 2007;2. https://doi.org/10.1371/journal.pone.0000718.
[31] Zhang H, Zhang F, Yu Y. A Comprehensive Online Database for Exploring ~20,000 Public Arabidopsis RNA-Seq Libraries. Mol Plant 2020;13:1231–3. https://doi.org/10.1016/j.molp.2020.08.001.
[32] Ma X, Denyer T, Timmermans MC. PscB: a browser to explore plant single cell RNA-sequencing data sets. Plant Physiol 2020;183:464–7.
[33] Proost S, Mutwil M. CoNekT: An open-source framework for comparative genomic and transcriptomic network analyses. Nucleic Acids Res 2018;46: W133–40. https://doi.org/10.1093/nar/gky336.
[34] Koh E, Goh W, Julca I, Villanueva E, Mutwil M. PEO: plant expression omnibus – a comparative transcriptomic database for 103 Archaeplastida. Plant J 2024;117: 1592–603. https://doi.org/10.1111/tpj.16566.
[35] Obayashi T, Aoki Y, Tadaka S, Kagaya Y, Kinoshita K. ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. Plant Cell Physiol 2018;59. https://doi.org/10.1093/pcp/pcx191.
[36] Austin RS, Hiu S, Waese J. New BAR tools for mining expression data and exploring Cis-elements in Arabidopsis thaliana. Plant J Cell Mol Biol 2016;88: 490–504. https://doi.org/10.1111/tpj.13261.
[37] Hooper CM, Castleden IR, Tanz SK, Aryamanesh N, Millar AH. SUBA4: the interactive data analysis centre for Arabidopsis subcellular protein locations. Nucleic Acids Res 2017;45:D1064–74. https://doi.org/10.1093/nar/gkw1041.
[38] Willems P, Horne A, Parys T. The Plant PTM Viewer, a central resource for exploring plant protein modifications. Plant J Cell Mol Biol 2019;99:752–62. https://doi.org/10.1111/tpj.14345.
[39] Yao Q, Ge H, Wu S. P3DB 3.0: from plant phosphorylation sites to protein networks. Nucleic Acids Res 2014;42. https://doi.org/10.1093/nar/gkt1135.
[40] Dong S, Lau V, Song R. Proteome-wide, structure-based prediction of protein-protein interactions/new molecular interactions viewer. Plant Physiol 2019;179: 1893–907. https://doi.org/10.1104/pp.18.01216.
[41] Li P, Zang W, Li Y, Xu F, Wang J, Shi T. AtPID: the overall hierarchical functional protein interaction network interface and analytic platform for arabidopsis. Nucleic Acids Res 2011;39:D1130–3. https://doi.org/10.1093/nar/gkq959.
[42] Jumper J, Evans R, Pritzel A. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583–9. https://doi.org/10.1038/s41586-021-03819-2.
[43] Li J-R, Liu C-C, Sun C-H, Chen Y-T. Plant stress RNA-seq Nexus: a stress-specific transcriptome database in plant cells. BMC Genom 2018;19:966. https://doi.org/10.1186/s12864-018-5367-5.
[44] Balaji J, Crouch JH, Petite PV, Hoisington DA. A database of annotated tentative orthologs from crop abiotic stress transcripts. Bioinformation 2006;1:225–7.
[45] Calle García J, Guadagno A, Paytuvi-Gallart A, Saera-Vila A, Amoroso CG, D'Esposito D, et al. PRGdb 4.0: an updated database dedicated to genes involved in plant disease resistance process. Nucleic Acids Res 2022;50:D1483–90. https://doi.org/10.1093/nar/gkab1087.
[46] Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön C-C, et al. DroughtDB: an expert-curated compilation of plant drought stress genes and their homologs in nine species. Database 2015;2015:bav046. https://doi.org/10.1093/database/bav046.
[47] Naika M, Shameer K, Mathew OK, Gowda R, Sowdhamini R. STIFDB2: an updated version of plant stress-responsive transcription factor database with additional stress signals, stress-responsive transcription factor binding sites and stress-responsive genes in arabidopsis and rice. Plant Cell Physiol 2013;54:e8. https://doi.org/10.1093/pcp/pcs185.
[48] Kumar S, Bhati J, Saha A, Lal SB, Pandey PK, Mishra DC, et al. CerealESTDb: a comprehensive resource for abiotic stress-responsive annotated ests with predicted genes, gene ontology, and metabolic pathways in major cereal crops. Front Genet 2022;13. https://doi.org/10.3389/fgene.2022.842868.
[49] Berz J, Simm S, Schuster S, Scharf K-D, Schleiff E, Ebersberger I. HEATSTER: a database and web server for identification and classification of heat stress transcription factors in plants. Bioinforma Biol Insights 2019;13: 1177932218821365. https://doi.org/10.1177/1177932218821365.
[50] Tareke Woldegiorgis S, Wang S, He Y, Xu Z, Chen L, Tao H, et al. Rice stress-resistant SNP database. Rice 2019;12:97. https://doi.org/10.1186/s12284-019-0356-0.
[51] Wu W, Wu Y, Hu D, Zhou Y, Hu Y, Chen Y, et al. PncStress: a manually curated database of experimentally validated stress-responsive non-coding RNAs in plants. Database 2020;2020:baaa001. https://doi.org/10.1093/database/baaa001.
[52] Wang K, Wang C, Guo B, Song K, Shi C, Jiang X, et al. CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress. Database 2019;2019:baz053. https://doi.org/10.1093/database/baz053.
[53] Guo X, Wang T, Jiang L, Qi H, Zhang Z. PlaASDB: a comprehensive database of plant alternative splicing events in response to stress. BMC Plant Biol 2023;23: 225. https://doi.org/10.1186/s12870-023-04234-7.
[54] Bleker C, Ramšak Ž, Bittner A, Podpečan V, Zagorščak M, Wurzinger B, et al. Stress knowledge map: a knowledge graph resource for systems biology analysis of plant stress responses. Plant Commun 2024;5. https://doi.org/10.1016/j.xplc.2024.100920.
[55] Fo K., Chuah Y.S., Foo H., Davey E.E., Fullwood M., Thibault G., et al. PlantConnectome: knowledge networks encompassing >100,000 plant article abstracts 2023:2023.07.11.548541. https://doi.org/⟨10.1101/2023.07.11.548541⟩.
[56] Arulprakasam K.R., Toh J.W.S., Foo H., Kumar M.R., Kutevska A.-N., Davey E.E., et al. Harnessing full-text publications for deep insights into C. elegans and Drosophila connectomes 2024:2024.04.13.588993. https://doi.org/⟨10.1101/2024.04.13.588993⟩.
[57] Church KW, Hanks P. Word association norms, mutual information, and lexicography. Comput Linguist 1990;16:22–9.
[58] Aka O, Burke K, Bauerle A, Greer C, Mitchell M. Measuring Model Biases in the Absence of Ground Truth. Proc. 2021 AAAIACM Conf. AI Ethics Soc. New York, NY, USA: Association for Computing Machinery; 2021. p. 327–35. https://doi.org/10.1145/3461702.3462557.
[59] Potamitis I. ChatGPT in the context of precision agriculture data analytics. ArXiv Prepr ArXiv231106390 2023.
[60] Araci D. Finbert: Financial sentiment analysis with pre-trained language models. ArXiv Prepr ArXiv190810063 2019.
[61] Cai X, Geng Y, Du Y, Westerman B, Wang D, Ma C, et al. Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation. medRxiv 2023. 2023–09.
[62] Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. J Am Chem Soc 2023;145: 18048–62.
[63] Mitchell T, Buchanan B, DeJong G, Dietterich T, Rosenbloom P, Waibel A. Machine learning. Annu Rev Comput Sci 1990;4:417–33.
[64] Murmu S, Sinha D, Chaurasia H, Sharma S, Das R, Jha GK, et al. A review of artificial intelligence-assisted omics techniques in plant defense: current trends and future directions. Front Plant Sci 2024;15:1292054.
[65] Neftci EO, Averbeck BB. Reinforcement learning in artificial and biological systems. Nat Mach Intell 2019;1:133–43.
[66] Kushwaha SK, Chauhan P, Hedlund K, Ahrén D. NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. Bioinformatics 2016;32:1223–5. https://doi.org/10.1093/bioinformatics/btv714.
[67] Hunt CH, Hayes BJ, van Eeuwijk FA, Mace ES, Jordan DR. Multi-environment analysis of sorghum breeding trials using additive and dominance genomic relationships. Theor Appl Genet 2020;133:1009–18. https://doi.org/10.1007/s00122-019-03526-7.
[68] Ma C, Xin M, Feldmann KA, Wang X. Machine learning–based differential network analysis: a study of stress-responsive transcriptomes in arabidopsis. Plant Cell 2014;26:520–37. https://doi.org/10.1105/tpc.113.121913.
[69] Yan J, Wang X. Machine learning bridges omics sciences and plant breeding. Trends Plant Sci 2023;28:199–210. https://doi.org/10.1016/j.tplants.2022.08.018.
[70] Devlin J., Chang M.-W., Lee K., Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Prepr ArXiv181004805 2018.
[71] Fedus W., Goodfellow I., Dai A.M. Maskgan: better text generation via filling in the_. ArXiv Prepr ArXiv180107736 2018.
[72] Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F.L., et al. Gpt-4 technical report. ArXiv Prepr ArXiv230308774 2023.
[73] Harfouche AL, Jacobson DA, Kainer D, Romero JC, Harfouche AH, Mugnozza GS, et al. Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. Trends Biotechnol 2019;37:1217–35. https://doi.org/10.1016/j.tibtech.2019.05.007.
[74] Weber L, Barth F, Lorenz L, Konrath F, Huska K, Wolf J, et al. PEDL+: protein-centered relation extraction from PubMed at your fingertip. Bioinformatics 2023; 39:btad603.
[75] Agathokleous E, Rillig MC, Peñuelas J, Yu Z. One hundred important questions facing plant science derived using a large language model. Trends Plant Sci 2023.
[76] Geitmann A, Bidhendi AJ. Plant blindness and diversity in AI language models. Trends Plant Sci 2023.
[77] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv Neural Inf Process Syst 2020;33:9459–74.
[78] Jin Q., Yang Y., Chen Q., Lu Z. GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information. ArXiv 2023: arXiv:2304.09667v3.
[79] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379: 1123–30. https://doi.org/10.1126/science.ade2574.
[80] Zhou Z., Ji Y., Li W., Dutta P., Davuluri R., Liu H. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome 2024. https://doi.org/⟨10.48550/arXiv.2306.15006⟩.
[81] Rao RM, Liu J, Verkuil R, Meier J, Canny J, Abbeel P, et al. MSA Transformer. In: Meila M, Zhang T, editors. Proc. 38th Int. Conf. Mach. Learn., vol. 139. PMLR; 2021. p. 8844–56.
[82] Lam HYI, Ong XE, Mutwil M. Large language models in plant biology. Trends Plant Sci 2024. https://doi.org/10.1016/j.tplants.2024.04.013.
[83] Wang S, You R, Liu Y, Xiong Y, Zhu S. NetGO 3.0: protein language model improves large-scale functional annotations. Genom Proteom Bioinforma 2023; 21:349–58. https://doi.org/10.1016/j.gpb.2023.04.001.
[84] del Alamo D, Sala D, Mchaourab HS, Meiler J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. eLife 2022; 11:e75751. https://doi.org/10.7554/eLife.75751.
[85] Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science 2024; 384:eadl2528. https://doi.org/10.1126/science.adl2528.
[86] Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024: 1–3.

[87] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature 2023;620:1089–100. https://doi.org/10.1038/s41586-023-06415-8.

[88] Bryant P, Pozzati G, Zhu W, Shenoy A, Kundrotas P, Elofsson A. Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nat Commun 2022;13:6028. https://doi.org/10.1038/s41467-022-33729-4.

[89] Cembrowska-Lech D, Krzemińska A, Miller T, Nowakowska A, Adamski C, Radaczyńska M, et al. An integrated multi-omics and artificial intelligence framework for advance plant phenotyping in horticulture. Biology 2023;12:1298. https://doi.org/10.3390/biology12101298.

[90] Rico-Chávez AK, Franco JA, Fernandez-Jaramillo AA, Contreras-Medina LM, Guevara-González RG, Hernandez-Escobedo Q. Machine learning for plant stress modeling: a perspective towards hormesis management. Plants 2022;11:970. https://doi.org/10.3390/plants11070970.

[91] Cruz DF, De Meyer S, Ampe J, Sprenger H, Herman D, Van Hautegem T, et al. Using single-plant-omics in the field to link maize genes to functions and phenotypes. Mol Syst Biol 2020;16:e9667. https://doi.org/10.15252/msb.20209667.

[92] Jogaiah S, Govind SR, Tran L-SP. Systems biology-based approaches toward understanding drought tolerance in food crops. Crit Rev Biotechnol 2013;33:23–39. https://doi.org/10.3109/07388551.2012.659174.

[93] Tan QW, Lim PK, Chen Z, Pasha A, Provart N, Arend M, et al. Cross-stress gene expression atlas of Marchantia polymorpha reveals the hierarchy and regulatory principles of abiotic stress responses. Nat Commun 2023;14:986. https://doi.org/10.1038/s41467-023-36517-w.

[94] Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 2021;18:1196–203. https://doi.org/10.1038/s41592-021-01252-x.

[95] Theodoris CV, Xiao L, Chopra A, Chaffin MD, Al Sayed ZR, Hill MC, et al. Transfer learning enables predictions in network biology. Nature 2023;618:616–24. https://doi.org/10.1038/s41586-023-06139-9.

[96] Benegas G, Batra SS, Song YS. DNA language models are powerful predictors of genome-wide variant effects. Proc Natl Acad Sci USA 2023;120:e2311219120. https://doi.org/10.1073/pnas.2311219120.

[97] Levy B., Xu Z., Zhao L., Kremling K., Altman R., Wong P., et al. FloraBERT: cross-species transfer learning withattention-based neural networks for geneexpression predictio 2022. ⟨https://doi.org/10.21203/rs.3.rs-1927200/v1⟩.

[98] Mendoza-Revilla J., Trop E., Gonzalez L., Roller M., Dalla-Torre H., Almeida B.P. de, et al. A Foundational Large Language Model for Edible Plant Genomes 2023: 2023.10.24.563624. ⟨https://doi.org/10.1101/2023.10.24.563624⟩.

[99] Yuan W, Wijewardane NK, Jenkins S, Bai G, Ge Y, Graef GL. Early prediction of soybean traits through color and texture features of canopy RGB imagery. Sci Rep 2019;9. https://doi.org/10.1038/s41598-019-50480-x.

[100] Barker J, Zhang N, Sharon J. Development of a field-based high-throughput mobile phenotyping platform. Comput Electron Agric 2016;122:74–85. https://doi.org/10.1016/j.compag.2016.01.017.

[101] Escalante HJ, Rodríguez-Sánchez S, Jiménez-Lizárraga M, Morales-Reyes A, Calleja J, Vazquez R. Barley yield and fertilization analysis from UAV imagery: a deep learning approach. Int J Remote Sens 2019;40:2493–516. https://doi.org/10.1080/01431161.2019.1577571.

[102] Ashourloo D, Mobasheri MR, Huete A. Evaluating the effect of different wheat rust disease symptoms on vegetation indices using hyperspectral measurements. Remote Sens 2014;6:5107–23. https://doi.org/10.3390/rs6065107.

[103] Lin K, Gong L, Huang Y, Liu C, Pan J. Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network. Front Plant Sci 2019;10. https://doi.org/10.3389/fpls.2019.00155.

[104] Ramcharan A, McCloskey P, Baranowski K. A mobile-based deep learning model for cassava disease diagnosis. Front Plant Sci 2019;10. https://doi.org/10.3389/fpls.2019.00272.

[105] Sandhu KS, Mihalyov PD, Lewien MJ, Pumphrey MO, Carter AH. Combining genomic and phenomic information for predicting grain protein content and grain yield in spring wheat. Front Plant Sci 2021;12. https://doi.org/10.3389/fpls.2021.613300.

[106] Gill T, Gill SK, Saini DK, Chopra Y, Koff JP, Sandhu KS. A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. Phenomics 2022;2:156–83. https://doi.org/10.1007/s43657-022-00048-z.

[107] van Klompenburg T, Kassahun A, Catal C. Crop yield prediction using machine learning: a systematic literature review. Comput Electron Agric 2020;177:105709. https://doi.org/10.1016/j.compag.2020.105709.

[108] Yang W, Feng H, Zhang X, Zhang J, Doonan JH, Batchelor WD, et al. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol Plant 2020;13:187–214. https://doi.org/10.1016/j.molp.2020.01.008.

[109] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[110] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69. https://doi.org/10.1093/bib/bbw068.

[111] Yağ İ, Altan A. Artificial intelligence-based robust hybrid algorithm design and implementation for real-time detection of plant diseases in agricultural environments. Biology 2022;11. https://doi.org/10.3390/biology11121732.

[112] Williamson HF, Brettschneider J, Caccamo M, Davey RP, Goble C, Kersey PJ, et al. Data management challenges for artificial intelligence in plant and agricultural research. F1000Research 2023;10:324. https://doi.org/10.12688/f1000research.52204.2.

[113] Harfouche AL, Petousi V, Jung W. AI ethics on the road to responsible AI plant science and societal welfare. Trends Plant Sci 2024;29:104–7. https://doi.org/10.1016/j.tplants.2023.12.016.

[114] Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell 2019;1:389–99. https://doi.org/10.1038/s42256-019-0088-2.

[115] Ryan M. The social and ethical impacts of artificial intelligence in agriculture: mapping the agricultural AI literature. AI Soc 2023;38:2473–85. https://doi.org/10.1007/s00146-021-01377-9.

[116] Northcutt C.G., Athalye A., Mueller J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks 2021.

[117] Budach L., Feuerpfeil M., Ihde N., Nathansen A., Noack N., Patzlaff H., et al. The Effects of Data Quality on Machine Learning Performance 2022. ⟨https://doi.org/10.48550/arXiv.2207.14529⟩.

[118] Gonçalves RS, Musen MA. The variable quality of metadata about biological samples used in biomedical experiments. Sci Data 2019;6:190021. https://doi.org/10.1038/sdata.2019.21.

[119] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet 2001;29:365–71. https://doi.org/10.1038/ng1201-365.

[120] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[121] Rädsch T, Reinke A, Weru V, Tizabi MD, Schreck N, Kavur AE, et al. Labelling instructions matter in biomedical image analysis. Nat Mach Intell 2023;5:273–83. https://doi.org/10.1038/s42256-023-00625-5.

[122] Vaneechoutte D, Vandepoele K. Curse: building expression atlases and co-expression networks from public RNA-Seq data. Bioinformatics 2019;35:2880–1. https://doi.org/10.1093/bioinformatics/bty1052.

[123] Zhu Y, Stephens RM, Meltzer PS, Davis SR. SRAdb: query and use public next-generation sequencing data from within R. BMC Bioinforma 2013;14:19. https://doi.org/10.1186/1471-2105-14-19.

[124] Goh W, Mutwil M. LSTrAP-Kingdom: an automated pipeline to generate annotated gene expression atlases for kingdoms of life. Bioinformatics 2021;37:3053–5. https://doi.org/10.1093/bioinformatics/btab168.

[125] Kaplan J., McCandlish S., Henighan T., Brown T.B., Chess B., Child R., et al. Scaling Laws for Neural Language Models 2020. ⟨https://doi.org/10.48550/arXiv.2001.08361⟩.

[126] Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 2024;21:1470–80. https://doi.org/10.1038/s41592-024-02201-0.

[127] Huang L., Yu W., Ma W., Zhong W., Feng Z., Wang H., et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions 2023. ⟨https://doi.org/10.48550/arXiv.2311.05232⟩.

[128] Renze M., Guven E. The Effect of Sampling Temperature on Problem Solving in Large Language Models 2024. ⟨https://doi.org/10.48550/arXiv.2402.05201⟩.

[129] Santu S.K.K., Feng D. TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks 2023. ⟨https://doi.org/10.48550/arXiv.2305.11430⟩.

[130] Li J., Chen J., Ren R., Cheng X., Zhao W.X., Nie J.-Y., et al. The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models 2024. ⟨https://doi.org/10.48550/arXiv.2401.03205⟩.

[131] Guu K, Lee K, Tung Z, Pasupat P, Chang M. Retrieval augmented language model pre-training. PMLR,; 2020. p. 3929–38.

[132] French RM. Catastrophic forgetting in connectionist networks. Trends Cogn Sci 1999;3:128–35.

[133] Yang G, Pan F, Gan W-B. Stably maintained dendritic spines are associated with lifelong memories. Nature 2009;462:920–4.

[134] Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci 2017;114:3521–6. https://doi.org/10.1073/pnas.1611835114.

[135] Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, et al. Explainable artificial intelligence (XAI): what we know and what is left to attain trustworthy artificial intelligence. Inf Fusion 2023;99:101805. https://doi.org/10.1016/j.inffus.2023.101805.

[136] Saranya A, Subhashini R. A systematic review of Explainable Artificial Intelligence models and applications: recent developments and future trends. Decis Anal J 2023;7:100230. https://doi.org/10.1016/j.dajour.2023.100230.

[137] Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy 2021;23. https://doi.org/10.3390/e23010018.

[138] Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., et al. Intriguing properties of neural networks 2014. ⟨https://doi.org/10.48550/arXiv.1312.6199⟩.

[139] Batzoglou S. Large language models in molecular biology. Medium 2023. ⟨https://towardsdatascience.com/large-language-models-in-molecular-biology-9eb6b65d8a30⟩ (accessed April 22, 2024).

[140] Love P., Arenas I de la T., Learner S., London S.J. in. How AI is decoding the animal kingdom 2024. ⟨https://ig.ft.com/ai-animals/⟩ (accessed April 22, 2024).