Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d6np00011h

Emerging technologies for the discovery of biosynthetic genes in plants

Anne Jaczkowski, †[ab] Arne Bültemeier, †[ab] Benedikt Seligmann, [a] Boas Pucker [c] and Jakob Franke *[ab]

Received 30th January 2026

DOI: 10.1039/d6np00011h

rsc.li/npr

Covering 1982 to 2026

Some of the most prominent natural products originate from plants. Discovering their biosynthetic genes has been a slow process. In contrast to microbial systems, co-expression analysis rather than genome mining has been the main strategy to elucidate biosynthetic pathways in plants. However, traditional co-expression analyses are limited in efficiency and often not as successful as desired. In this review, we describe emerging technologies to improve or replace traditional co-expression analyses, for example based on genome or protein data. Furthermore, we critically discuss the current state and impact of artificial intelligence and machine learning in the field. Our review will help to select the most efficient approaches for elucidation of biosynthetic pathways in plants for future work. Additionally, we highlight areas that require further methodological improvements to guide future research.

1. Introduction

Some of the most prominent natural products used in medicine, agriculture, and other fields of industry originate from plants, for example morphine, azadirachtin, and menthol. For decades, there have been extensive efforts to understand the biosynthetic pathways by which plants produce such natural products. The motivation behind these studies has been twofold: first, to improve our fundamental understanding of biochemical pathways in plants. Second, and perhaps more importantly, as a starting point for biotechnology and metabolic engineering efforts to improve access to natural products from plants, as many of these occur only in minor quantities in plants and are therefore expensive.[1,2] In microorganisms, genome mining is the single major strategy to elucidate biosynthetic pathways of natural products.[3–5] This is in sharp contrast to plants, where genome mining has found very limited use so far. Originally, the major limitation preventing efficient application

[a]Institute of Botany, Leibniz University Hannover, Hannover 30419, Germany. E-mail: jakob.franke@botanik.uni-hannover.de

[b]Centre of Biomolecular Drug Research, Leibniz University Hannover, Hannover 30167, Germany

[c]Institute for Cellular and Molecular Botany, University of Bonn, Bonn 53115, Germany

† These authors contributed equally.

of genome mining in plants was simply the lack of genomic data of sufficient quality. However, even today, with improved sequencing technologies and hundreds of plant genome sequences having become available,[6–8] genome mining in plants is still limited in relevance, as many currently elucidated biosynthetic genes in plants are not physically clustered by pathway.[9] Instead, for the past two decades, the central paradigm for gene discovery in biosynthetic pathways in plants has been the so-called "guilt-by-association" principle:[10] It is assumed that genes in the same biosynthetic pathway share the same expression pattern across tissues or experimental conditions. At least one known pathway gene, termed "bait gene", is then required as a reference point to "fish" for further genes with a similar expression pattern. Indeed, searching for genes that are co-expressed with one or several bait genes from the target pathway has proven to be a very efficient strategy forming the foundation of many breakthrough studies in recent years. Thereby, co-expression analyses have largely superseded other established approaches for discovering biosynthetic genes such as classical genetics.[11] However, this traditional co-expression analysis is an inherently slow and not always reliable process, which can either lead to very long (often >100) candidate lists[12] or to overlooking genes which are not commonly associated

Anne Jaczkowski

Anne Jaczkowski is a PhD student at Leibniz University Hannover, working in Prof. Jakob Franke's group on the biochemistry of specialised metabolites in plants. After completing her B.Sc. in Molecular Biotechnology at Dresden University of Technology, she pursued an M.Sc. in Plant Biotechnology at Leibniz University Hannover, driven by her interest in plant biology. Her current interdisciplinary research combines biology, biochemistry, and bioinformatics, aiming to discover plant-derived enzymes for sustainable modification of alkaloids with medicinal relevance.

Arne Bültemeier

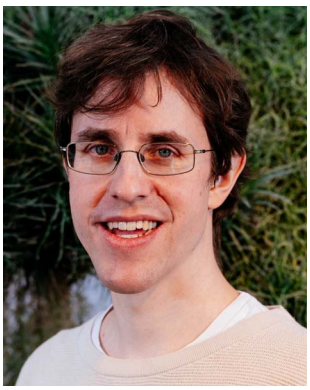
Arne Bültemeier is a PhD student in Jakob Franke's "Biochemistry of Plant Specialised Metabolites" research group. He earned his B. Sc. and M. Sc. diploma in Life Sciences from Leibniz University Hannover and joined the group during his master's studies to work on the elucidation of steroid biosynthetic pathways in Solanaceae *plants. The focus of his research lies in the identification and characterisation of genes involved in plant specialised metabolism, with a special emphasis on gene clustering.*

Benedikt Seligmann

Benedikt Seligmann is a PhD student at Leibniz University Hannover and former member of Prof. Jakob Franke's group. He obtained both his B. Sc. and M. Sc. in Biotechnology from the Technical University of Braunschweig, with a focus on cell biology and bioprocess engineering during his master's studies. His research interests shifted towards plant systems, while maintaining a strong emphasis on biotechnological approaches. His current work focuses on yeast metabolic engineering for the production of plant specialised metabolites, integrating methods from molecular biology, biochemistry, and biotechnology.

Boas Pucker

Boas Pucker is a plant scientist and bioinformatician investigating the evolution of plant specialised metabolism. He conducted research at Bielefeld University, University of Cambridge, and TU Braunschweig. His work integrates genomics, phylogenetics, and transcriptomics to uncover how biosynthetic pathways, particularly flavonoid metabolism, evolve across species. By combining large-scale sequencing data with public datasets, he advances big data up-cycling approaches to extract new biological insights. His Plant Biotechnology and Bioinformatics group was established at TU Braunschweig in 2021 and moved with him to the University of Bonn in 2025.



Jakob Franke

Jakob Franke is a biochemist and natural product chemist with a focus on specialised metabolites from plants. After receiving a B.Sc. in Biochemistry and a M.Sc. in Chemistry in Munich, he joined the group of Christian Hertweck in Jena for his doctoral studies on microbial natural product biosynthesis. Switching to plants, he then carried out postdoctoral research in the group of Sarah O'Connor at the John Innes Centre in Norwich, UK. In 2017, Jakob started his independent career at Leibniz University Hannover, Germany. His lab is focused on elucidating biosynthetic pathways in plants and harnessing this knowledge for metabolic engineering.

with specialised metabolism or not tightly correlated with the pathway of interest at the transcriptional level. To improve the efficiency of biosynthetic pathway elucidation in plants, there is therefore an urgent need to improve traditional co-expression analyses or to find alternative strategies that can complement or extend them based on modern methodological developments.

In this review, we will first describe current strategies that have been used to improve the power of co-expression analyses (Section 2), followed by Sections 3–5 focussing on orthogonal discovery strategies. In Section 3, we discuss in which ways and to what extent genome data can facilitate the discovery of biosynthetic genes in plants. Section 4 covers current approaches that are focused directly on biosynthetic proteins rather than at the gene level. In Section 5 we describe efforts to leverage artificial intelligence for the study of biochemical pathways in plants, including a critical discussion of current limitations. Finally, Section 6 provides a conclusion with a critical comparison of these different emerging technologies which facilitate the discovery of biosynthetic genes in plants.

2. Improving co-expression analyses

As co-expression analyses have been the dominant strategy to discover biosynthetic genes in plants, an obvious way forward is to further improve existing methods (Fig. 1). The statistical power of co-expression analyses and therefore its capacity to reveal a restricted set of gene candidates with high confidence is strongly determined by the number of samples and the variation in expression levels of target genes within these samples. Ideal datasets for co-expression analysis include some samples with very high expression of target genes and other samples with very low values. For example, several triterpenoid biosynthesis genes were discovered based on their strong root-specific expression.[13] Likewise, defence-related biosynthesis genes in bread wheat showed strong induction upon treatment with pathogens.[14] As a consequence, traditional co-expression analyses tend to fail in cases where biosynthetic pathways merely show constant basal expression throughout a plant without any obvious pattern, or if expression patterns are not uniform within a pathway. Several studies suggest that in many cases where initially no clear expression pattern can be observed, there are indeed hidden patterns which are averaged out in bulk transcriptome sequencing data.[15,16] A major strategy to improve co-expression analyses is therefore to refine datasets to increase the resolution of expression data. In this section, we will discuss different approaches to achieve such improved resolution, either based on integration of further omics data, increase of the spatial or temporal resolution, or by comparing expression patterns across species borders (Fig. 1).

2.1. Multi-omics strategies

On its own, transcriptomic analysis is typically limited to a one-sided perspective on complex, biological systems which neglects downstream regulatory layers. This can easily result in missing critical connections that would be of interest for discovering biosynthetic genes. To expand the capacity of co-expression analysis and to overcome its limitations, it has become increasingly popular to combine transcriptomics with other layers of omics data, such as genomics, proteomics and/or metabolomics (Fig. 1A). These multi-omics analyses offer a more comprehensive view on complex biological processes and are especially useful when a pathway is still completely unknown or when functional relationships are not clearly reflected in expression data. Several reviews have already highlighted the rise of multi-omics analysis from different perspectives, *e.g.*, illustrating applications with specific examples,[17] discussing challenges of handling high-dimensional biological datasets,[18] and providing potential guidelines for the design and execution of multi-omics studies.[19]

While (epi)genomics and proteomics will be discussed in more detail in Sections 3 and 4, we focus here on one of the most widely applied multi-omics strategies for pathway elucidation, the integration of transcriptomics and metabolomics. The reason for the prevalence of this combination is probably that metabolite analysis is inherently required for pathway elucidation, for example to identify pathway intermediates, to

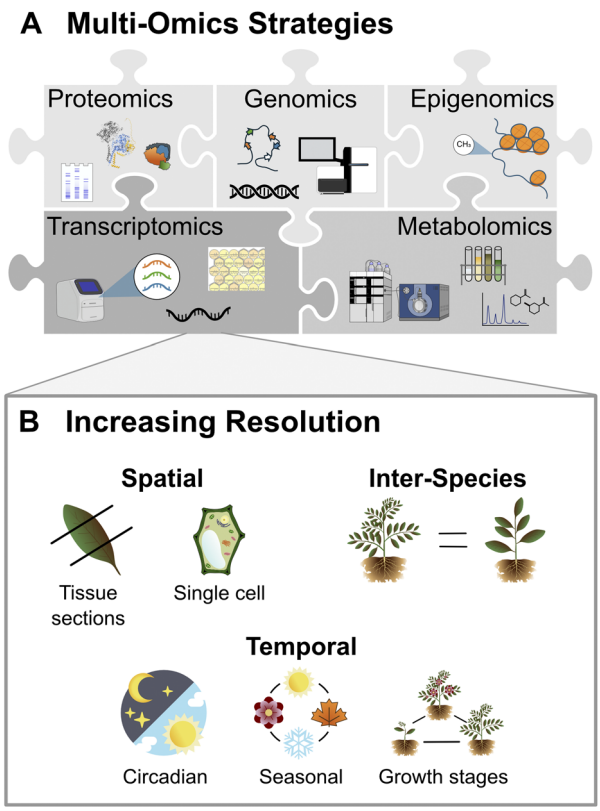


Fig. 1 Improving traditional co-expression analysis by integration of multi-omics strategies and approaches to increase its resolution. (A) Combination of transcriptomics with other omics techniques. (B) Increasing the resolution of co-expression analyses by integrating spatial and temporal information or by inter-species comparison (phylotranscriptomics). Icons from BioIcons and FreeSVG were used: 'qpcr_machine' by KeHan (CC0), 'LC-MS system' by funkyfoodscience (Public Domain) and 'genomesequencer-2' by DBCLS (CC-BY 4.0, modified).

monitor enzyme assays, or to analyse heterologous reconstitution efforts. Consequently, metabolomics data and equipment are often already accessible. In its simplest form, this combination of omics datasets can be merely conceptual, based on separate analysis of the datasets followed by a joint interpretation.[20,21] Even such a conceptual integration can strengthen the selection process of candidate genes when co-expression patterns align with metabolite abundance or when tissues in which relevant genes are likely to be active are highlighted. Two recent examples showcasing the application of this approach include the elucidation of ipecac alkaloid biosynthesis[22] and saponin biosynthesis.[23] A particularly important application of metabolomics analyses is to decrypt effects from metabolite transport throughout tissues, which often blur any differences in biosynthetic activity of different plant parts. A powerful strategy to counter such transport effects and to reveal tissue sections which are biosynthetically active has been recently reported based on feeding of isotope-labelled intermediates or even simply deuterated water.[24,25] This enabled the identification of plant parts that actively produce metabolites to inform the selection of samples for transcriptome sequencing, resulting in successful gene discovery.

Beyond the simple, solely conceptual combination of transcriptomic and metabolomic datasets, such data can also be integrated and analysed collectively within a single statistical framework.[20] This statistical integration can reveal relationships and associations between different omics datasets based on statistical methods. Cavill *et al.* provide an in-depth overview of different concepts how such statistical integration can be achieved, for example *via* correlation-based data integration.[20] Focusing on examples of pathway elucidation, statistical integration of transcriptomics and metabolomics data has been applied to propose candidate genes in dihydrochalcone biosynthesis in sweet tea[26] and in neolignan biosynthesis in *Magnolia*.[27] In both cases, candidate genes were only predicted but not experimentally validated, highlighting that statistical integration for direct pathway elucidation remains relatively rare compared to simpler conceptual approaches, potentially caused by a gap between researchers with advanced statistical and computational expertise and experimental scientists. Nevertheless, statistical integration is frequently used in related research topics, such as exploring system-wide responses to stress,[28,29] pathway regulation mechanisms,[30–32] and evolutionary analyses,[33] offering potentially transferable strategies for multi-omics integration.

A few of these studies, for example the investigation of heat stress responses in tea plants,[28] illustrate the use of untargeted, large-scale metabolomic approaches integrated with transcriptomics. Rather than focusing on specific metabolite classes, such approaches aim at a broader, system-level characterisation of metabolic changes. This more comprehensive metabolomic coverage might also be very valuable for biosynthetic pathway elucidation, especially if substrates, intermediates, or products are unknown, transient, or distributed across interconnected metabolic networks. Untargeted metabolomics can therefore provide unbiased insights into metabolic alterations and may help reveal connections between biosynthetic pathways and other metabolic processes that would not be captured by targeted analyses alone.

Despite the growing availability of multi-omics datasets in plant research, integrating transcriptomic, metabolomic, and other omics layers remains challenging. The data are high-dimensional, heterogeneous, and often interdependent, limiting the effectiveness of simpler statistical approaches.[34] Computational tools and platforms are being developed rapidly to manage, visualise, and partially integrate these datasets. Notable approaches from the perspective of plant science with recent applications include PaintOmics,[35,36] OmicsAnalyst,[37] and Holomics.[38] On the web server PaintOmics, multi-omics data can be integrated by mapping them onto biological pathways.[35,36] PaintOmics has been used, for example, to analyse the phenylpropanoid pathway in *Sorghum bicolor*, enabling assessment of both gene expression and metabolite changes under different conditions;[39] in another example, proteomic and metabolomic data from safflower seeds were integrated, providing pathway-level visualisation of coordinated alterations across metabolic processes.[40] The web-based platform OmicsAnalyst focuses on statistical integration approaches, such as correlation-based analysis, combined with multiple visualisation options such as heatmaps and scatter plots.[37] For example, Berková *et al.* used OmicsAnalyst to investigate cadmium-induced changes in protein abundance and metabolite levels in flax.[41] Similarly, Holomics can be used for statistical integration of multi-omics data, as demonstrated for studying the effects of hormone levels on other signalling pathways in germinating barley embryos.[42] Compared to the previous two web-based platforms, Holomics is provided as an easy-to-use R application.[38]

Although the availability of such user-friendly tools for multi-omics integration is improving, fully harnessing the potential of multi-omics data and deriving meaningful biological insights often requires even more advanced approaches. In Section 5, we will examine machine learning methods that build on these statistical foundations to perform integration of multi-omics data.

2.2. Increasing the temporal and spatial resolution

To successfully improve co-expression analyses, samples should be included in which a given biosynthetic pathway is highly active, together with samples where the biosynthetic pathway is dormant. This can be achieved by finding conditions that induce a biosynthetic pathway, for example based on phytohormone treatment,[43] abiotic stress,[43,44] or biotic stress such as pathogen treatment.[14,45] Theoretically, as many biosynthetic pathways exhibit differential activity throughout the developmental cycle of a plant, increasing the temporal resolution of expression data by adding additional time points would be a conceivable option (Fig. 1B).[46,47] However, focusing on the temporal variation of biosynthetic pathways in plants is rarely used for gene discovery, possibly because the effect size is often rather small. One of the few successful examples comes from olives, where secoiridoid levels vary during maturation and ripening.[48] Instead, probably the most reliable approach to

obtain expression data with higher resolution is to improve the spatial resolution. Instead of bulk tissues, it is possible to analyse tissue sections, which can be prepared either manually or more precisely using laser microdissection.[49–51] Ideally, the selection of tissue parts is guided by spatially resolved metabolomics data that reveals metabolite distribution patterns and ideally also sites of active biosynthesis. Mass spectrometry imaging is a particularly powerful approach for this purpose.[52] The combination of laser-capture microdissection and mass spectrometry imaging was recently applied to obtain fundamental insights into quinolizidine alkaloid biosynthesis in lupins.[53]

Ultimately, this approach of improving the spatial resolution of expression data leads to single-cell or single-nucleus RNA-seq, techniques which are becoming more and more accessible and affordable in plant science.[54–56] Indeed, recent breakthrough studies highlight that expression patterns that are not obvious at the bulk tissue level often become much more pronounced at the single-cell level, revealing that often specific cell types are crucial for a biosynthetic pathway. For example, biosynthesis of the antidepressant natural product hyperforin is governed by specialised cells termed Hyper cells in *Hypericum perforatum*.[57] Single-cell transcriptomics was also successfully used in a tobacco species to study flower scent biosynthesis, which only occurs in distinct cells in the corolla.[16] Recently, single-cell transcriptomics helped to discover key genes for securinega alkaloid biosynthesis, which are expressed in a vasculature-associated cell type.[58] In monoterpene indole alkaloid biosynthesis, single-cell RNA-seq revealed even multiple different specialised cell types which co-operate in a coordinated manner.[59–61] Impressively, such single-cell transcriptomics data can now even be connected with metabolomics data of the same single cells to further improve predictions of which cell types are particularly relevant for a given pathway.[62]

Even at the single-cell level, not all biosynthetic pathways are easily resolved. Using single-cell data for gene discovery is particularly successful in cases where biosynthetic genes are expressed in a homogenous cluster of cells that corresponds to one discrete cell type. However, if biosynthetic genes are expressed in diverse cell types, it can again be difficult to identify clear patterns. In such cases, a powerful bioinformatics method termed consensus non-negative matrix factorisation (cNMF) can be applied to distinguish between identity programmes (which determine the cell type) and activity programmes (such as biosynthesis).[63] This cNMF method was recently successfully employed in a breakthrough study to identify previously missing genes from paclitaxel (Taxol) biosynthesis in yew tree needles.[15]

At the moment, single-cell and single-nucleus RNA-seq is still a fairly rare approach to discover biosynthetic genes in plants. The main reasons are that sequencing costs are still high compared to bulk RNA-seq and that experimental protocols to get high quality data are relatively challenging.[54] However, data analysis is becoming more streamlined, as databases and resources such as PlantscRNAdb 4.0 are developed which help to rapidly assign plant cell types.[64]

In addition to sequencing of individual cells or nuclei outside of the original tissue context, modern techniques for spatial transcriptomics are starting to provide spatially resolved expression data for intact tissues; an example is the commercial 10X Visium platform.[65,66] While earlier methods were limited in spatial resolution,[66] modern approaches such as Stereo-seq or the image-based MERFISH method can achieve cellular and even sub-cellular resolution.[65,67] Despite their potential, spatial transcriptomics techniques have not yet been applied for gene discovery in plants. One reason is that plant cells possess challenging features such as thick cell walls and autofluorescence which hinder common spatial transcriptomics protocols.[65] So far, two studies focused on known pathways such as flavonoid or carotenoid biosynthesis.[68,69] Considering that spatial transcriptomics can also be combined with other omics data such as metabolomics data from mass spectrometry imaging or single-cell RNA-seq,[65] we expect that it will become an important tool for gene discovery in the future.

As discussed above, it is possible that even at the single-cell level no apparent biosynthesis expression patterns can be observed unless suitable statistical methods to extract such patterns are applied. Furthermore, single-cell analysis is only advantageous if a biosynthetic pathway is cell-type specific; if this is not given, obtaining more bulk RNA-seq data is economically preferable. Nonetheless, we anticipate that further improvements in single-cell/single-nucleus RNA-seq as well as spatial transcriptomics and particularly better algorithms for data analysis will establish pathway elucidation at the single-cell level as a central strategy in the future.

2.3. Co-expression across species borders – phylotranscriptomics

Traditionally, co-expression analysis is carried out within a single species. Nonetheless, in principle, transcriptomics can also be applied – with certain restrictions – across different species or other taxonomic units, an approach typically termed phylotranscriptomics. In the simplest form, phylotranscriptomics can be used similarly to phylogenomics to reconstruct phylogenetic relationships by comparing transcript rather than gene sequences. When based on rigorous orthologue identification, sequence-based phylotranscriptomic trees can closely align with phylogenomic trees but offer a more cost-effective alternative.[70]

More compellingly, phylotranscriptomics can extend beyond sequence analysis and enable comparison of expression data across different species, which has the potential to uncover conserved patterns of gene activity and regulatory mechanisms. Although cross-species co-expression analyses are still rarely applied directly for pathway elucidation and candidate gene discovery, some intriguing examples from related research areas highlight the potential of expression-based phylotranscriptomics. For example, this approach was used to study the evolution of terpene biosynthesis in Pinaceae[71] and of allium flavour precursors in Asparagales.[72] Expression-based phylotranscriptomics has also been applied to elucidate regulatory mechanisms of plant pathways, as demonstrated in the genera

Plantago[73] and *Datura*.[74] However, in existing studies, cross-species co-expression analysis often remains at a conceptual level by comparing data at the conclusion level rather than through statistical co-analysis, reflecting the same trend described above for multi-omics. A major limitation is that the pathway of interest needs to be present and conserved across many plant species, which is often not the case in specialised metabolism. A further challenge for this approach is to find suitable normalisation methods to enable a valid comparison of expression data from different species. Therefore, statistical methods for integrating expression data from different plant species are highly valuable. For example, Lee *et al.* applied OrthoClust[75] to fuse species-specific orthology-based networks, enabling the identification of conserved co-expression modules.[76] Similarly, Passalacqua and Gillis aligned single-cell datasets across species by first measuring how genes were co-expressed within each species, then using these co-expression patterns to bring the datasets into a common framework based on functional similarity and to identify conserved expression patterns.[77] Although neither study was focused on biosynthetic pathway elucidation, both investigated conserved inter-species expression patterns, and their methods could potentially be applied to pathway gene identification.

Building on these methodological advances, a promising new tool called CoExpPhylo might have the potential to bridge the gap between phylotranscriptomics as a solely conceptual approach and actual gene discovery.[78] This computational pipeline integrates species-specific co-expression networks with phylogenetic relationships. By clustering orthologous genes that are both co-expressed and evolutionarily conserved, CoExpPhylo identifies candidate genes for specialised metabolic pathways across multiple species. It demonstrates the relevance and potential of phylotranscriptomics for the field of pathway elucidation.

3. Mining genome and other large sequence datasets

As outlined in Section 2, co-expression analysis, particularly when implementing recent improvements, is still the major approach for identifying biosynthetic genes in plants. However, an inherent limitation is the underlying assumption of the "guilt-by-association" principle. In reality, many of the genes that are co-expressed within a limited dataset might be functionally unrelated, possibly leading to incorrect gene candidates. More problematically, genes from the same biosynthetic pathway might have different expression patterns, for example if one gene is involved in multiple pathways, and would therefore be missed by co-expression analysis. Therefore, expression-based approaches alone are not an ideal basis for discovering new biosynthetic genes.

Genomic data offer a promising complementary approach to transcriptomics for identifying the genes underlying specialised metabolism (Fig. 2): they can provide other types of association among genes than co-expression and capture the full set of enzymatic functions encoded in an organism. Since the publication of the first plant genome sequence in 2000, assembly quality has improved dramatically with a concomitant drop in sequencing costs. Initial plant genome sequences were still highly fragmented, but long-read sequencing technologies such as PacBio and Oxford Nanopore Technologies sequencing, scaffolding techniques like Hi-C and Pore-C, and improved assembly algorithms now enable the generation of nearly gapless assemblies.[7,79] Additionally, further types of genomic data, such as epigenetic modifications, can complement a genome sequence. In this section, we introduce how genomic data can be leveraged to identify biosynthetic genes in plant specialised metabolism (Fig. 2).

3.1. Biosynthetic gene clusters in plants

It is well-established that biosynthetic genes in bacteria and fungi are commonly located next to each other in the genome. Such biosynthetic gene clusters (BGCs) were historically believed to be rare in plants; instead, it was assumed that the genes of biosynthetic pathways in plants would generally be scattered throughout the genome. This view was shaped by the fact that the genes from the first biosynthetic pathways that were elucidated in plants, for example for flavonoids, carotenoids and glucosinolates, were not clustered. However, this assumption has been continuously challenged by the increasing number of highly contiguous plant genome sequences and improving knowledge of biosynthetic pathways in plant specialised metabolism.[80] Initiated by the discovery of the BGC for the benzoxazinoid DIMBOA in *Zea mays* in 1997,[81] more than 40 biosynthetic gene clusters have been characterised in plants so far.[82] They span many compound classes like terpenoids, alkaloids, and cyanogenic glycosides and have been identified in a wide range of plant species. Some of these plant gene clusters can be visually explored in the Minimum Information about a Biosynthetic Gene cluster (MIBiG) repository.[83] It appears that BGCs often represent evolutionary innovations in plant defence.[84]

The generally accepted definition of a BGC states that it should contain at least three non-homologous biosynthetic genes participating in the same pathway.[85] In contrast, gene arrays in plant genomes comprise tandem gene duplicates involving about 65% of all annotated plant genes.[86] Similar to microbial BGCs, plant BGCs also often contain a characteristic signature gene, initiating a specialised metabolic pathway (Fig. 3). However, gene clusters from microbes and plants show substantial architectural differences: those found in plants more frequently contain intervening, unrelated genes (Fig. 3). Additionally, it is less common for plant BGCs to contain the full set of genes required for a biosynthetic pathway. Such BGCs have been termed "compact",[87,88] while those containing only a portion of the genes constituting a pathway are referred to as "loose"[88] or "split"[80] clusters. The core of a loose cluster can be complemented by another BGC, pairs of pathway genes called satellite (sub)groups, or single peripheral genes (Fig. 3B).[87] The size of plant BGCs depends on the definition, but usually ranges from several dozen kbp[89] to over one Mbp.[90]

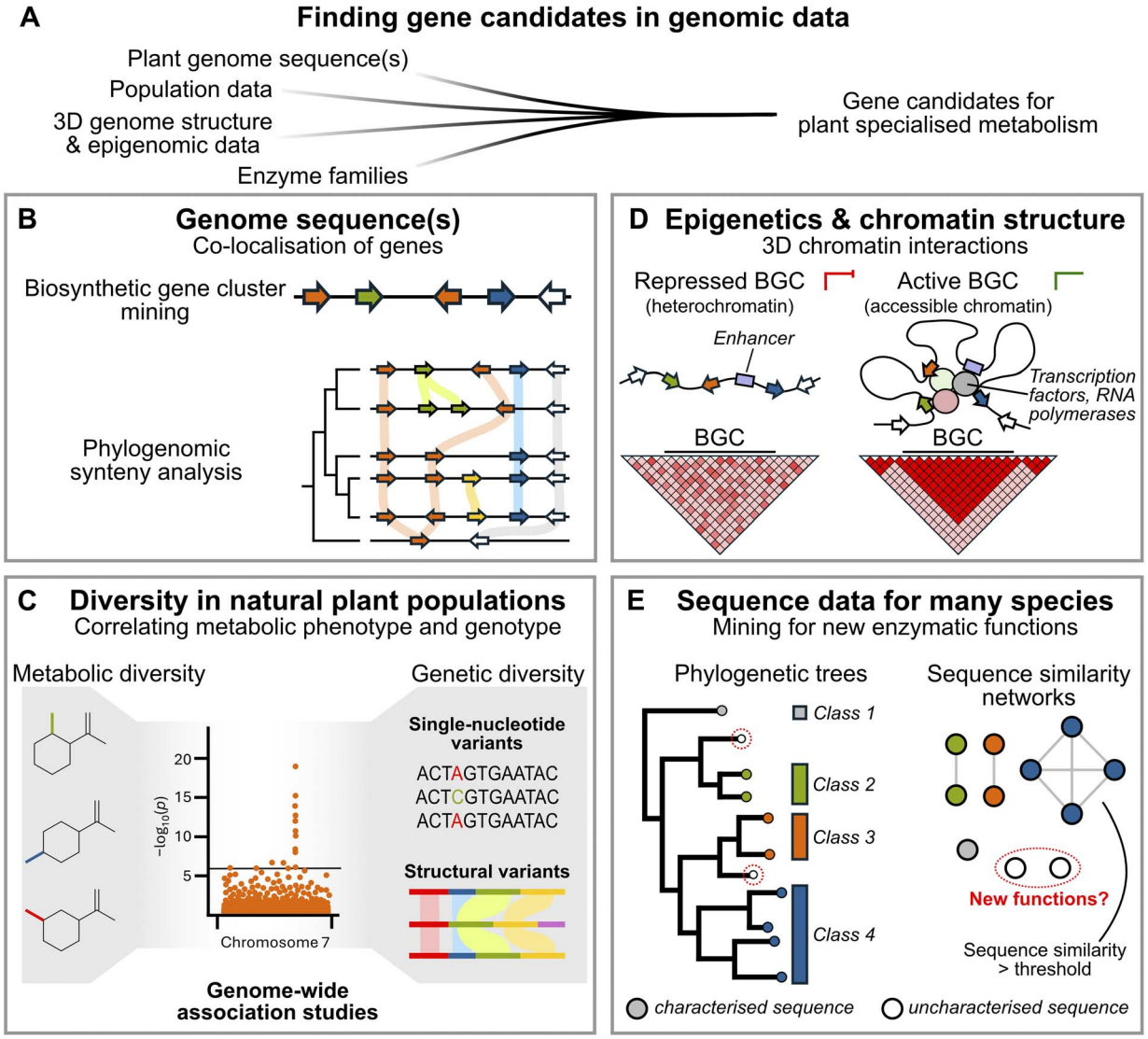


Fig. 2 Genomics-based approaches for discovering genes involved in specialised metabolism in plants. (A) Overview of different genome-related data types that can be leveraged. (B) Approaches based on co-localisation of genes. (C) Linking phenotype and genotype. (D) Biosynthetic genes in the context of epigenomics. (E) Finding exotic members of a gene family using large-scale phylogenetic trees or sequence similarity networks.

While the evolutionary emergence of gene clusters in plants requires further investigation, there are two plausible scenarios, which differ in the order of gene movement and neo-functionalisation: (1) genes evolve while dispersed and are then brought together as a BGC or (2) genes move first into a cluster region, where they evolve into a pathway.[80] Clustering of pathway genes into a BGC is hypothesised to offer several advantages for plants. It ensures that the whole set of genes can be inherited together, avoiding the accumulation of toxic intermediates formed by partial pathways.[80,91] Physical proximity of genes with a shared function could also facilitate their co-regulation: BGCs have been shown to form distinct chromatin domains which share epigenetic markers.[80]

The fact that biosynthetic genes in plants can be physically clustered in the genome offers a complementary alternative to co-expression analysis for identifying biosynthetic genes. The following subsections will introduce different approaches to identify and characterise biosynthetic gene clusters in plants based on genomic data.

3.2. Finding biosynthetic gene clusters in plant genomes

With increasing numbers of plant genome sequences available, more and more biosynthetic gene clusters (BGCs) encoding specialised metabolites have been identified in plants.[80] Searching for such potential BGCs is a major alternative to finding biosynthetic genes *via* co-expression analysis. In the following, we will describe the two major approaches that can be used to explore gene clustering (Fig. 4): the first is to focus on the functional annotations of adjacent genes in only one plant species, which is frequently employed to mine genome

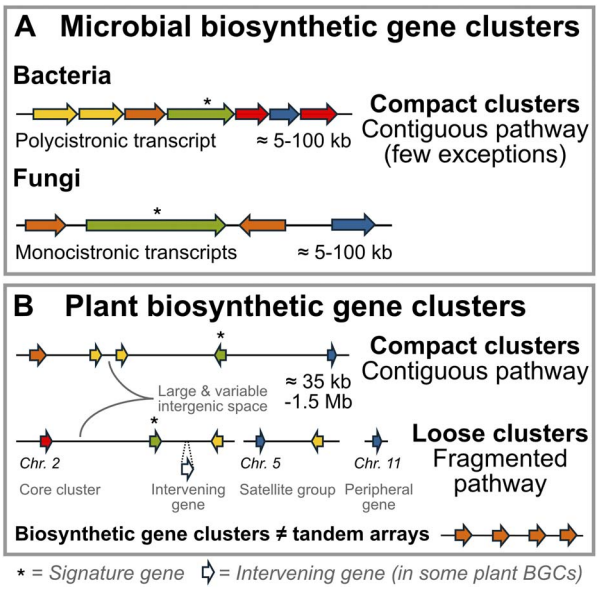


Fig. 3 Comparison of typical biosynthetic gene clusters (BGCs) from microbes (A) and from plants (B). In contrast to microbial BGCs, plant BGCs contain large and variable intergenic regions and pathways are often fragmented.

sequences for BGCs (Fig. 4A). The second, called comparative genomic or phylogenomic approach, adds another layer of information by comparing the organisation of adjacent genes across different species (Fig. 4B).

3.2.1. Genome mining for BGCs. Once a genome sequence of the target plant species is at hand, the generation of a structural and functional annotation is required to enable screening for putative BGCs. In case one or more pathway genes have already been identified before, a good approach is to manually inspect the genomic neighbourhood of these bait genes. If one or several neighbouring genes belong to gene families common in specialised metabolism, these are promising candidates for functional characterisation. This simple approach has enabled the initial identification of many clustered biosynthetic genes in plants. One example is the biosynthesis of the alkaloid gramine in barley (*Hordeum vulgare*): while the enzyme performing the second step of this short two-step pathway had been known since 2006,[92] the enzyme responsible for the first step stayed elusive much longer. An old hypothesis about the involvement of a pyridoxal phosphate-dependent enzyme was refuted in 2024, when Dias *et al.* discovered the oxidase CYP76M57 based on gene clustering.[93]

When no dedicated pathway enzyme has been reported yet, this targeted genomic neighbourhood analysis cannot be applied. For such cases, several tools to mine plant genomes for BGCs without known bait genes have been developed. These are strongly inspired by bacterial and fungal genome mining tools such as antiSMASH,[94] which are nowadays more or less indispensable for discovering new natural products and elucidating pathways in microbial natural product research. All these tools follow the same principle, which is to detect clustering of genes with functional annotations related to specialised metabolism. Such genome mining tools use unannotated genome sequences as their main input and identify genes from relevant families using profile Hidden Markov Models (pHMM). These algorithms use characteristic motifs, which have to be defined in advance, to infer gene families. The approach of tools like antiSMASH has been termed "cluster-first" or "function-centric", since they heavily rely on such pre-defined motifs for expected gene families. Since plant genomes differ strongly from their bacterial and fungal counterparts, tools which follow the original logic but are adapted for plants have been created: the most highly cited one, plantiSMASH (online and local, now V2.0),[95] is directly based on the antiSMASH framework.[94] Examples of the main adjustments implemented include the addition of plant-specific gene families to the motif catalogue and respecting the increased intergenic space in plant genomes. The PhytoClust tool[96] used a very similar logic like plantiSMASH but is not maintained anymore. A slightly different but also function-centric detection approach was implemented in the PlantClusterFinder algorithm, which is available as local software.[97] To increase confidence in its predictions, plantiSMASH offers the CoExpress module to integrate co-expression data after BGC detection. Since genes from metabolic gene clusters are usually strongly co-expressed, such an analysis should increase the accuracy of predictions. Apart from co-expression analysis within a BGC, the CoExpress module can also detect further BGCs or individual genes that show expression correlation with a certain BGC.[98] The plantiSMASH pipeline also makes it possible to compare detected BGCs across different species using its ClusterBLAST function. To increase the power of this comparative approach, plantiSMASH hosts precomputed gene cluster predictions for *ca.* 430 species (in V2.0), corresponding to 30 423 putative BGCs in the plantiSMASH 2.0 database.[95] Additionally, the plantiSMASH pipeline offers two more modules to aid in prioritising BGC candidates for functional characterisation: one infers enzyme substrate classes by grouping enzymes into characterised enzyme family subclasses with known substrate types. The other detects transcription factor binding sites, which allows researchers to choose candidate BGCs based on regulatory patterns. Finally, plantiSMASH 2.0 offers a comprehensive developer wiki, enabling researchers to define their own detection rules and enzyme subfamilies and to build custom BGC databases.[95]

The predicted gene clusters from such plant-specific genome mining tools have been experimentally verified in several cases, which led to the elucidation of several pathways. Successful discoveries using plantiSMASH include two BGCs for hyperforin biosynthesis in St John's wort (*Hypericum perforatum*)[99] and two gene clusters involved in QS saponin production in the Chilean soapbark tree (*Quillaja saponaria*).[100] A BGC predicted from the tomato (*Solanum lycopersicum*) genome using PlantClusterFinder led to the characterization of an unusual enzymatic activity of a chalcone synthase-like enzyme, possessing aminoacylation activity during the production of a hydroxycinnamic acid amide.[101] However, this genome mining workflow is still not employed as routinely as for microbes.

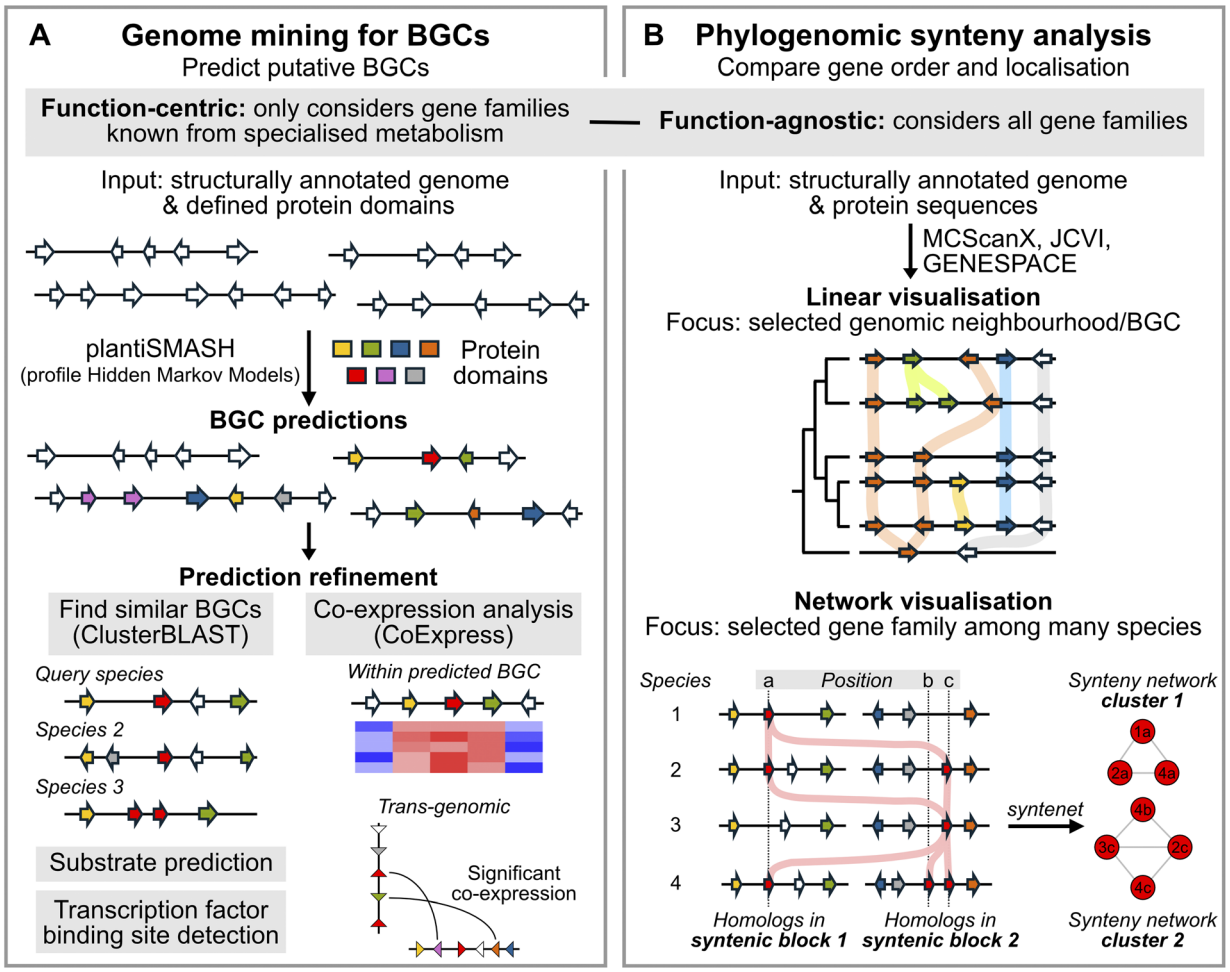


Fig. 4 Workflows of the two major approaches for genomics-based discovery of biosynthetic genes from plants. (A) Genome mining to detect BGCs based on a function-centric approach. (B) Synteny analysis incorporating phylogenomic information as a function-agnostic approach.

To conclude, the described function-centric BGC mining tools provide a valuable toolkit to quickly analyse plant genome sequences to predict potential BGCs. While their focus on functional annotation allows them to detect many BGC classes reliably, their main drawback is that they fail to consider non-canonical gene families.[102] For such more unusual biosynthetic systems, it can be worthwhile to employ phylogenomic approaches to explore gene clustering from a different viewpoint.

3.2.2. Phylogenomic and synteny approaches. Tools like plantiSMASH use Hidden Markov Models to identify putative biosynthetic gene clusters in a plant genome using functional gene annotations.[95] They can detect a large variety of BGCs using only one genome sequence as input. Results are often restricted to known biosynthetic gene families to maintain a high specificity, which can be a limitation for clusters that contain unexpected gene families. Considering that BGCs in plants show substantial plasticity,[80] this is a major challenge. To detect non-canonical BGCs and to explore BGC architecture and evolution in more detail, another level of information is therefore required to reach reliable predictions. This can be provided by comparing the genome architecture of closely related species against more distant species – an approach described as comparative genomics or phylogenomics. In contrast to transcriptomes, genomes offer additional information about the position, order, orientation, and structure of genes. Many methods and tools exist to compare these across species. Conserved gene order among genomic sequences is called collinearity. The term synteny, which is often used interchangeably, describes the fact that collinear genomic segments are derived from a common ancestor and thereby retain their organisation.[103] The latter is usually divided into macrosynteny, referring to chromosome-scale observations, while conservation at the level of several genes in a local context is termed microsynteny.[104] In this section, we focus on gene-anchored synteny analysis. Plant genomes are more plastic than those of other eukaryotes.[105] New genes and even gene clusters can form at new genomic positions by genomic rearrangements.[80,87] This can result in novel traits, such as the production of certain specialised metabolites. Conversely, genes can be readily lost when selection pressure on them disappears.[106] Therefore, genes or a BGC required for formation of a certain natural product might be found by comparing genome sequences of producing species with non-producing species as outgroups, as

non-producing species should lack the required gene set partially or completely. This approach can be termed "function-agnostic", since it compares gene order for all gene families without being limited to specific gene families. If a pathway gene is already known, it can be used as a starting point to compare the collinearity at this genomic location between producing and non-producing species. Many suitable synteny analysis tools have been developed, most notably MCScanX.[107] MCScanX only requires structurally annotated genome sequences and lists of all encoded protein sequences to detect syntenic relationships among species. In the first step, MCScanX performs an all-versus-all BLASTP search to identify homologous gene pairs. The runtime of this step can be very high, especially when many genome sequences are to be compared.[108] Therefore, several implementations of MCScanX employ faster sequence aligners, like LAST[109] in the JCVI library[104] and DIAMOND[110] in TBtools-II.[111] MCScanX and its JCVI implementation were recently applied to characterise the BGC involved in the biosynthesis of withanolide-type steroids in *Solanaceae*.[90,112,113] After locating the only previously known pathway gene in the genome, the authors of these respective studies recognised clustering of genes from gene families typical for specialised metabolism in its neighbourhood. Synteny analysis revealed that this cluster was absent in related non-producing species such as tomato, increasing confidence in the BGC.

In principle, synteny analysis of BGCs from related species opens the door to studying species-specific differences in a pathway and its evolution. However, synteny analyses become increasingly complicated when more distantly related species are compared. This is especially the case for plants, where whole-genome duplications, deletions, and other large-scale genome rearrangements are very common. These processes confound the distinction between orthologues, *i.e.*, homologues resulting from a speciation event, and paralogues, which are created by gene or genome duplication.[114] Differentiating these correctly is important for comparative genomic studies. To tackle this problem, the GENESPACE tool was developed.[115] It integrates information about synteny (conserved gene order) into the process of orthologue inference (sequence similarity), which makes the tool capable of comparing even complex genomes from distantly related plant species.

To enable large-scale synteny comparison of gene families, Zhao *et al.*[116,117] created a visualisation and analysis workflow based on networks (*syntenet*, available as an R package[118]). Their so-called synteny networks consist of nodes, which represent genes, and edges, which connect those nodes having a syntenic relationship. Such a network builds on pairwise genome comparisons performed by MCScanX but extends the results by abstracting the detected synteny blocks into a complex network. This approach is valuable to study the evolution of large gene families, as synteny networks can reveal genes which have been recruited to new genomic contexts. More importantly for gene discovery, synteny clusters (genes at a syntenic position) specific for a certain lineage can be identified, to visualise the specific genomic neighbourhood of a query gene. Li *et al.*[119] used this synteny neighbourhood network strategy to analyse the flanking genes around oxidosqualene cyclase genes in 122 plant species and revealed several associated gene families, particularly CYP716s. Such an investigation can thereby provide insights into lineage-specific gene associations and BGC evolution.

Finally, to develop improved strategies for discovering BGCs in plants, it could be worthwhile to look at similar approaches in fungi rather than bacteria, as fungi are also eukaryotes. Several tools for automated gene cluster prediction based on conserved synteny have been developed for fungal genomes.[120–122] Since function-centric prediction algorithms like antiSMASH/plantiSMASH only detect BGCs similar to canonical ones, the motivation behind these synteny-based tools is to uncover more cryptic gene clusters. The most recent of these algorithms, CLOCI (Co-occurrence Locus and Orthologous Cluster Identifier), identifies loci with unexpectedly shared microsynteny and compares these among species using alignments and phylogenetic methods.[122] In this way, it was able to detect gene clusters missed by the function-centric antiSMASH. Such phylogenomics-based prediction algorithms require a sufficient collection of genome sequences and have not been adapted for plants yet. However, the growing number of available plant genome sequences might pave the way for similar plant-suitable tools, possibly leading to the prediction of unknown BGC classes in plants.

To summarise, a phylogenomic approach based on synteny analysis can serve as a promising tool to find and analyse biosynthetic gene clusters in plants.

3.3. Linking genotype and metabolic phenotype in plant populations

Genomic data are a powerful resource for the discovery of biosynthetic gene clusters in plants. While the described approaches to find BGCs are exclusively sequence-based, genomic data can also be integrated with metabolic information to aid gene discovery. This principle has been employed to reveal genotype-phenotype connections in plant populations. To identify genes contributing to a quantitative phenotype, *e.g.*, the production of a specialised metabolite, populations with broad genetic variation are phenotyped and genotyped. Quantitative trait loci (QTL) harbour the gene(s) responsible for a certain trait and are often identified through genome-wide association studies (GWAS) or mapping-by-sequencing (MBS). MBS typically utilises populations derived from bi- or multi-parental crosses, for which parents with diverse phenotypes and genotypes are selected, while GWAS harness natural intraspecific variation. Traditionally, the genetic data used for GWAS were single nucleotide variants (SNVs) which were called using microarrays. As soon as a reference genome sequence is available for a species, members of a population can be re-sequenced and the reads can be aligned to the reference to call SNVs.[123] When the observed phenotype is a metabolite, studying the association between metabolite quantity and genetic markers is called metabolic GWAS (mGWAS)[124] or qualitative trait GWAS (QT-GWAS)[125] when the phenotype is binary. Such an approach can for example identify knockout mutations that are causal for loss of production in

a population.[125,126] An advantage of GWAS for finding biosynthetic genes is that it is an untargeted method without prior assumptions about the involved gene families. However, conducting such association studies requires populations with intraspecific phenotypic variation. Therefore, this approach has mainly been adapted for crop plants or model species for which large germplasm collections exist. During the last decade, another innovation in plant genomics has revolutionised association studies. The availability of genome sequences of multiple accessions of the same species has led to the generation of plant pangenomes.[127] These datasets revealed extensive intraspecific structural variation and, by combining the genetic information from all accessions, cover the full set of genes and non-coding sequences of a species. Since structural variants (SVs) can strongly influence the biosynthesis of a specialised metabolite,[128] the structural variations identified in a pangenome can be leveraged for association studies.[127] Using this approach, Zhu *et al.*[129] investigated the genes involved in metabolite biosynthesis in potato tubers. They constructed a pangenome from 29 genome assemblies and more than 200 re-sequenced potato accessions and complemented this genetic resource with metabolomic analyses. By this approach, the researchers identified more than 9000 structural variations significantly associated with hundreds of metabolites.[129] While employing SVs instead of SNVs can improve the accuracy of GWAS,[129,130] identifying and using such SVs in plant pangenomes have their own challenges.[129,131] Graph-based pangenome representations have proven superior over linear reference genome sequences for SV identification, but their construction and application can be difficult; this is especially true for species with complex genomes due to polyploidy, high heterozygosity or a large genome size. Additionally, the number of genomes that can be processed is limited by the available computational resources. Therefore, researchers have to balance computational capacities and desired accuracy when constructing a graph-based pangenome representation for SV identification.[131] Furthermore, when short-read sequencing is used for genotyping large populations, resolving complex SVs exhibits limited reliability in highly repetitive regions. Despite these challenges, GWAS and other methods which harness the diversity of plant populations offer a valuable approach to shortlist gene candidates for metabolic pathways.

3.4. Epigenomics

As outlined, the genome sequence of a plant species enables the identification of biosynthetic gene candidates by BGC mining, synteny analysis and metabolite–gene associations. However, a genome sequence only sets the starting point for genome-based analyses. An additional layer of information can be data about epigenetic modifications and the 3D chromosomal structure – both providing insight into the regulatory processes governing gene expression.[132]

Well-characterised epigenetic features include DNA methylation, typically analysed *via* whole-genome bisulfite sequencing (WGBS)[133] or increasingly *via* long-read platforms such as Oxford Nanopore Technologies and PacBio.[132] Furthermore, histone modifications and variants are explored by techniques such as ChIP-Seq or CUT&Tag.[132] Information about these epigenetic markers can be obtained at the whole-genome scale[132] and could therefore be used for epigenome-wide association studies (EWAS), which link epialleles in a population with phenotypic variation.[134] This approach can be adapted to identify gene candidates for specialised metabolism in cases where phenotypic variation is not based on genomic sequence variation but on epigenetic differences. For example, Guo *et al.*[135] performed a metabolome-based epigenome-wide association study (mEWAS) using WGBS data in tomato and compared it with traditional SNP-based mGWAS. Their mEWAS identified several glycosyltransferase genes in differentially methylated regions, which were subsequently confirmed to be involved in specialised metabolism.[135]

Several histone modifications have been shown to influence chromatin accessibility and gene expression: for example, the histone methylation mark H3K27me3 is known to repress the expression of a region (in facultative heterochromatin), while H3K4me and histone acetylation have an activating effect (in euchromatin).[136] In human genetics, epigenome-wide association studies using histone modifications have helped to characterise genes involved in diseases and disorders.[137] While no association studies with histone modifications and phenotypes have been performed in plants yet,[134] their influence on the formation of complex chromatin structures in plant BGCs has already been investigated.[78,125]

Chromosome conformation capture techniques based on Hi-C allow for the direct analysis of such 3D genomic structures.[138] Chromatin is organised into different conformations at hierarchical scales: at the largest scale (tens of megabases), a chromosome can be divided into accessible euchromatic A and inaccessible heterochromatic B compartments.[139,140] At a finer scale (hundreds of kilobases), topologically associating domains (TADs) – or TAD-like domains in plants – define genomic regions that interact more frequently with themselves than neighbouring regions.[139,140]

Several plant BGCs have been found to form TAD-like domains, which may facilitate the shared regulation of their constituent genes.[89] Studies in *Arabidopsis thaliana*[141] and tomato[142] showed that these BGCs, when active, reside in strongly interacting regions of accessible chromatin. The interactions among the promoter regions in these BGCs are facilitated by the formation of short chromatin loops between genes. Enhancers play an important role in the formation of these interactions: by recruiting transcription factors and RNA polymerases to promoters of a BGC, they allow transcriptional boosting of their target genes.[142,143] The resulting complexes of regulated genes, enhancers, and proteins have been termed transcriptional condensates[144] or, in analogy to prokaryotes, topological operons.[145] In some cases, like the withanolide BGC in *Physalis grisea*, biosynthetic gene clusters can even comprise several TAD-like domains, reflecting different tissue-specific expression patterns.[113] Information about the physical interaction of chromatin regions can be leveraged for prioritising gene candidates for specialised metabolism: Li *et al.*[59] found that a gene encoding a transporter is clustered with two known

monoterpene indole alkaloid biosynthetic genes in *Catharanthus roseus*. The fact that all three genes were part of the same TAD-like domain led them to characterise the transporter, finding that it is specific for the pathway intermediate secologanin. In the future, long read-based approaches using Pore-C[146] or CiFi[147] could supplement or even replace Hi-C in the chromatin structure analysis.

Beyond the active physical interactions of genes within a BGC, the clustered genes can also show long-range intra- and interchromosomal contacts with BGCs or gene arrays outside of the cluster. Such long-range chromatin loops are linked to heterochromatic regions and the characteristic repressive histone mark H3K27me3. Using this mechanism, plants[139,148] and other organisms[149] can co-silence functionally related BGCs or genes which are distantly located in the genome sequence. For example, while analysing monoterpenoid biosynthesis in *Agastache rugosa*, Liu *et al.*[150] found that the pulegone monoterpenoid BGC on chromosome 8 significantly interacts with an array of known pulegone reductase genes on chromosome 2 which participate in the same pathway.

Given that co-regulated biosynthetic genes can interact over long ranges, 3D genome data holds potential for identifying new pathway candidates based on physical interactions with bait genes. To date, no biosynthetic gene in plants has been discovered using such a strategy. However, in human genetics, Hi-C data has already been employed to identify candidate genes involved in complex diseases *via* their interactions with enhancers.[151]

To conclude, epigenetic data and information about chromatin topology can help to understand the regulation of specialised metabolism. Additionally, they might offer potential to identify new gene candidates *via* association studies and the analysis of gene interactions.

3.5. Mining enzyme families

Lastly, the availability of large amounts of sequence data from plants also enables applications that extend beyond the common focus on smaller lineages, gene clusters, and synteny. One of the most promising approaches is to investigate enzyme families which share the same substrate but lead to diverse products. With large sequence data pools, such enzyme families can be mined to discover rare representatives with new biochemistry. This was demonstrated by two recent studies[152,153] which independently investigated the sequence diversity of oxidosqualene cyclases in plants, enzymes which catalyse the entry step into triterpenoid biosynthesis and are crucial to generate polycyclic carbon skeletons from the simple precursor oxidosqualene. In both cases, oxidosqualene cyclases capable of producing previously inaccessible or even new carbon skeletons were discovered from hundreds of sequence datasets.[152,153] A practical challenge is to process the thousands of homologous enzyme sequences that can be extracted by such an approach and to find effective strategies to predict which of these represent viable candidates for functional investigation. The study by Stephenson *et al.* used a phylogenetic tree to select enzymes from divergent branches or underrepresented clades,[153] whereas the study by Hakim *et al.* used sequence similarity networks for a similar purpose. While phylogenetic trees are highly useful to reflect the evolutionary history of the gene family, sequence similarity networks are easier to compute at scale and enable seamless metadata integration and visualisation in a network format.[152] Databases which collect information about certain enzyme families (*e.g.*, MARTS-DB focused on terpene synthases[154]) are a very useful source of reference data for such analyses. It can be expected that these approaches will be further expanded to other enzyme classes and integrated with other uses of sequence data such as searching for associated biosynthetic gene clusters in the future.

4. Identifying biosynthetic enzymes at the protein level

Since the advent of next-generation sequencing, biosynthetic gene discovery has been carried out almost exclusively at the gene level, fuelled by the increasing accessibility of gene sequences and expression data. As biochemical pathways revolve around enzymes, studying enzymes directly rather than genes would be advantageous. As enzyme levels can be highly dynamic, proteomics studies can provide important complementary information to transcriptome and genome data; however, the analysis can be technically demanding.[155–157] Traditionally, before modern sequencing technologies, direct purification of biosynthetic enzymes was a major, albeit very slow, method for pathway elucidation.[158,159] In this section, we will describe modern technologies such as affinity probes and proximity labelling that potentially enable pathway elucidation directly at the protein level. These methods offer greater power and efficiency than traditional protein-based approaches and can complement gene-based methods (Fig. 5).

4.1. Correlating protein levels in biosynthetic pathways

Gene-based co-expression analysis is based on the assumption that equal gene expression levels reflect equal protein activity, which is desirable for fruitful interactions of biosynthetic enzymes and efficient metabolic flux.[160] For many pathways, this assumption seems to hold true, as transcriptome-level analyses have been successfully used for the selection of candidates many times. However, there is an inherent risk that biosynthetic enzymes exhibit weak correlation between transcript and protein levels,[155,156,161] possibly in cases where pathway activity has to be modulated extremely quickly (Fig. 5). Such weak correlation could lead to an improper selection of candidate genes either by selecting false positive candidates that are not involved in the pathway or, more problematically, by missing candidates for which the genes falsely appear not to have a matching expression pattern. In such cases, correlating protein levels directly could improve the candidate selection;[157] at the same time, such an analysis can also give insights into translation efficiency, post-translational modifications and degradation.[160] Accessing such information could open relevant new avenues, as post-translational modifications are well-known regulatory elements in primary metabolism[162] but have

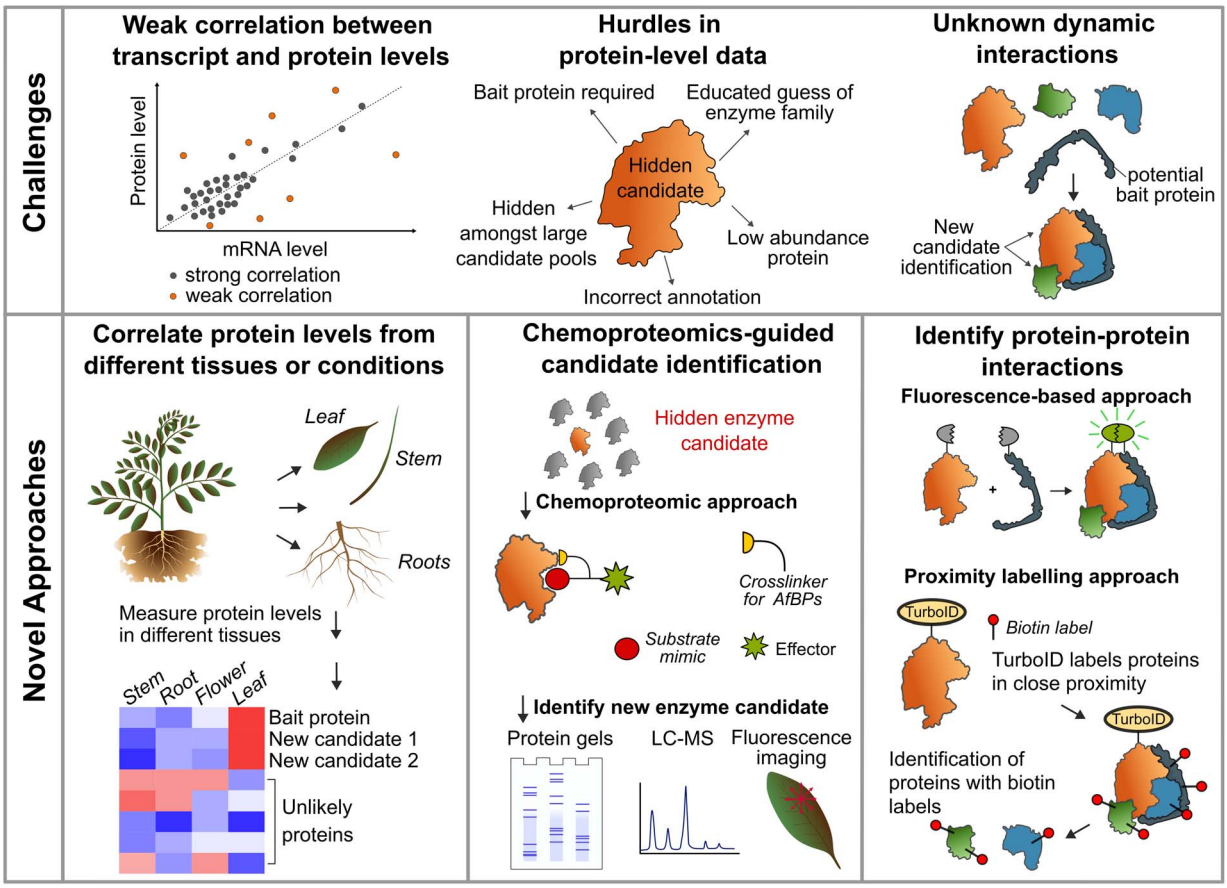


Fig. 5 Identifying biosynthetic enzymes at the protein level. Correlating protein levels can help identify overlooked enzyme candidates in cases of weak correlations between transcript and protein levels. Additionally, chemical probes for chemoproteomics and analyses of protein–protein interactions provide complementary approaches for enzyme identification. AfBP: Affinity-based probes.

been largely neglected for pathways from specialised metabolism.[163]

To determine protein levels, the first step is to identify proteins in complex biological samples. In principle, this could be achieved in a targeted or untargeted way. Conventional targeted methods like Western-Blot analyses require a specific antibody for a known enzyme or the introduction of a protein tag for detection and are therefore severely limited in throughput. For an untargeted and more unbiased approach, proteomics based on protein mass spectrometry is preferable, since mass spectrometry-based proteomics enables identification and quantification of thousands of proteins from complex samples without prior knowledge.[164] The workflow for such a proteomics approach includes sample preparation (cell/tissue disruption), proteolysis of proteins, separation of peptides by liquid chromatography and finally mass spectrometry detection. In the crucial data analysis step, protein abundances are determined based on associated peptide abundance as a proxy. Using a shotgun proteomics approach, the localisation and abundance of three key enzymes from morphine biosynthetic pathway could be identified in *Papaver somniferum*.[165] In another study, a cytochrome P450 monooxygenase with epoxidase activity was identified by comparing protein abundance patterns of known pathway enzymes with potential enzyme candidates.[166] Neopinone isomerase catalyses a reaction in opiate alkaloid biosynthesis that was originally assumed to occur spontaneously; this enzyme was identified based on protein abundance in different protein fractions such as latex and from different chemotypes, supported by gene expression data.[167] It is important to keep in mind that – in analogy to gene-based co-expression analyses – at least one suitable bait protein is required for correlating protein abundance patterns across different samples to find new biosynthetic enzymes by this approach (Fig. 5).

Two major limiting factors can hinder such a protein level correlation analysis, first the lack of suitable bait proteins, second proteins with very low abundance or proteins from rare cell types. These limitations create the need for additional methods to overcome these problems.

4.2. Trapping biosynthetic enzymes with chemical probes

A challenge of gene discovery based on traditional co-expression analysis is that often very large numbers of gene candidates are obtained, increasing the workload of functional screening.[12] Furthermore, co-expression analysis as well as many complementary techniques are based on prior knowledge of at least

one bait protein from the pathway and informed predictions of putative enzyme classes for the enzymatic reaction in question (Fig. 5). A possible complementary strategy to overcome these issues is to employ chemical probes.[168,169] These probes contain a binding moiety that targets the active site of a target enzyme, for example by acting as a substrate mimic. A functional substrate mimic probe must still be able to bind to the target enzyme; without prior knowledge of the enzyme–substrate interaction, it is therefore advantageous to design multiple probe variants to increase the chances of success. In general, this is similar to identifying the molecular targets of drugs;[170] therefore, some of these methods can be adapted for the identification of substrate–enzyme pairs in metabolic pathways. The probes can be classified by their type of binding and the type of reporter for subsequent detection of bound proteins. A major group are activity-based probes (ABPs), which comprise small molecules that covalently bind to and modify the active sites of enzymes; these are commonly linked to fluorescent dyes or biotin as reporters.[171] Affinity-based probes (AfBPs) additionally contain a photo-crosslinking group which can be activated by UV radiation to covalently link the probe with the target enzyme (Fig. 5).[171] In both cases, the reactive group is connected to the reporter *via* a linker. Reporters are commonly either fluorescence tags or affinity tags such as biotin. By using affinity tags, proteins bound to the probe can be purified, visualised on protein gels, and analysed using protein mass spectrometry. This allows the identification of previously unknown substrate–enzyme pairs. Furthermore, fluorescent probes can be leveraged for bioimaging to determine the localisation of target enzymes in tissues or cells.[172] To date, these approaches have been mostly used in drug research for identifying their cellular targets,[173] but there are also examples where the use of proteomics probes led to the identification of new biosynthetic enzymes in plant pathways. Using an affinity-based probe containing a photo-crosslinking group, a new UDP-glycosyltransferase was identified from *Stevia rebaudiana*.[174] This probe system contained a click chemistry system to flexibly add various reporters in a second step.[174] A similar approach was used to discover a novel FAD-dependent Diels-Alderase from *Morus alba*; this work exemplifies the discovery of a biosynthetic enzyme for which no enzyme class could have been reliably predicted *a priori* and traditional approaches would therefore have likely failed.[175] In an effort to elucidate the biosynthetic pathway of the anticancer drug precursor camptothecin, an epoxidase was found in *Camptotheca acuminata* using a chemoproteomics approach.[166] An activity-based fluorogenic probe was used to monitor the activity of *O*-methyltransferases directly in *Arabidopsis thaliana* based on fluorescence microscopy.[172] An issue of *in vivo* applications in plants is that bulky reporters or substrates might not be taken up by cells;[171] therefore, many published strategies include the use of lysed samples.[174] However, tissue lysis disrupts the cellular organisation, which is a limitation if different reactions of a pathway need to take place in different cell types. Methods used in other organism groups could also inspire further research in plants; for example, in the bacterium *Staphylococcus aureus*, a pyridoxal phosphate probe was used to access and annotate the PLP-dependent enzyme proteome.[176]

Although chemoproteomics probes could be a very promising complementary approach to discover biosynthetic enzymes in plants, there are some limitations of the system. Certain protein classes are more difficult to analyse by such an approach due to their membrane localisation or possible aggregation. Most importantly, the probe design is crucial to determine the success of such an endeavour. The probes have to be sensitive enough for proteins with low abundance; at the same time, the specificity has to be high enough to limit the number of unspecific candidates. The reporters and linker parts of the probes also must not prevent the binding of the probes to the target proteins. It is therefore good practice to test different probe designs for maximum chances of success. Chemical synthesis of the probes can also be challenging, as probes are inherently reactive, particularly when photo-crosslinking groups are integrated. Many synthetic routes start from the native substrate; it is therefore also essential that sufficient amounts of starting material are available for multi-step semi-synthetic modifications. Hence, probe-based enzyme discovery campaigns are highly dependent on successful interdisciplinary combination of proteomics and synthetic methodologies.[168]

4.3. Finding biosynthetic enzymes by protein–protein interactions

Biosynthetic enzymes in plants can be organised as protein complexes or at least as dynamic metabolons to ensure efficient metabolite production and control of enzymatic reactions.[177,178] For practical reasons, these interactions of proteins in complex pathways have often been neglected due to the difficult analysis. Nevertheless, studying such protein–protein interactions could not only help to improve our understanding of how such pathways work in plants but could also help to identify previously overlooked biosynthetic enzymes by probing the interactions of a bait protein with other proteins (Fig. 5).

Various methods are commonly used to study protein–protein interactions, most importantly biophysical methods based on fluorescence.[179] For example, a dynamic metabolon from *Sorghum bicolor* was characterised using Förster resonance energy transfer (FRET) with different fluorophores which were coupled to the proteins that were analysed.[180] With this technique, it can be monitored whether proteins are in close proximity to each other, as only then is fluorescence energy transferred.[179] In an alternative approach, the protein–protein interactions of a dynamic plant complex were studied using bimolecular fluorescence complementation (BiFC) in yeast.[181] In this work, the target enzymes were fused to non-fluorescent fragments of a fluorescent protein; only when the target enzymes were close to each other, the fragments could interact and reconstitute the original fluorescence protein activity.[181] While these fluorescence-based methods are crucial to provide evidence for physical contact between two proteins, they suffer from inherently low throughput. Therefore, only a very limited number of protein candidates can be tested by this approach.[181] Further improvements or alternative methods are therefore

needed to harness the potential of protein–protein interaction-based discovery of biosynthetic enzymes.

A possible alternative without any prior protein candidate selection could be proximity labelling (Fig. 5).[182] The idea behind this technique is that all enzymes in the proximity of a bait protein are enzymatically tagged, for example with biotin.[182] One possibility to achieve this enzymatic tagging is to fuse the bait protein with a biotin ligase such as TurboID.[182,183] All proteins that come close to the biotin ligase-fused bait protein will then get biotinylated. The labelled enzymes can later be identified by western blot analysis, proteomics, or even purified using the biotin label.[183] This approach was used to investigate the phenylpropanoid pathway in *Petunia inflata* protoplasts.[184] In this work, the cinnamic acid hydroxylase (C4H) was fused to TurboID; labelled target proteins were identified using mass spectrometry.[184] All candidates from the proximity labelling were independently validated by BiFC.[184] This work is an excellent demonstration that proximity labelling can in principle be used to find interacting enzymes in a biosynthetic pathway in plants and could thereby help to accelerate the discovery of new pathway enzymes. A main limitation is that proximity labelling requires a working transformation system and is therefore not applicable to many non-model plants. Furthermore, proximity labelling is a very challenging method and requires careful optimisation.[182] Nonetheless, the examples mentioned here show that searching for interacting proteins of a bait pathway enzyme can be a powerful strategy to find further pathway proteins. We anticipate that further methodological improvements will help to establish interaction-based methods as major alternatives to gene-based pathway elucidation.

5. Artificial intelligence-guided discovery of biosynthetic genes

The rapid expansion of omics technologies has made plant research increasingly data-rich, with thousands of sequenced genomes, extensive transcriptomic surveys, and growing metabolomic and proteomic datasets available.[7,185,186] While previous sections of this review highlighted how various omics technologies are used and combined for pathway elucidation, traditional mathematical models often face limitations when applied to complex datasets. In particular, multi-omics analyses involve multi-layered and heterogeneous data, and biological relationships may involve combinations of features across different molecular layers rather than simple pairwise associations.[187–189] As a result, extracting connections from multi-omics data remains difficult, even when conventional computational strategies are applied.

This complexity points to the potential of advanced computational approaches, such as artificial intelligence (AI) including machine learning (ML), which may help identify hidden patterns or relationships.[190,191] AI is defined as a computational system capable of intelligent decision-making, while ML, as a subfield, refers to algorithms that learn patterns from data.[192] In the field of plant specialised metabolism, such methods might help to integrate multiple omics layers, detect latent patterns, and select gene candidates in ways not achievable with conventional approaches.[193,194] Several recent reviews have emphasised how current AI tools can already accelerate biosynthetic pathway elucidation, guide hypothesis generation, and facilitate integration of diverse data sources.[18,187,189,191,193,195] In this section, we examine AI-driven approaches that are not yet employed in plant specialised metabolism but have the potential to support pathway elucidation, with a particular emphasis on methods that are already used in other areas of plant science or show clear promise for gene discovery. We distinguish between explorative AI approaches that can help reveal patterns in complex biological datasets and predictive AI approaches that infer unknown gene functions or pathway components by learning from known data (Fig. 6). Finally, we critically assess where these approaches currently fall short for plant research and which plant-specific constraints limit their practical utility.

5.1. Explorative AI to spot hidden gene candidates

We define explorative AI here as computational tools which are used to detect hidden patterns in heterogeneous biological datasets. Its goal is to guide hypothesis generation and to uncover nonlinear relationships that complement or surpass traditional methods not based on AI. Explorative methods typically rely on unsupervised learning – such as clustering, dimensionality reduction, or latent-factor analysis – to reveal functional patterns within a dataset of interest (Fig. 6).

A relevant explorative AI approach illustrating this capability is the use of self-organising maps (SOMs), an unsupervised neural network method developed in the 1980s.[196,197] SOMs project high-dimensional input data onto a two-dimensional grid, grouping nodes with shared characteristics. When applied to transcriptomic datasets, they enable clustering of genes with similar expression patterns, helping to identify co-expression modules and prioritise candidate biosynthetic genes for pathway discovery. Recent studies on alkaloid and terpene biosynthesis have shown how SOM-based clustering can successfully guide the discovery of biosynthetic enzymes and transporters involved in a pathway of interest.[13,198–200] In plant biology in general, SOMs have also been applied to areas like metabolomic profiles,[201] mass spectrometry datasets,[202] and phenotypic measurements,[203] further underscoring their utility for explorative analysis beyond transcriptomic data.

While SOMs are typically applied to single datasets, other machine learning approaches are designed to integrate multiple omics layers. Most established tools for multi-omics integration originated in human or animal research, particularly in cancer and disease studies.[204,205] Prominent examples include MOFA,[206] DIABLO,[207] iCluster,[208] MOGONET,[209] SNF,[210] and MOMA,[211] which span a diverse range of methodologies such as latent factor modelling, network-based data integration, and deep learning approaches. Despite the increasing availability of plant multi-omics datasets, computational analyses that truly integrate different omics layers remain relatively uncommon in plant research, and the use of ML approaches to

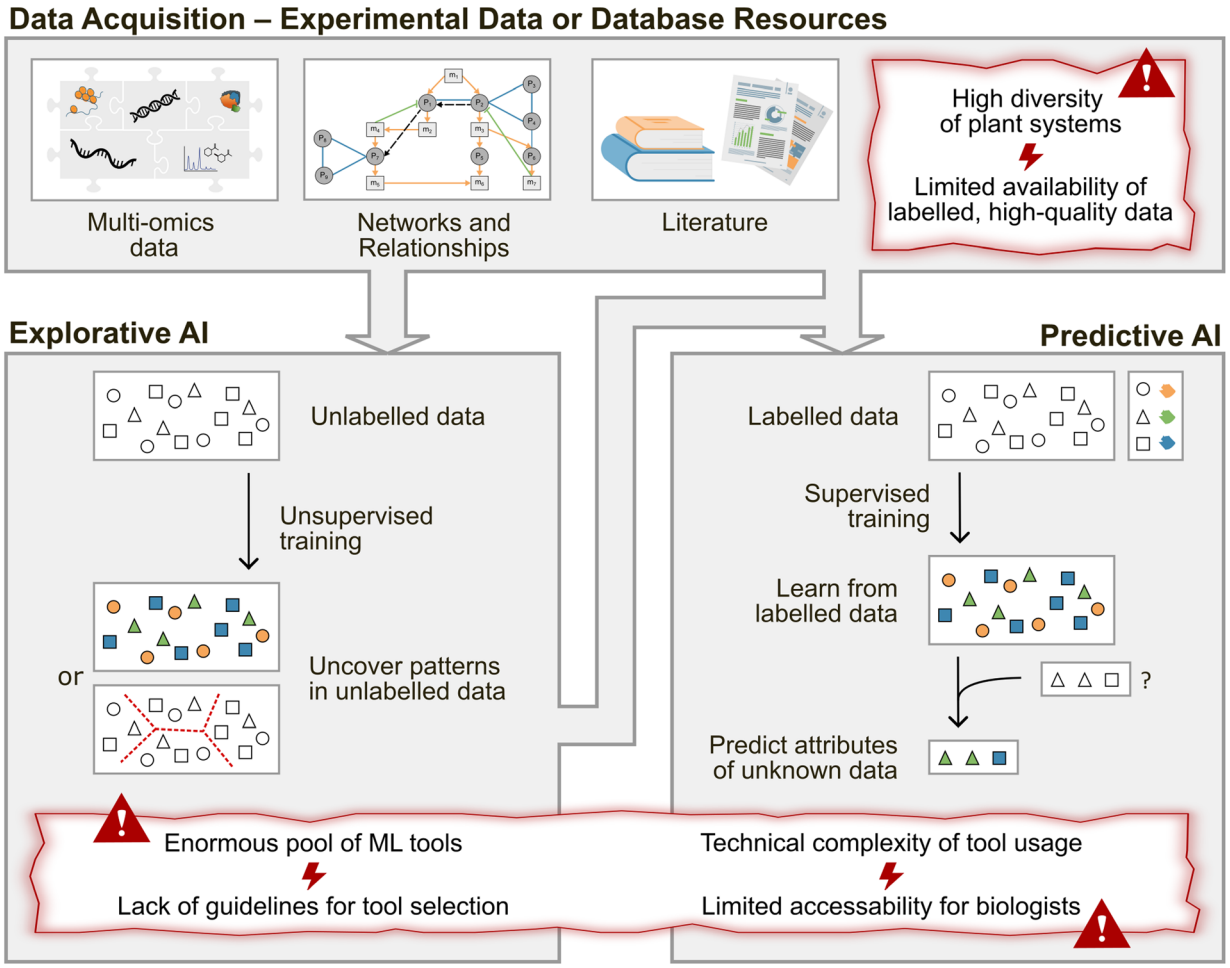


Fig. 6 Distinction between explorative AI and predictive AI in this review, alongside major challenges impeding the use of AI for gene discovery in plant specialised metabolism.

achieve this goal is even rarer. Nevertheless, several studies demonstrate that tools developed in human research can be successfully transferred to plant systems for broader biological questions. For example, MOFA has been used to integrate proteomics and metabolomics data to study host–pathogen interactions in *Quercus ilex*.[212] Likewise, DIABLO has been applied to wheat drought response studies by combining transcriptomic and metabolomic datasets.[213] These applications, although they do not focus on pathway reconstruction, illustrate that ML-based multi-omics tools developed for one kingdom may be sufficiently general to allow application to other biological kingdoms and questions. On the other hand, a recent tool called MEANtools[214] has been developed specifically to explore multi-omics data for plant pathway elucidation using ML. It integrates transcriptomics and metabolomics data based on publicly available knowledge of general biochemical reaction rules and metabolite structures to identify plausible transcript–metabolite relationships associated with a pathway of interest. This approach highlights a potential avenue for plant pathway research, as it enables systematic hypothesis generation for a target metabolic plant pathway directly from integrated multi-omics data.

Beyond numerical datasets, explorative AI approaches can also be applied to textual and semantic information. These models, commonly referred to as language models (LM), learn statistical patterns, context, and relationships from text.[215] They vary in the amount of training data used ("large" *vs.* "small" language models) and in their degree of specialisation for particular biological questions. General purpose LMs such as GPT can support scientific work by facilitating computational tasks, assisting with literature exploration, and easing academic writing.[216,217] More tailored models, on the other hand, may provide higher accuracy for kingdom-specific questions. For example, a LM trained on plant science texts called Plant-LLaMA claims improved performance on plant-related text tasks.[218] Within plant specialised metabolism and trait-based ecology, kingdom-adapted LMs have been used to automatically extract enzyme–product relationships[219,220] and metabolite–taxon or trait–taxon associations, respectively.[221–223] Although language models are not directly applied for identifying new candidate genes, they might be useful to indirectly influence and guide pathway elucidation by generating structured knowledge, automating workflows, curating annotation data, and preparing training data for downstream predictive

models. However, many of the mentioned studies also highlight that model accuracy depends strongly on aspects such as quality of the training data and formulation of prompts, and that careful human curation is essential to ensure reliability.[219,221,222] Furthermore, outputs from general purpose LMs such as GPT are fundamentally constrained by the availability of data and may be affected by non-deterministic behaviour and occasional inconsistencies or hallucinations, highlighting the importance of validation by human experts.[216,217]

5.2. Predictive AI to find biosynthetic enzymes and elucidate pathways

In contrast to explorative approaches, predictive AI focuses on deducing patterns learned from given annotated datasets to make inferences about new or unseen data (Fig. 6). These methods rely primarily on supervised learning, where models are trained on known examples, such as gene–pathway relationships, annotated expression profiles, or biochemical data, and then used to generalise beyond the training set. While the diversity of predictive AI tools for biological research is enormous, we here focus on tools and aspects of particular interest for plant pathway elucidation.

A fundamental application of predictive AI is the functional annotation of genes and proteins. Traditionally, functions are inferred either from sequence homology to annotated genes or proteins (*e.g.*, BLAST,[224] OrthoFinder[225]) or from the presence of motifs and domains associated with specific activities (*e.g.*, HMMER profiles, InterPro annotation[226]). Reliable functional annotations are crucial for pathway elucidation, especially when candidate selection focuses on common multi-gene families such as cytochrome P450 monooxygenases. However, these traditional approaches face clear limitations when sequence similarity is low or when functions are insufficiently represented in existing databases. ML approaches have been proposed to address these gaps by learning features beyond obvious similarity and by identifying hidden relationships which alignment-based methods may miss.[227,228] ML approaches could be a solution to condense the enormous number of characterised sequences and the associated knowledge into compact models that facilitate functional annotation. Over the past years, numerous ML tools and architectures have emerged aiming to improve protein classification and functional annotation,[227,228] such as DeepFam,[229] ProtCNN,[230] or DeepGOPlus.[231] Functional annotation with ML can also incorporate structural information in addition to sequence data, potentially increasing prediction capability (*e.g.*, ProSST[232]). Understanding the relevance of individual amino acids in protein structures could be a major step forward for more accurate predictions. Despite these methodological novelties, traditional homology- and domain-based strategies continue to dominate in plant science, and the adoption of ML-based function prediction remains limited. When applied, these tools still primarily complement rather than replace conventional annotation methods.[233,234]

Much more widely adopted is the application of ML tools for predicting protein features such as subcellular localisation, membrane association, and three-dimensional structure (Fig. 7). For membrane association, a well-known ML tool is DeepTMHMM, which has been reported to outperform its non-ML predecessor TMHMM.[235] For subcellular localisation, several ML-based tools are available, including SignalP,[236] DeepLoc,[237] and MULocDeep.[238] Several studies in the field of plant specialised metabolism illustrate the integration of these tools into the workflow for *in silico* protein characterisation.[239–241] Additionally, the quality of predictions of three-dimensional structures of proteins has seen a dramatic improvement in recent years. Prominent tools include AlphaFold3 (AF3),[242] ESMFold,[243] and RoseTTAFold,[244] which have frequently been applied in plant pathway studies for structural predictions and visualisation of candidate proteins.[245–247] However, protein features predicted by ML tools are typically not used as a primary criterion for candidate selection; rather, they serve to support hypotheses, for example by mapping protein–substrate co-localisation or comparing structural features between a candidate and a known homologue. Therefore, they remain a complementary approach in pathway research.

ML approaches are not restricted to predicting features of individual proteins in isolation; they can also be used to explore potential protein–substrate and protein–protein interactions *in silico* (Fig. 7). Many ML tools have been developed to structurally predict binding poses between an enzyme and a ligand of interest – commonly referred to as molecular docking – such as DiffDock,[248] GNINA,[249] or Interformer.[250] Standard structure prediction tools, including AlphaFold3,[242] can also contribute to this task. However, in plant specialised metabolism, traditional physics-based docking methods such as AutoDock Vina[251] and SwissDock[252] still dominate.[25,253,254] When applied to pathway elucidation, the primary role of docking is to support

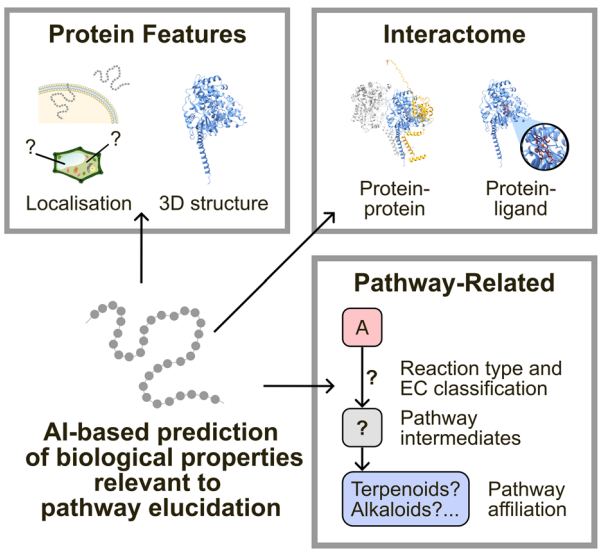


Fig. 7 Common AI applications to predict biological properties relevant for elucidation of biosynthetic pathways. The icon 'emptycell-membrane-oval' by Servier (CC-BY 3.0, cropped) from BioIcons was used.

hypothesis generation and provide mechanistic insights rather than to perform candidate selection. On the other hand, docking is also widely used in drug discovery to explore inhibitory interactions.[255,256] This reflects an important methodological constraint: docking – whether ML-based or physics-based – predicts binding rather than catalysis.[257] Therefore, docking may not reliably distinguish whether a ligand will undergo catalytic turnover or merely bind as an inhibitor, a distinction that often requires additional mechanistic or biochemical evidence. Developing new tools that can not only predict binding but also catalysis is therefore crucial to overcome this limitation. The physics-based docking tool EnzyDock is a promising step in this direction.[258,259] It performs multistate docking along a known reaction pathway, predicting catalytically plausible orientations of given substrates, intermediates, transition states, and products. While this approach might help to flag ligands that are unlikely to adopt catalytically competent poses to distinguish inhibitors from true substrates, it still requires prior knowledge of the reaction states and does not guarantee actual turnover. Therefore, future improvements are still urgently needed to develop a universal and reliable prediction tool for enzyme substrates.

More tailored ML approaches aim specifically to predict whether an enzyme can recognise a substrate of interest, for example tools such as DeepMolecules[260] or MPEK.[261] ML predictors focused on specific enzyme families have also been developed, such as DeepP450 and[262] CYPstrate[263] for cytochrome P450 monooxygenases. However, the applicability of these tools depends heavily on the composition of their training sets, and reported examples of their practical use – especially in plant biosynthesis research – remain limited. One notable application combined AlphaFold3 structure prediction, docking with SwissDock, and ML-based interaction scoring with MPEK to identify UDP-glycosyltransferases potentially involved in ginsenoside biosynthesis.[264] Although experimental validation is still pending, this study illustrates how ML-based tools might be integrated into a useful candidate selection pipeline.

A related avenue focuses on predicting protein–protein interactions. As described in Section 4, these are central for coordinating enzymatic pathways and regulating plant metabolism and therefore represent an enormous opportunity for pathway elucidation when recognised. Tools such as AlphaFold3[242] and AlphaFold-Multimer[265] provide structural predictions of protein complexes. In addition, plant-specific predictors exist – including DWPPI,[266] PlantPathoPPI,[267] DeepAraPPI,[268] and ESMAraPPI[269] – which are trained either for *Arabidopsis* specifically or for broader plant applications. Several studies illustrate the potential of these tools: AlphaFold-Multimer has been used to model a ternary complex in the soybean isoflavonoid pathway, supporting hypotheses of metabolon-like assemblies.[270] In plant pathogen research, AlphaFold-Multimer enabled large-scale *in silico* screening of thousands of protein pairs to predict cross-kingdom interactions between pathogen-secreted proteins and host hydrolases, with several novel interactions experimentally validated.[271] Similarly, high-throughput AF3 pipelines applied to microbial biosynthetic gene clusters predicted tens of thousands of interactions, including previously uncharacterised proteins, demonstrating the potential of predicting protein–protein interactions with ML to uncover hidden functional relationships.[272]

Besides the prediction of protein–protein interactions, there are further ML approaches which do not focus on individual catalytic steps but rather take a broader perspective, aiming to elucidate entire pathways rather than single reactions (Fig. 7). ML-guided tools such as READRetro[273] or DeepMol[274] predict likely precursors and pathway intermediates for biosynthetic pathways, offering guidance regarding the sequence and type of reactions – and therefore the classes of enzymes – potentially involved in a pathway. Other strategies differentiate genes and proteins based on their affiliation with primary or specialised metabolism[275] or directly assign candidates to specialised pathways by integrating features from multiple omics layers[276] or by using sequence-derived descriptors.[277]

Despite this rapidly expanding range of predictive tools based on ML and AI approaches, practical applications in plant specialised metabolism that offer a substantial added value over traditional approaches remain scarce. This gap between computational potential and experimental reality highlights the challenges of translating predictions into actionable insights, which will be discussed in detail in the following subsection.

5.3. Overcoming challenges to unlock the potential of AI in plant sciences

Despite the rapidly growing landscape of machine learning tools in computational biology, applying such tools to plant pathway elucidation remains a multifaceted challenge. First, one of the main hurdles is simply navigating through the extensive repertoire of available ML tools. For nearly every type of prediction – localisation, functional annotation, structure, or interaction – multiple ML tools exist; some of these have a broad scope, while some are tailored to a specific subproblem. However, high-quality pioneer work showcasing successful application of these tools is very often still lacking, especially in specific areas such as plant pathway elucidation. Furthermore, systematic comparisons or clear guidelines for selecting the right tool for a given scientific problem are largely missing. Even when such in-depth methodological reviews exist, they become outdated quickly due to the rapid pace of new developments and are often not focussed on plant research.[205,227] Moreover, many models were trained on datasets with limited or no plant representation, raising concerns regarding their reliability when applied to plant research. In one study that specifically evaluated the transferability of a model trained on human proteins to a plant dataset, the results showed poor performance, highlighting the need for tools specifically developed for plant research.[262]

A related constraint is the enormous diversity of plant omics data, which is of limited suitability as a comprehensive training dataset for many ML applications in the field of plant specialised metabolism. Unlike human or microbial systems, from which many ML tools originate, plant specialised metabolism does not revolve around a single reference species or small

genomes, respectively, but spans more than 370 000 species.[278] Less than 3000 plant genomes have been sequenced, and these are dominated by a few families rich in crops and medicinal species.[7] Additionally, many pathways involved in specialised metabolism in plants still lack experimental characterisation. This means that the data available to train effective and precise ML models tailored to biosynthetic pathways in plants are not only skewed but also insufficient in size. Even for well-studied pathways and plant groups, for which relevant data exist, they are typically scattered across numerous databases, each with different formats, quality standards, and accessibility. Consequently, obtaining comprehensive, reliable, and high-quality plant-specific training data remains a challenging and labour-intensive process.

A further limitation concerns practical usability. Many ML tools are not available as user-friendly web applications, or if they are, they may be restricted in functionality, throughput, or input size. Alternatively, running ML tools locally is an option, which, however, introduces several complications. Primarily, accessibility of ML tools is linked to clear, complete and publicly available documentation, but this is not always guaranteed. Furthermore, installation and usage often require substantial technical skills, including managing package dependencies, using command-line interfaces, and performing manual pre-processing of input data. This is exemplified when installing and running the latest version of AlphaFold 3 locally, which only supports Linux and thus relies on command-line usage. As a result, a significant gap persists between developers of computational tools and actual potential end users in plant science. Bridging this divide will require continued efforts in interdisciplinary collaboration, training in bioinformatics, and the willingness to explore and invest time in unfamiliar computational methods. It also should be kept in mind that the hardware requirements for many deep learning-based tools can be substantial, which becomes more obvious when tools are run locally. For AlphaFold 3, for example, a high-end GPU, at least 64 GB of RAM, and up to 1 TB of storage to accommodate genetic databases are required. These hardware requirements not only come with high investment costs but also entail high energy consumption – particularly when performing complex predictions at large scale – which represents an additional financial factor and has considerable environmental implications, both when running the tools locally and on cloud-based servers.

Nevertheless, there is cause for optimism. With the continuous evolution of technologies, we can expect a gradual bridging of the knowledge and data gaps in plant specialised metabolism. Efforts to standardise data formats will further help to improve the re-usability of omics data, particularly metabolomics, for ML training. For example, the adoption of the FAIR principles – which aim to make data Findable, Accessible, Interoperable, and Reusable – and platforms such as DataPLANT's PLANTdataHUB provide concrete guidance and infrastructure for improving the sharing of plant omics datasets.[279,280] Over time, ML tools should therefore become more accurate, more accessible, and more widely applied. Until then, ML already plays a crucial role by helping researchers navigate the expanding scale of plant-related data and by facilitating the integration of complex computational tools.

6. Critical discussion and conclusions

All emerging technologies presented in this review exhibit different strengths, limitations, and practical impact (Table 1). Overall, co-expression-based methods will remain a major strategy for discovering biosynthetic genes in plants, as their power can be substantially enhanced by integrating additional omics layers or by improving the resolution of data analysis. Aligning insights from multi-omics data at a conceptual level to formulate a shared hypothesis is already common in plant pathway research, but true statistical integration of diverse omics datasets remains challenging and rare. The growing availability of high-quality public omics datasets, along with bioinformatics tools for integrated analysis and visualisation, will likely facilitate more effective multi-omics approaches in the future.

Single-cell transcriptomics holds great potential for improving co-expression analyses. Although still relatively expensive and technically demanding, this approach has already demonstrated utility as an emerging strategy for successful pathway elucidation in cases where traditional co-expression analyses have faced major barriers. Therefore, single-cell transcriptomics will likely become one of the most powerful strategies for pathway elucidation in plants.

Cross-species co-expression analyses, or phylotranscriptomics, offer an intriguing avenue for studying conserved pathways. While this approach currently lacks robust case studies, it has shown promising results in related fields, and new tools may increase the applicability of this strategy in the future.

Currently, one of the most powerful complements to co-expression analysis comes from genomic data, particularly as high-quality plant genomes become increasingly available. Function-centric approaches – which search for biosynthetic gene clusters (BGCs) based on gene families common for specialised metabolism – are highly powerful but are only applicable if at least a subset of pathway genes is physically clustered. This strategy has proven successful for a diverse set of pathways. Synteny comparison, either as an alternative or in combination with function-centric analysis, can also identify cryptic or dispersed pathway genes with less dependence on known gene functions. Plant-adapted tools for BGC discovery are becoming more widely available and established, though tools for synteny analysis are still limited.

Beyond gene clusters, GWAS provide a complementary strategy to link genomic variation with specific phenotypes, such as the ability to produce specialised metabolites. However, GWAS not only require detailed population data and high-quality datasets, but execution and analyses are also computationally intensive, limiting widespread adoption. Nonetheless, GWAS can offer valuable complementary insights, particularly when pathway genes are not physically clustered. Similarly, epigenomics approaches are technically complex and data

Table 1 Critical comparison of emerging technologies for the discovery of biosynthetic genes in plants

Method	Strengths	Limitations	Practical impact
Improved transcriptomic and multi-omics approaches			
Multi-omics	Offers multi-layer perspective; insights at system level	Data integration can be complex	Increasingly used; tools tailored to plants increasingly available
Single-cell transcriptomics	High resolution; uncovers correlations hidden in bulk data	Sample preparation and data analysis challenging; expensive	Emerging and powerful strategy; strong case studies
Phylo-transcriptomics	Insights into evolution and cross-species pathway conservation	Restricted to conserved pathways; normalisation challenging	Little used; studies from other fields suggest potential
Genomic and population strategies			
Genome mining for BGCs	Direct detection of co-localised biosynthetic genes	Genes must be clustered; requires high-quality genome sequence	Established and increasingly used; strong case studies
Synteny analysis	Discovery of gene associations beyond canonical BGCs; insights into evolution	Highly dependent on high-quality genome sequences and conserved synteny	Not yet widely used; could offer complementary evidence for BGC mining
GWAS	Direct genotype–phenotype association in populations	High computational demand; large population datasets required	Potentially strong complementary strategy; mainly used for crop plants, but limited general use
Epigenomics	Insights into regulatory state beyond DNA sequence	Very data intensive; provides only indirect link to pathways	Limited for pathway elucidation; mostly regulation studies
Mining enzyme families	Explores full sequence diversity of an enzyme family	Restricted to certain enzyme families; large candidate space	Still rarely used; high potential for specific enzyme families
Protein-based strategies			
Protein-level correlation	Revealing further correlations compared to expression data	Requires bait protein; higher experimental effort	Niche; can provide complementary data
Chemo-proteomics	No prior knowledge about target enzymes required	Complex experimental design; not suitable for all enzyme classes; requires synthesis expertise	So far very rarely used; very high potential as complementary approach
Protein–protein interactions	Additionally reveals functional enzyme networks	Limited throughput; complex experimental setup	Rarely used for enzyme discovery; very high potential
AI-based strategies			
Explorative AI	Detects hidden patterns across large datasets	Very data intensive; application complex	Still emerging; few successful case studies available (*e.g.*, SOMs)
Predictive AI	Predicting cryptic properties better than human experts and non-AI tools	Very data intensive, yet limited plant pathway training data; application complex; tool selection	Still emerging; mainly used in combination with conventional approaches; high potential

intensive. So far, they have been primarily applied to gene regulation rather than direct pathway gene identification.

Another strategy is enzyme family mining, which involves screening entire enzyme families for candidate genes. This approach is independent of gene co-expression, co-regulation, or co-localisation. However, it is so far only applicable to enzyme families for which biochemical characterisation is straightforward, for example due to a single shared substrate. Furthermore, additional filtering strategies are required to narrow down enzyme numbers to a reasonable level for experimental validation.

Efforts to discover biosynthetic enzymes directly at the protein level remain relatively niche, largely due to the complexity of experimental design, setup, and data analysis. Protein-level correlation analyses can complement gene expression data, particularly when transcript–protein correlations are weak; however, they also require a suitable bait protein and exhibit limitations for low-abundance proteins. Chemoproteomics allows the identification of uncommon enzymes without prior knowledge, providing access to targets not easily captured by other approaches; however, designing and carrying out such experiments is highly challenging and requires suitable interdisciplinary expertise. Protein–protein interaction-based strategies reveal functional biochemical networks and thereby capture relationships that may be invisible from gene expression or protein abundance data alone. However, the throughput of such interaction studies is currently too low for efficient enzyme discovery; in addition, the experiments are also technically demanding. While none of these protein-based methods is likely to replace transcriptomic or genomic approaches in the near future, each offers distinct advantages and can serve as a valuable complement when other strategies reach their limits.

Finally, AI and ML tools hold great promise also for plant specialised metabolism, but neither exploratory nor predictive approaches have yet achieved major breakthroughs in the field. Some exploratory methods, such as self-organising maps, are already more widely used, but their scope and impact remain

limited. Predictive tools rarely function effectively as standalone approaches and are typically applied only in combination or in comparison to conventional methods. Across both categories, progress is constrained particularly by the limited availability of high-quality training datasets specific for plant specialised metabolism. A key challenge, but also the beauty of this field, is that individual biosynthetic pathways are highly diverse; even minor evolutionary events can influence the metabolic outcome of a pathway, while in other cases independently evolved pathways can converge on the same metabolites. This makes it much more challenging to provide generalisable training data for ML models. However, once pioneering studies demonstrate clear, reproducible successes, ML tools are likely to become a more integral part of pathway research.

To better illustrate these methodological developments and improvements over the past decades, efforts to investigate the biosynthesis of monoterpenoid indole alkaloids (MIAs) serve as an exemplary blueprint for the historical progress in the field. Early stages of pathway elucidation relied primarily on forward genetics, often including direct purification of biosynthetic enzymes from the native producers and cDNA library screening, a slow and low-throughput process.[281] This was subsequently accelerated by sequence-based approaches enabled by expressed sequence tags and transcript data, which allowed candidate gene discovery based on homology to known gene families and expression in relevant tissues.[282] With the increasing availability and quality of transcriptomic resources, co-expression analysis became a powerful strategy which enabled the discovery of novel enzymes beyond previously recognised families.[283] Sequencing of plant genomes, such as that of *Catharanthus roseus* in 2015, enabled genome-wide analyses, revealing partial clustering and duplication of MIA biosynthetic genes.[284] However, as these MIA-associated clusters did not represent entire pathways, alternative approaches were still required.[284] Therefore, comprehensive multi-omics and single-cell approaches have recently been employed to resolve the spatial and regulatory complexity of MIA biosynthesis, as demonstrated in studies by Li *et al.*[59] and Stander *et al.*[285] In parallel, proteomics-based identification of enzyme complexes, for example in *Mitragyna speciosa*, has highlighted the applicability of protein–protein interaction screening for pathway discovery.[181] Machine learning approaches based on expression data have also been applied in MIA pathway research to prioritise candidate biosynthetic genes, complementing co-expression-based strategies.[285,286]

In conclusion, these developments in MIA biosynthesis research demonstrate that – despite increasing data availability and methodological sophistication – complete pathway elucidation remains challenging due to factors such as spatial compartmentalisation, non-canonical enzymes, and pathway complexity. This underscores the value of complementary approaches for gene discovery.

Therefore, overall, we consider a combination of enhanced co-expression analyses and genome-based approaches to be the current state-of-the-art strategy for discovering biosynthetic genes in plants; other non-ML methods may serve as valuable complements under suitable circumstances. While machine learning approaches are expected to gradually improve in performance and applicability, non-ML strategies are likely to remain central to the field in the near future.

7. Author contributions

All authors performed literature research, wrote, edited, and approved the final manuscript.

8. Conflicts of interest

There are no conflicts to declare.

9. Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

10. Acknowledgements

JF and BP gratefully acknowledge joint funding by the German Research Foundation (DFG) (FR 3720/7-1 and PU 718/2-1). JF additionally acknowledges financial support by the DFG *via* the Emmy Noether programme (FR 3720/3-1) and the NSERC-DFG SUSTAIN programme (FR 3720/8-1).

11. Notes and references

1 V. Courdavault, S. E. O'Connor, M. K. Jensen and N. Papon, *Nat. Prod. Rep.*, 2021, **38**, 2145–2153, https://pubs.rsc.org/en/content/articlehtml/2021/np/d0np00092b.
2 X. Zhu, X. Liu, T. Liu, Y. Wang, N. Ahmed, Z. Li and H. Jiang, *Plant Commun.*, 2021, **2**, 100229, https://www.cell.com/plant-communications/fulltext/S2590-3462(21)00131-0.
3 E. Kenshole, M. Herisse, M. Michael and S. J. Pidot, *Curr. Opin. Chem. Biol.*, 2021, **60**, 47–54, https://www.sciencedirect.com/science/article/pii/S136759312030106X.
4 K. Scherlach and C. Hertweck, *Nat. Commun.*, 2021, **12**, 3864, https://www.nature.com/articles/s41467-021-24133-5.
5 K. D. Bauman, K. S. Butler, B. S. Moore and J. R. Chekan, *Nat. Prod. Rep.*, 2021, **38**, 2100–2129, https://pubs.rsc.org/en/content/articlehtml/2021/xx/d1np00032b.
6 J. Chen, Y. Zhang, Y. Li, Y. Liu, Q. Li, Z. Lv, M. I. Georgiev and P. Liao, *Engineering*, 2025, DOI: 10.1016/j.eng.2025.09.024.
7 R. Schwacke, M. E. Bolger and B. Usadel, *Front. Plant Sci.*, 2025, **16**, 1603547.
8 L. Xie, X. Gong, K. Yang, Y. Huang, S. Zhang, L. Shen, Y. Sun, D. Wu, C. Ye, Q.-H. Zhu and L. Fan, *Nat. Plants*, 2024, **10**, 551–566, https://www.nature.com/articles/s41477-024-01655-6.
9 M. H. Medema, T. de Rond and B. S. Moore, *Nat. Rev. Genet.*, 2021, **22**, 553–571, https://www.nature.com/articles/s41576-021-00363-7.

10 X. Rao and R. A. Dixon, *Acta Biochim. Biophys. Sin.*, 2019, **51**, 981–988, DOI: 10.1093/abbs/gmz080.
11 A. R. Fernie and T. Tohge, *Annu. Rev. Genet.*, 2017, **51**, 287–310, DOI: 10.1146/annurev-genet-120116-024640.
12 A. Goossens, *Mol. Plant*, 2015, **8**, 2–5, https://www.cell.com/molecular-plant/fulltext/S1674-2052(14)00012-4.
13 L. Chuang, S. Liu and J. Franke, *J. Am. Chem. Soc.*, 2023, **145**, 5083–5091.
14 G. Polturak, M. Dippe, M. J. Stephenson, R. Chandra Misra, C. Owen, R. H. Ramirez-Gonzalez, J. F. Haidoulis, H.-J. Schoonbeek, L. Chartrain, P. Borrill, D. R. Nelson, J. K. M. Brown, P. Nicholson, C. Uauy and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2123299119, DOI: 10.1073/pnas.2123299119.
15 C. J. McClune, J. C.-T. Liu, C. Wick, R. de La Peña, B. M. Lange, P. M. Fordyce and E. S. Sattely, *Nature*, 2025, **643**, 582–592, https://www.nature.com/articles/s41586-025-09090-z.
16 M. Kang, Y. Choi, H. Kim and S.-G. Kim, *New Phytol.*, 2022, **234**, 527–544, DOI: 10.1111/nph.17992.
17 J. Han, E. P. Miller and S. Li, *Curr. Opin. Biotechnol.*, 2024, **87**, 103137.
18 M. McConnachie, T.-A. M. Nguyen, T. Kim, T.-D. Nguyen and T.-T. T. Dang, *Plant J.*, 2025, **122**, e70288.
19 F. C. Wolters, E. Del Pup, K. S. Singh, K. Bouwmeester, M. E. Schranz, J. J. J. van der Hooft and M. H. Medema, *Curr. Opin. Plant Biol.*, 2024, **82**, 102657, https://www.sciencedirect.com/science/article/pii/S1369526624001481.
20 R. Cavill, D. Jennen, J. Kleinjans and J. J. Briedé, *Briefings Bioinf.*, 2016, **17**, 891–901.
21 T. M. Ebbels and R. Cavill, *Prog. Nucl. Magn. Reson. Spectrosc.*, 2009, **55**, 361–374.
22 M. Colinas, C. Morweiser, O. Dittberner, B. Chioca, R. Alam, H. Leucke, Y. Nakamura, D. A. Serna Guerrero, S. Heinicke, M. Kunert, J. Wurlitzer, K. Ploss, B. Hong, V. Grabe, A. A. Lopes and S. E. O'Connor, *Nat. Chem. Biol.*, 2025, **21**, 1794–1805.
23 S. Jo, A. El-Demerdash, C. Owen, V. Srivastava, D. Wu, S. Kikuchi, J. Reed, H. Hodgson, A. Harkess, S. Shu, C. Plott, J. Jenkins, M. Williams, L.-B. Boston, E. Lacchini, T. Qu, A. Goossens, J. Grimwood, J. Schmutz, J. Leebens-Mack and A. Osbourn, *Nat. Chem. Biol.*, 2025, **21**, 215–226, https://www.nature.com/articles/s41589-024-01681-7.
24 R. S. Nett, X. Guan, K. Smith, A. M. Faust, E. S. Sattely and C. R. Fischer, *AIChE J.*, 2018, **64**, 4319–4330, DOI: 10.1002/aic.16413.
25 N. Mehta, Y. Meng, R. Zare, R. Kamenetsky-Goldstein and E. Sattely, *Cell*, 2024, **187**, 5620–5637.e10.
26 S. Ling, Q. Lin, B. Zhou, Y. Liang, W. Luo, Z. Shen, J. Wang, J. Niu, L. Qiao, B. Wang and H. Liu, *Front. Plant Sci.*, 2025, **16**, 1629266, DOI: 10.3389/fpls.2025.1629266.
27 M. Rai, A. Rai, T. Yokosaka, T. Mori, R. Nakabayashi, M. Nakamura, H. Suzuki, K. Saito and M. Yamazaki, *Int. J. Mol. Sci.*, 2025, **26**, 1068.
28 F. Huang, Y. Lei, J. Duan, Y. Kang, Y. Luo, D. Ding, Y. Chen and S. Li, *Sci. Rep.*, 2024, **14**, 10023.
29 Z. Zhu, Y. Zhou, X. Liu, F. Meng, C. Xu and M. Chen, *Plant Biotechnol. J.*, 2025, **23**, 715–730.
30 P. Zhang, S. Xu, L. Zhang, X. Li, J. Qi, L. Weng, S. Cai and J. Wang, *BMC Plant Biol.*, 2025, **25**, 729.
31 J. Wei, X. Mu, S. Wang, Q. Wei, L. Zhu, X. Zhang, J. Zhang, X. Liu, B. Wen, M. Li and J. Liu, *Food Res. Int.*, 2025, **201**, 115542.
32 X. Zhao, W. Ge and Z. Miao, *Sci. Rep.*, 2024, **14**, 8644.
33 D. Zhao, Y. Zhang, H. Ren, Y. Shi, D. Dong, Z. Li, G. Cui, Y. Shen, Z. Mou, E. J. Kennelly, L. Huang, J. Ruan, S. Chen, D. Yu and Y. Cun, *J. Integr. Plant Biol.*, 2023, **65**, 2320–2335.
34 C. N. Hayes, H. Nakahara, A. Ono, M. Tsuge and S. Oka, *Genes*, 2024, **15**, 1551.
35 F. García-Alcalde, F. García-López, J. Dopazo and A. Conesa, *Bioinformatics*, 2011, **27**, 137–139.
36 T. Liu, P. Salguero, M. Petek, C. Martinez-Mira, L. Balzano-Nogueira, Ž. Ramšak, L. McIntyre, K. Gruden, S. Tarazona and A. Conesa, *Nucleic Acids Res.*, 2022, **50**, W551–W559.
37 G. Zhou, J. Ewald and J. Xia, *Nucleic Acids Res.*, 2021, **49**, W476–W482.
38 K. Munk, D. Ilina, L. Ziemba, G. Brader and E. M. Molin, *BMC Bioinf.*, 2024, **25**, 93.
39 J. B. Fontanet-Manzaneque, N. Laibach, I. Herrero-García, V. Coleto-Alcudia, D. Blasco-Escámez, C. Zhang, L. Orduña, S. Alseekh, S. Miller, N. Bjarnholt, A. R. Fernie, J. T. Matus and A. I. Caño-Delgado, *Plant Biotechnol. J.*, 2024, **22**, 3406–3423.
40 D. Vincent, P. Reddy and D. Isenegger, *Biomolecules*, 2024, **14**, 414.
41 V. Berková, M. Berka, M. Griga, R. Kopecká, M. Prokopová, M. Luklová, J. Horáček, I. Smýkalová, P. Čičmanec, J. Novák, B. Brzobohatý and M. Černý, *Plants*, 2022, **11**, 2931.
42 E. Sybilska, B. S. Haddadi, L. A. J. Mur, M. Beckmann, S. Hryhorowicz, J. Suszynska-Zajczyk, M. Knaur, A. Pławski and A. Daszkowska-Golec, *BMC Plant Biol.*, 2025, **25**, 619.
43 X. Pu, H.-C. Gao, M.-J. Wang, J.-H. Zhang, J.-H. Shan, M.-H. Chen, L. Zhang, H.-G. Wang, A.-X. Wen, Y.-G. Luo and Q.-M. Huang, *Front. Plant Sci.*, 2022, **13**, 851077, DOI: 10.3389/fpls.2022.851077.
44 A. Muchlinski, M. Jia, K. Tiedge, J. S. Fell, K. A. Pelot, L. Chew, D. Davisson, Y. Chen, J. Siegel, J. T. Lovell and P. Zerbe, *Plant J.*, 2021, **108**, 1053–1068, DOI: 10.1111/tpj.15492.
45 J. E. Jeon, J.-G. Kim, C. R. Fischer, N. Mehta, C. Dufour-Schroif, K. Wemmer, M. B. Mudgett and E. Sattely, *Cell*, 2020, **180**, 176–187.e19, https://www.cell.com/cell/fulltext/S0092-8674(19)31322-4.
46 Z. Wang, Q. Yun, J. Hu, Z. Wei, D. Feng, N. Li, H. Xu, L. Fu, Z. Wang, S. Li, F. Liu, Y. Wang, B. Cong and B. Wang, *Sci. Rep.*, 2025, **15**, 27406, https://www.nature.com/articles/s41598-025-11942-7.
47 V. Tzin, N. Fernandez-Pozo, A. Richter, E. A. Schmelz, M. Schoettner, M. Schäfer, K. R. Ahern, L. N. Meihls, H. Kaur, A. Huffaker, N. Mori, J. Degenhardt, L. A. Mueller and G. Jander, *Plant Physiol.*, 2015, **169**,

1727–1743, https://academic.oup.com/plphys/article/169/3/1727/6113904?login=true.
48 O. Calderini, M. O. Kamileen, Y. Nakamura, S. Heinicke, R. M. Alam, B. Hong, Y. Jiang, A. Gutiérrez-Vences, F. Alagna, F. Paolocci, M. C. Valeri, E. Franco, S. Mousavi, R. Mariotti, L. Caputi, S. E. O'Connor and C. E. Rodríguez-López, *Plant Commun.*, 2026, 101713, https://www.cell.com/plant-communications/fulltext/S2590-3462(26)00021-0.
49 J. Fang and B. Schneider, *Phytochem. Anal.*, 2014, **25**, 307–313, DOI: 10.1002/pca.2477.
50 E. Abbott, D. Hall, B. Hamberger and J. Bohlmann, *BMC Plant Biol.*, 2010, **10**, 106, DOI: 10.1186/1471-2229-10-106.
51 T. Nelson, S. L. Tausta, N. Gandotra and T. Liu, *Annu. Rev. Plant Biol.*, 2006, **57**, 181–201, DOI: 10.1146/annurev.arplant.56.032604.144138.
52 P. J. Horn and K. D. Chapman, *J. Exp. Bot.*, 2024, **75**, 1654–1670, https://academic.oup.com/jxb/article/75/6/1654/7331171?login=true.
53 K. M. Frick, M. D. B. B. Lorensen, N. Micic, E. Esteban, A. Pasha, A. Schulz, N. J. Provart, H. H. Nour-Eldin, N. Bjarnholt, C. Janfelt and F. Geu-Flores, *New Phytol.*, 2025, **245**, 2052–2068, DOI: 10.1111/nph.20384.
54 N. Ntelkis, C. R. Buell and A. Goossens, *Plant Physiol.*, 2025, **199**, kiaf375, https://academic.oup.com/plphys/article/199/1/kiaf375/8245308?login=true.
55 M. S. Rhaman, M. Ali, W. Ye and B. Li, *Genom. Proteom. Bioinform.*, 2024, **22**, qzae026, https://academic.oup.com/gpb/article/22/2/qzae026/7630181?login=true.
56 Y. Sun, J. Sun, C. Lin, J. Zhang, H. Yan, Z. Guan and C. Zhang, *Cells*, 2024, **13**, https://www.mdpi.com/2073-4409/13/18/1561.
57 S. Wu, A. L. M. Morotti, J. Yang, E. Wang and E. C. Tatsis, *Mol. Plant*, 2024, **17**, 1439–1457, https://linkinghub.elsevier.com/retrieve/pii/S1674205224002582.
58 S. Choung, G. Kang, T. Kim, S. Kim, H. Yun, Y. Hwang, H. Kim, H. Lim, K. Moon, S. Han and S.-G. Kim, *Nat. Commun.*, 2026, **17**, 1954, https://www.nature.com/articles/s41467-026-68816-3#peer-review.
59 C. Li, J. C. Wood, A. H. Vu, J. P. Hamilton, C. E. Rodriguez Lopez, R. M. E. Payne, D. A. Serna Guerrero, K. Gase, K. Yamamoto, B. Vaillancourt, L. Caputi, S. E. O'Connor and C. Robin Buell, *Nat. Chem. Biol.*, 2023, **19**, 1031–1041, https://www.nature.com/articles/s41589-023-01327-0.
60 V.-H. Bui, J. C. Wood, B. Vaillancourt, J. P. Hamilton, L. H. Carlton, T.-T. T. Dang, C. R. Buell and C. Li, *Plant Biotechnol. J.*, 2026, **24**(2), 918–920, DOI: 10.1111/pbi.70386.
61 S. Sun, X. Shen, Y. Li, Y. Li, S. Wang, R. Li, H. Zhang, G. Shen, B. Guo, J. Wei, J. Xu, B. St-Pierre, S. Chen and C. Sun, *Nat. Plants*, 2023, **9**, 179–190, https://www.nature.com/articles/s41477-022-01291-y.
62 M. Kang, A. H. Vu, A. L. Casper, R. Kim, J. Wurlitzer, S. Heinicke, A. Yeroslaviz, L. Caputi and S. E. O'Connor, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2512828122.
63 D. Kotliar, A. Veres, M. A. Nagy, S. Tabrizi, E. Hodis, D. A. Melton and P. C. Sabeti, *eLife*, 2019, https://elifesciences.org/articles/43803.
64 D. Zheng, X. Lu, Y. Lu, P. Liang, N. Shang, J. Xu, J. Yao, F. Mo, Q. Chu, L. Fan and H. Chen, *Mol. Plant*, 2026, **19**, 673–688, https://www.cell.com/molecular-plant/fulltext/S1674-2052(25)00456-3.
65 J. Yao, A. P. Marand, Y. Bai, R. J. Schmitz and L. Fan, *Trends Plant Sci.*, 2025, https://www.cell.com/trends/plant-science/fulltext/S1360-1385(25)00287-0.
66 S. Giacomello, *Curr. Opin. Plant Biol.*, 2021, **60**, 102041, https://www.sciencedirect.com/science/article/pii/S1369526621000418.
67 K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang and X. Zhuang, *Science*, 2015, **348**, aaa6090, DOI: 10.1126/science.aaa6090.
68 T. A. Lee, N. Illouz-Eliaz, T. Nobori, J. Xu, B. Jow, J. R. Nery and J. R. Ecker, *Nat. Plants*, 2025, **11**, 1960–1975, https://www.nature.com/articles/s41477-025-02072-z.
69 U. K. Reddy, K. S. Karnatam, A. Talavera-Caro, C. Lopez-Ortiz, K.-M. Ku, S. R. Chinreddy, S. Ramireddy, P. Natarajan, V. Kumar, S. S. Kadiyala, P. Somagattu, R. Duhan, N. Balagurusamy, V. A. Benedito, D. A. Adjeroh and P. Nimmakayala, *Hortic. Res.*, 2025, **12**, uhaf243, https://academic.oup.com/hr/article/12/12/uhaf243/8254149?guestAccessKey=.
70 S. Cheon, J. Zhang and C. Park, *Mol. Biol. Evol.*, 2020, **37**, 3672–3683, https://academic.oup.com/mbe/article/37/12/3672/5870839.
71 K. Jiang, C. Du, L. Huang, J. Luo, T. Liu and S. Huang, *Front. Plant Sci.*, 2023, **14**, 1114579.
72 X.-X. Wang, C.-H. Huang, D. F. Morales-Briones, X.-Y. Wang, Y. Hu, N. Zhang, P.-G. Zhao, X.-M. Wei, K.-H. Wei, X. Hemu, N.-H. Tan, Q.-F. Wang and L.-Y. Chen, *Nat. Commun.*, 2024, **15**, 9663.
73 S. Gupta, R. Singh, P. Paul, S. Kaul, S. K. Lattoo and M. K. Dhar, *Plant Biotechnol. Rep.*, 2024, **18**, 75–89.
74 E. Kariñho Betancourt, N. Calderón Cortés, R. Tapia López, I. De-la-Cruz, J. Núñez Farfán and K. Oyama, *Evol. Ecol.*, 2024, **14**, e11496.
75 K.-K. Yan, D. Wang, J. Rozowsky, H. Zheng, C. Cheng and M. Gerstein, *Genome Biol.*, 2014, **15**, R100.
76 J. Lee, L. S. Heath, R. Grene and S. Li, *Plant Methods*, 2019, **15**, 61.
77 M. J. Passalacqua and J. Gillis, *Nat. Plants*, 2024, **10**, 1075–1080.
78 N. Grünig and B. Pucker, *BMC Genom.*, 2025, **26**, 807.
79 J. A. V. S. de Oliveira, N. Choudhary, S. N. Meckoni, M. S. Nowak, M. Hagedorn and B. Pucker, *BMC Genom.*, 2026, **27**, 231, DOI: 10.1186/s12864-026-12623-z.
80 S. J. Smit and B. R. Lichman, *Nat. Prod. Rep.*, 2022, **39**, 1465–1482, https://pubs.rsc.org/en/content/articlehtml/2022/np/d2np00005a.
81 M. Frey, P. Chomet, E. Glawischnig, C. Stettner, S. Grün, A. Winklmair, W. Eisenreich, A. Bacher, R. B. Meeley, S. P. Briggs, K. Simcox and A. Gierl, *Science*, 1997, **277**, 696–699.
82 X. Qiao, A. Houghton, J. Reed, B. Steuernagel, J. Zhang, C. Owen, A. Leveau, A. Orme, T. Louveau, R. Melton,

B. B. H. Wulff and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2417588122.

83 M. M. Zdouc, K. Blin, N. L. L. Louwen, J. Navarro, C. Loureiro, C. D. Bader, C. B. Bailey, L. Barra, T. J. Booth, K. A. J. Bozhüyük, J. D. D. Cediel-Becerra, Z. Charlop-Powers, M. G. Chevrette, Y. H. Chooi, P. M. D'Agostino, T. de Rond, E. Del Pup, K. R. Duncan, W. Gu, N. Hanif, E. J. N. Helfrich, M. Jenner, Y. Katsuyama, A. Korenskaia, D. Krug, V. Libis, G. A. Lund, S. Mantri, K. D. Morgan, C. Owen, C.-S. Phan, B. Philmus, Z. L. Reitz, S. L. Robinson, K. S. Singh, R. Teufel, Y. Tong, F. Tugizimana, D. Ulanova, J. M. Winter, C. Aguilar, D. Y. Akiyama, S. A. A. Al-Salihi, M. Alanjary, F. Alberti, G. Aleti, S. A. Alharthi, M. Y. A. Rojo, A. A. Arishi, H. E. Augustijn, N. E. Avalon, J. A. Avelar-Rivas, K. K. Axt, H. B. Barbieri, J. C. J. Barbosa, L. G. Barboza Segato, S. E. Barrett, M. Baunach, C. Beemelmanns, D. Beqaj, T. Berger, J. Bernaldo-Agüero, S. M. Bettenbühl, V. A. Bielinski, F. Biermann, R. M. Borges, R. Borriss, M. Breitenbach, K. M. Bretscher, M. W. Brigham, L. Buedenbender, B. W. Bulcock, C. Cano-Prieto, J. Capela, V. J. Carrion, R. S. Carter, R. Castelo-Branco, G. Castro-Falcón, F. O. Chagas, E. Charria-Girón, A. A. Chaudhri, V. Chaudhry, H. Choi, Y. Choi, R. Choupannejad, J. Chromy, M. S. C. Donahey, J. Collemare, J. A. Connolly, K. E. Creamer, M. Crüsemann, A. A. Cruz, A. Cumsille, J.-F. Dallery, L. C. Damas-Ramos, T. Damiani, M. de Kruijff, B. D. Martín, G. Della Sala, J. Dillen, D. T. Doering, S. R. Dommaraju, S. Durusu, S. Egbert, M. Ellerhorst, B. Faussurier, A. Fetter, M. Feuermann, D. P. Fewer, J. Foldi, A. Frediansyah, E. A. Garza, A. Gavriilidou, A. Gentile, J. Gerke, H. Gerstmans, J. P. Gomez-Escribano, L. A. González-Salazar, N. E. Grayson, C. Greco, J. E. G. Gomez, S. Guerra, S. G. Flores, A. Gurevich, K. Gutiérrez-García, L. Hart, K. Haslinger, B. He, T. Hebra, J. L. Hemmann, H. Hindra, L. Höing, D. C. Holland, J. E. Holme, T. Horch, P. Hrab, J. Hu, T.-H. Huynh, J.-Y. Hwang, R. Iacovelli, D. Iftime, M. Iorio, S. Jayachandran, E. Jeong, J. Jing, J. J. Jung, Y. Kakumu, E. Kalkreuter, K. B. Kang, S. Kang, W. Kim, G. J. Kim, H. Kim, H. U. Kim, M. Klapper, R. A. Koetsier, C. Kollten, Á. T. Kovács, Y. Kriukova, N. Kubach, A. M. Kunjapur, A. K. Kushnareva, A. Kust, J. Lamber, M. Larralde, N. J. Larsen, A. P. Launay, N.-T.-H. Le, S. Lebeer, B. T. Lee, K. Lee, K. L. Lev, S.-M. Li, Y.-X. Li, C. Licona-Cassani, A. Lien, J. Liu, J. A. V. Lopez, N. V. Machushynets, M. I. Macias, T. Mahmud, M. Maleckis, A. M. Martinez-Martinez, Y. Mast, M. F. Maximo, C. M. McBride, R. M. McLellan, K. M. Bhatt, C. Melkonian, A. Merrild, M. Metsä-Ketelä, D. A. Mitchell, A. V. Müller, G.-S. Nguyen, H. T. Nguyen, T. H. J. Niedermeyer, J. H. O'Hare, A. Ossowicki, B. O. Ostash, H. Otani, L. Padva, S. Paliyal, X. Pan, M. Panghal, D. S. Parade, J. Park, J. Parra, M. P. Rubio, H. T. Pham, S. J. Pidot, J. Piel, B. Pourmohsenin, M. Rakhmanov, S. Ramesh, M. H. Rasmussen, A. Rego, R. Reher, A. J. Rice, A. Rigolet, A. Romero-Otero, L. R. Rosas-Becerra, P. Y. Rosiles, A. Rutz, B. Ryu, L.-A. Sahadeo, M. Saldanha, L. Salvi, E. Sánchez-Carvajal, C. Santos-Medellin, N. Sbaraini, S. M. Schoellhorn, C. Schumm, L. Sehnal, N. Selem, A. D. Shah, T. K. Shishido, S. Sieber, V. Silviani, G. Singh, H. Singh, N. Sokolova, E. C. Sonnenschein, M. Sosio, S. T. Sowa, K. Steffen, E. Stegmann, A. B. Streiff, A. Strüder, F. Surup, T. Svenningsen, D. Sweeney, J. Szenei, A. Tagirdzhanov, B. Tan, M. J. Tarnowski, B. R. Terlouw, T. Rey, N. U. Thome, L. R. Torres Ortega, T. Tørring, M. Trindade, A. W. Truman, M. Tvilum, D. W. Udwary, C. Ulbricht, L. Vader, G. P. van Wezel, M. Walmsley, R. Warnasinghe, H. G. Weddeling, A. N. M. Weir, K. Williams, S. E. Williams, T. E. Witte, S. M. W. Rocca, K. Yamada, D. Yang, D. Yang, J. Yu, Z. Zhou, N. Ziemert, L. Zimmer, A. Zimmermann, C. Zimmermann, J. J. J. van der Hooft, R. G. Linington, T. Weber and M. H. Medema, *Nucleic Acids Res.*, 2025, **53**, D678–D690, https://academic.oup.com/nar/article/53/D1/D678/7919508.

84 G. Polturak and A. Osbourn, *PLoS Pathog.*, 2021, **17**, e1009698, DOI: 10.1371/journal.ppat.1009698.

85 H.-W. Nützmann and A. Osbourn, *Curr. Opin. Biotechnol.*, 2014, **26**, 91–99, https://www.sciencedirect.com/science/article/pii/S0958166913006836.

86 N. Panchy, M. Lehti-Shiu and S. Shiu, *Plant Physiol.*, 2016, **171**, 2294–2316, https://academic.oup.com/plphys/article/171/4/2294/6115338?login=true.

87 H.-W. Nützmann, A. Huang and A. Osbourn, *New Phytol.*, 2016, **211**, 771–789, DOI: 10.1111/nph.13981.

88 C. Zhan, S. Shen, C. Yang, Z. Liu, A. R. Fernie, I. A. Graham and J. Luo, *Trends Plant Sci.*, 2022, **27**, 981–1001, https://www.cell.com/trends/plant-science/fulltext/S1360-1385(22)00055-3.

89 H.-W. Nützmann, D. Doerr, A. Ramírez-Colmenero, J. E. Sotelo-Fonseca, E. Wegel, M. Di Stefano, S. W. Wingett, P. Fraser, L. Hurst, S. L. Fernandez-Valverde and A. Osbourn, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 13800–13809, DOI: 10.1073/pnas.1920474117.

90 S. E. Hakim, N. Choudhary, K. Malhotra, J. Peng, A. Bültemeier, A. Arafa, R. Friedhoff, M. Bauer, J. Eikenberg, C.-P. Witte, M. Herde, P. Heretsch, B. Pucker and J. Franke, *Nat. Commun.*, 2025, **16**, 6367, https://www.nature.com/articles/s41467-025-61686-1.

91 W. Ji, A. Osbourn and Z. Liu, *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, 2024, **379**, 20230359.

92 K. A. E. Larsson, I. Zetterlund, G. Delp and L. M. V. Jonsson, *Phytochemistry*, 2006, **67**, 2002–2008, https://www.sciencedirect.com/science/article/pii/S0031942206003852.

93 S. L. Dias, L. Chuang, S. Liu, B. Seligmann, F. L. Brendel, B. G. Chavez, R. E. Hoffie, I. Hoffie, J. Kumlehn, A. Bültemeier, J. Wolf, M. Herde, C.-P. Witte, J. C. D'Auria and J. Franke, *Science*, 2024, **383**, 1448–1454.

94 K. Blin, S. Shaw, L. Vader, J. Szenei, Z. L. Reitz, H. E. Augustijn, J. D. D. Cediel-Becerra, V. de Crécy-Lagard, R. A. Koetsier, S. E. Williams, P. Cruz-Morales,

S. Wongwas, A. E. Segurado Luchsinger, F. Biermann, A. Korenskaia, M. M. Zdouc, D. Meijer, B. R. Terlouw, J. J. J. van der Hooft, N. Ziemert, E. J. N. Helfrich, J. Masschelein, C. Corre, M. G. Chevrette, G. P. van Wezel, M. H. Medema and T. Weber, *Nucleic Acids Res.*, 2025, **53**, W32–W38, https://academic.oup.com/nar/article/53/W1/W32/8119805.

95 E. Del Pup, C. Owen, Z. Luo, H. E. Augustijn, A. Draisma, G. Polturak, S. A. Kautsar, A. Osbourn, J. J. J. van der Hooft and M. H. Medema, *J. Mol. Biol.*, 2026, 169798, https://www.sciencedirect.com/science/article/pii/S0022283626001713.

96 N. Töpfer, L.-M. Fuchs and A. Aharoni, *Nucleic Acids Res.*, 2017, **45**, 7049–7063, https://academic.oup.com/nar/article/45/12/7049/3806663.

97 P. Schläpfer, P. Zhang, C. Wang, T. Kim, M. Banf, L. Chae, K. Dreher, A. K. Chavali, R. Nilo-Poyanco, T. Bernard, D. Kahn and S. Y. Rhee, *Plant Physiol.*, 2017, **173**, 2041–2059, https://academic.oup.com/plphys/article/173/4/2041/6115986#236478650.

98 S. A. Kautsar, H. G. Suarez Duran, K. Blin, A. Osbourn and M. H. Medema, *Nucleic Acids Res.*, 2017, **45**, W55–W63, https://academic.oup.com/nar/article/45/W1/W55/3769247.

99 S. Wu, A. L. Malaco Morotti, S. Wang, Y. Wang, X. Xu, J. Chen, G. Wang and E. C. Tatsis, *New Phytol.*, 2022, **235**, 646–661, DOI: 10.1111/nph.18138.

100 J. Reed, A. Orme, A. El-Demerdash, C. Owen, L. B. B. Martin, R. C. Misra, S. Kikuchi, M. Rejzek, A. C. Martin, A. Harkess, J. Leebens-Mack, T. Louveau, M. J. Stephenson and A. Osbourn, *Science*, 2023, **379**, 1252–1264.

101 D. Kong, S. Li and C. D. Smolke, *Sci. Adv.*, 2020, **6**, eabd1143.

102 L. K. Caesar, R. Montaser, N. P. Keller and N. L. Kelleher, *Nat. Prod. Rep.*, 2021, **38**, 2041–2065, https://pubs.rsc.org/en/content/articlehtml/2021/np/d1np00036e.

103 E. Lyons, B. Pedersen, J. Kane, M. Alam, R. Ming, H. Tang, X. Wang, J. Bowers, A. Paterson, D. Lisch and M. Freeling, *Plant Physiol.*, 2008, **148**, 1772–1781, https://academic.oup.com/plphys/article/148/4/1772/6107434.

104 H. Tang, V. Krishnakumar, X. Zeng, Z. Xu, A. Taranto, J. S. Lomas, Y. Zhang, Y. Huang, Y. Wang, W. C. Yim, J. Zhang and X. Zhang, *iMeta*, 2024, **3**, e211, DOI: 10.1002/imt2.211.

105 H. Tang, J. E. Bowers, X. Wang, R. Ming, M. Alam and A. H. Paterson, *Science*, 2008, **320**, 486–488, https://pubmed.ncbi.nlm.nih.gov/18436778/.

106 B. Pucker, N. Walker-Hale, J. Dzurlic, W. C. Yim, J. C. Cushman, A. Crum, Y. Yang and S. F. Brockington, *New Phytol.*, 2024, **241**, 471–489, DOI: 10.1111/nph.19341.

107 Y. Wang, H. Tang, J. D. DeBarry, X. Tan, J. Li, X. Wang, T. Lee, H. Jin, B. Marler, H. Guo, J. C. Kissinger and A. H. Paterson, *Nucleic Acids Res.*, 2012, **40**, e49, https://academic.oup.com/nar/article/40/7/e49/1202057.

108 Y. Wang, H. Tang, X. Wang, Y. Sun, P. V. Joseph and A. H. Paterson, *Nat. Protoc.*, 2024, **19**, 2206–2229, https://www.nature.com/articles/s41596-024-00968-2?fromPaywallRec=false.

109 S. M. Kiełbasa, R. Wan, K. Sato, P. Horton and M. C. Frith, *Genome Res.*, 2011, **21**, 487–493, https://pubmed.ncbi.nlm.nih.gov/21209072/.

110 B. Buchfink, K. Reuter and H.-G. Drost, *Nat. Methods*, 2021, **18**, 366–368, https://www.nature.com/articles/s41592-021-01101-x.

111 C. Chen, Y. Wu, J. Li, X. Wang, Z. Zeng, J. Xu, Y. Liu, J. Feng, H. Chen, Y. He and R. Xia, *Mol. Plant*, 2023, **16**, 1733–1742, https://www.sciencedirect.com/science/article/pii/S1674205223002812.

112 E. E. Reynolds, M. Trauger, F.-S. Li, J. Huang, T. Moss, B. Christ, M. Xu, E. Knoch and J.-K. Weng, *Nat. Plants*, 2026, **12**, 432–446, https://www.nature.com/articles/s41477-026-02220-z.

113 S. Priego-Cubero, E. Knoch, Z. Wang, S. Alseekh, K.-H. Braun, P. Chapman, A. R. Fernie, C. Liu and C. Becker, *Proc. Natl. Acad. Sci. U. S. A.*, 2025, **122**, e2420164122.

114 E. V. Koonin, *Annu. Rev. Genet.*, 2005, **39**, 309–338, DOI: 10.1146/annurev.genet.39.073003.114725.

115 J. T. Lovell, A. Sreedasyam, M. E. Schranz, M. Wilson, J. W. Carlson, A. Harkess, D. Emms, D. M. Goodstein and J. Schmutz, *eLife*, 2022, **11**, e78526.

116 T. Zhao and M. E. Schranz, *Curr. Opin. Plant Biol.*, 2017, **36**, 129–134, https://www.sciencedirect.com/science/article/pii/S1369526616302230.

117 T. Zhao, R. Holmer, S. de Bruijn, G. C. Angenent, H. A. van den Burg and M. E. Schranz, *Plant Cell*, 2017, **29**, 1278–1292, https://academic.oup.com/plcell/article/29/6/1278/6099368.

118 F. Almeida-Silva, T. Zhao, K. K. Ullrich, M. E. Schranz and Y. van de Peer, *Bioinformatics*, 2023, **39**, https://academic.oup.com/bioinformatics/article/39/1/btac806/6947985.

119 H. Li, J. Li, X. Li, J. Li, D. Chen, Y. Zhang, Q. Yu, F. Yang, Y. Liu, W. Dai, Y. Sun, P. Li, M. E. Schranz, F. Ma and T. Zhao, *New Phytol.*, 2025, **245**, 2150–2169, DOI: 10.1111/nph.20357.

120 M. Marcet-Houben and T. Gabaldón, *Bioinformatics*, 2020, **36**, 1265–1266, https://academic.oup.com/bioinformatics/article/36/4/1265/5575072.

121 E. Gluck-Thaler, S. Haridas, M. Binder, I. V. Grigoriev, P. W. Crous, J. W. Spatafora, K. Bushley and J. C. Slot, 2020.

122 Z. Konkel, L. Kubatko and J. C. Slot, *Nucleic Acids Res.*, 2024, **52**, e75, https://academic.oup.com/nar/article/52/16/e75/7715716.

123 B. Song, W. Ning, Di Wei, M. Jiang, K. Zhu, X. Wang, D. Edwards, D. A. Odeny and S. Cheng, *Mol. Plant*, 2023, **16**, 1252–1268, https://www.sciencedirect.com/science/article/pii/S1674205223002113.

124 C. Fang and J. Luo, *Plant J.*, 2019, **97**, 91–100, https://pubmed.ncbi.nlm.nih.gov/30231195/.

125 M. Brouckaert, M. Peng, R. Höfer, I. El Houari, C. Darrah, V. Storme, Y. Saeys, R. Vanholme, G. Goeminne, V. I. Timokhin, J. Ralph, K. Morreel and W. Boerjan, *Mol. Plant*, 2023, **16**, 1212–1227, https://www.cell.com/molecular-plant/fulltext/S1674-2052(23)00170-3.

126 D. Mancinotti, K. Czepiel, J. L. Taylor, H. Golshadi Galehshahi, L. A. Møller, M. K. Jensen, M. S. Motawia, B. Hufnagel, A. Soriano, L. Yeheyis, L. Kjaerulff, B. Péret, D. Staerk, T. Wendt, M. N. Nelson, M. Kroc and F. Geu-Flores, *Sci. Adv.*, 2023, **9**, eadg8866.

127 M. Jayakodi, H. Shim and M. Mascher, *Annu. Rev. Plant Biol.*, 2025, **76**, 663–686, DOI: 10.1146/annurev-arplant-090823-015358.

128 H. M. Schilbert, B. Pucker, D. Ries, P. Viehöver, Z. Micic, F. Dreyer, K. Beckmann, B. Wittkop, B. Weisshaar and D. Holtgräwe, *Genes*, 2022, **13**, 1131, https://www.mdpi.com/2073-4425/13/7/1131.

129 X. Zhu, R. Yang, Q. Liang, Y. Yu, T. Wang, L. Meng, P. Wang, S. Wang, X. Li, Q. Yang, H. Guo, Q. Sui, Q. Wang, H. Du, Q. Chen, Z. Liang, X. Wu, Q. Zeng and B. Huang, *Mol. Plant*, 2025, **18**, 590–602, https://www.cell.com/molecular-plant/fulltext/S1674-2052(25)00038-3.

130 F. He, S. Chen, Y. Zhang, K. Chai, Q. Zhang, W. Kong, S. Qu, L. Chen, F. Zhang, M. Li, X. Wang, H. Lv, T. Zhang, X. He, X. Li, Y. Li, X. Li, X. Jiang, M. Xu, B. Sod, J. Kang, X. Zhang, R. Long and Q. Yang, *Nat. Genet.*, 2025, **57**, 1262–1273, https://www.nature.com/articles/s41588-025-02164-8.

131 Z.-Z. Du, J.-B. He and W.-B. Jiao, *aBIOTECH*, 2025, **6**, 361–376, https://www.sciencedirect.com/science/article/pii/S266217382500205X.

132 S. Fan, H. Yang, Y. Hu, L. Zhang and M. Huang, *Plant Stress*, 2026, **19**, 101144, https://www.sciencedirect.com/science/article/pii/S2667064X25004117.

133 P. Hüther, J. Hagmann, A. Nunn, I. Kakoulidou, R. Pisupati, D. Langenberger, D. Weigel, F. Johannes, S. J. Schultheiss and C. Becker, *Quant. Plant Biol.*, 2022, **3**, e19, https://www.cambridge.org/core/journals/quantitative-plant-biology/article/methylscore-a-pipeline-for-accurate-and-contextaware-identification-of-differentially-methylated-regions-from-populationscale-plant-wholegenome-bisulfite-sequencing-data/D241679DECD6275134588BA15A421752.

134 S. N. Can, A. Nunn, D. Galanti, D. Langenberger, C. Becker, K. Volmer, K. Heer, L. Opgenoorth, N. Fernandez-Pozo and S. A. Rensing, *Epigenomes*, 2021, **5**, 12, https://www.mdpi.com/2075-4655/5/2/12.

135 H. Guo, P. Cao, C. Wang, J. Lai, Y. Deng, C. Li, Y. Hao, Z. Wu, R. Chen, Q. Qiang, A. R. Fernie, J. Yang and S. Wang, *Sci. China Life Sci.*, 2023, **66**, 1888–1902.

136 H. Le, C. H. Simmons and X. Zhong, *Annu. Rev. Plant Biol.*, 2025, **76**, 551–578, DOI: 10.1146/annurev-arplant-083123-070919.

137 C. G. Bell, *Cell. Mol. Life Sci.*, 2024, **81**, 178, DOI: 10.1007/s00018-024-05206-2.

138 K. Domb, N. Wang, G. Hummel and C. Liu, *Annu. Rev. Plant Biol.*, 2022, **73**, 173–200, DOI: 10.1146/annurev-arplant-102720-022810.

139 S. Xiao, L. Luo, M. Yang, H. He and Y. Zhou, *Curr. Opin. Plant Biol.*, 2025, **88**, 102786, https://www.sciencedirect.com/science/article/pii/S1369526625001001.

140 H. Lee and P. J. Seo, *J. Exp. Bot.*, 2025, **77**, 134–140.

141 H. Zhao, M. Yang, J. Bishop, Y. Teng, Y. Cao, B. D. Beall, S. Li, T. Liu, Q. Fang, C. Fang, H. Xin, H.-W. Nützmann, A. Osbourn, F. Meng and J. Jiang, *Proc. Natl. Acad. Sci. U. S. A.*, 2022, **119**, e2215328119, DOI: 10.1073/pnas.2215328119.

142 F. Bai, P. Shu, H. Deng, Y. Wu, Y. Chen, M. Wu, T. Ma, Y. Zhang, J. Pirrello, Z. Li, Y. Hong, M. Bouzayen and M. Liu, *Nat. Commun.*, 2024, **15**, 2894, https://www.nature.com/articles/s41467-024-47292-7.

143 Y. Huang, J. An, S. Sircar, C. Bergis, C. D. Lopes, X. He, B. Da Costa, F.-Q. Tan, J. Bazin, J. Antunez-Sanchez, M. F. Mammarella, R. Devani, R. Brik-Chaouche, A. Bendahmane, F. Frugier, C. Xia, C. Rothan, A. V. Probst, Z. Mohamed, C. Bergounioux, M. Delarue, Y. Zhang, S. Zheng, M. Crespi, S. Fragkostefanakis, M. M. Mahfouz, F. Ariel, J. Gutierrez-Marcos, C. Raynaud, D. Latrasse and M. Benhamed, *Nat. Commun.*, 2023, **14**, 469, https://www.nature.com/articles/s41467-023-36227-3.

144 G. Pei, H. Lyons, P. Li and B. R. Sabari, *Nat. Rev. Mol. Cell Biol.*, 2025, **26**, 213–236, https://www.nature.com/articles/s41580-024-00789-x.

145 M. Levo, J. Raimundo, X. Y. Bing, Z. Sisco, P. J. Batut, S. Ryabichko, T. Gregor and M. S. Levine, *Nature*, 2022, **605**, 754–760, https://www.nature.com/articles/s41586-022-04680-7.

146 A. S. Deshpande, N. Ulahannan, M. Pendleton, X. Dai, L. Ly, J. M. Behr, S. Schwenk, W. Liao, M. A. Augello, C. Tyer, P. Rughani, S. Kudman, H. Tian, H. G. Otis, E. Adney, D. Wilkes, J. M. Mosquera, C. E. Barbieri, A. Melnick, D. Stoddart, D. J. Turner, S. Juul, E. Harrington and M. Imieliński, *Nat. Biotechnol.*, 2022, **40**, 1488–1499, https://www.nature.com/articles/s41587-022-01289-z.

147 S. P. McGinty, G. Kaya, S. B. Sim, A. Makunin, R. L. Corpuz, M. A. Quail, M. Abuelanin, M. K. N. Lawniczak, S. M. Geib, J. Korlach and M. Y. Dennis, *Nat. Commun.*, 2025, **17**, 215, https://www.nature.com/articles/s41467-025-66918-y.

148 L. Sun, Y. Cao, Z. Li, Y. Liu, X. Yin, X. W. Deng, H. He and W. Qian, *J. Integr. Plant Biol.*, 2023, **65**, 1966–1982.

149 T. Pollex, R. Marco-Ferreres, L. Ciglar, Y. Ghavi-Helm, A. Rabinowitz, R. R. Viales, C. Schaub, A. Jankowski, C. Girardot and E. E. M. Furlong, *Mol. Cell*, 2024, **84**, 822–838.e8, https://www.cell.com/molecular-cell/fulltext/S1097-2765(23)01042-0?uuid=uuid3A6fbc9f1c-d33f-4db2-b406-05857f575a01.

150 C. Liu, D. Li, J. Dang, J. Shu, S. J. Smit, Q. Wu and B. R. Lichman, *Hortic. Res.*, 2025, **12**, uhaf034, https://academic.oup.com/hr/article/12/5/uhaf034/7994528.

151 P. Martin, A. McGovern, G. Orozco, K. Duffus, A. Yarwood, S. Schoenfelder, N. J. Cooper, A. Barton, C. Wallace, P. Fraser, J. Worthington and S. Eyre, *Nat. Commun.*, 2015, **6**, 10069, https://www.nature.com/articles/ncomms10069#Sec8.

152 S. E. Hakim, S. Liu, R. Herzog, A. Arafa, J. de Vries, G. Dräger and J. Franke, *J. Am. Chem. Soc.*, 2025, **147**, 10320–10330.

153 M. J. Stephenson, C. Owen, J. Reed and A. Osbourn, *Nat. Chem. Biol.*, 2025, https://www.nature.com/articles/s41589-025-02034-8.
154 M. Engst, M. Brokeš, T. Čalounová, R. Samusevich, R. Bushuiev, A. Bushuiev, R. Chatpatanasiri, A. Tajovská, S. M. Akmeșe, M. Perković, M. Soldát, J. Sivic and T. Pluskal, *BMC Bioinf.*, 2025, **27**, 10, DOI: 10.1186/s12859-025-06341-8.
155 I. Feussner and A. Polle, *Curr. Opin. Plant Biol.*, 2015, **26**, 26–31.
156 Y. Liu, A. Beyer and R. Aebersold, *Cell*, 2016, **165**, 535–550.
157 M. J. Martínez-Esteso, J. Morante-Carriel, A. Samper-Herrero, A. Martínez-Márquez, S. Sellés-Marchart, H. Nájera and R. Bru-Martínez, *Biomolecules*, 2024, **14**, 1539.
158 N. Samanani and P. J. Facchini, *J. Biol. Chem.*, 2002, **277**, 33878–33883, https://www.jbc.org/article/S0021-9258(20)74275-4/fulltext.
159 P. Steffens, N. Nagakura and M. H. Zenk, *Phytochemistry*, 1985, **24**, 2577–2583, https://www.sciencedirect.com/science/article/pii/S003194220080672X.
160 S. Yan, R. Bhawal, Z. Yin, T. W. Thannhauser and S. Zhang, *Mol. Hortic.*, 2022, **2**, 17.
161 L. Ponnala, Y. Wang, Q. Sun and K. J. van Wijk, *Plant J.*, 2014, **78**, 424–440.
162 G. Friso and K. J. van Wijk, *Plant Physiol.*, 2015, **169**, 1469–1487, https://academic.oup.com/plphys/article/169/3/1469/6114070?login=true.
163 N. Ntelkis, A. Goossens and K. Šola, *Curr. Opin. Plant Biol.*, 2024, **81**, 102575, https://www.sciencedirect.com/science/article/pii/S1369526624000669.
164 R. Aebersold and M. Mann, *Nature*, 2003, **422**, 198–207, https://www.nature.com/articles/nature01511.
165 A. Onoyovwe, J. M. Hagel, X. Chen, M. F. Khan, D. C. Schriemer and P. J. Facchini, *Plant Cell*, 2013, **25**, 4110–4122.
166 X. Pu, M. Wang, M. Chen, X. Lin, M. Lei, J. Zhang, S. Yang, H. Wang, J. Liao, L. Zhang and Q. Huang, *ACS Chem. Biol.*, 2023, **18**, 1772–1785.
167 M. Dastmalchi, X. Chen, J. M. Hagel, L. Chang, R. Chen, S. Ramasamy, S. Yeaman and P. J. Facchini, *Nat. Chem. Biol.*, 2019, **15**, 384–390, https://www.nature.com/articles/s41589-019-0247-0.
168 B. Seligmann, S. Liu and J. Franke, *Curr. Opin. Plant Biol.*, 2024, **80**, 102554.
169 Q. Yin and M. Yang, *Front. Plant Sci.*, 2024, **15**, 1506569.
170 M. H. Wright and S. A. Sieber, *Nat. Prod. Rep.*, 2016, **33**, 681–708.
171 H. Fang, B. Peng, S. Y. Ong, Q. Wu, L. Li and S. Q. Yao, *Chem. Sci.*, 2021, **12**, 8288–8310, https://xlink.rsc.org/?DOI=D1SC01359A.
172 X.-L. Jia, L. Zhu, Y. Li, P. Zhang, X. Chen, K. Shao, J. Feng, S. Qiu, J. Geng, Y. Yang, Z. Wu, J. Xue, P. Wang, W. Chen and Y. Xiao, *New Phytol.*, 2024, **244**, 1901–1915, DOI: 10.1111/nph.20104.
173 X. Chen, Y. Wang, N. Ma, J. Tian, Y. Shao, B. Zhu, Y. K. Wong, Z. Liang, C. Zou and J. Wang, *Signal Transduction Targeted Ther.*, 2020, **5**, 72, https://www.nature.com/articles/s41392-020-0186-y.
174 W. Li, Y. Zhou, W. You, M. Yang, Y. Ma, M. Wang, Y. Wang, S. Yuan and Y. Xiao, *ACS Chem. Biol.*, 2018, **13**, 1944–1949, DOI: 10.1021/acschembio.8b00285.
175 L. Gao, C. Su, X. Du, R. Wang, S. Chen, Y. Zhou, C. Liu, X. Liu, R. Tian, L. Zhang, K. Xie, S. Chen, Q. Guo, L. Guo, Y. Hano, M. Shimazaki, A. Minami, H. Oikawa, N. Huang, K. N. Houk, L. Huang, J. Dai and X. Lei, *Nat. Chem.*, 2020, **12**, 620–628.
176 A. Hoegl, M. B. Nodwell, V. C. Kirsch, N. C. Bach, M. Pfanzelt, M. Stahl, S. Schneider and S. A. Sieber, *Nat. Chem.*, 2018, **10**, 1234–1245.
177 T. Obata, *Phytochem. Rev.*, 2019, **18**, 1483–1507, DOI: 10.1007/s11101-019-09619-x.
178 K. Jørgensen, A. V. Rasmussen, M. Morant, A. H. Nielsen, N. Bjarnholt, M. Zagrobelny, S. Bak and B. L. Møller, *Curr. Opin. Plant Biol.*, 2005, **8**, 280–291, https://www.sciencedirect.com/science/article/pii/S1369526605000464.
179 V. I. Strotmann and Y. Stahl, *J. Exp. Bot.*, 2022, **73**, 3866–3880, https://academic.oup.com/jxb/article/73/12/3866/6565416?login=false.
180 T. Laursen, J. Borch, C. Knudsen, K. Bavishi, F. Torta, H. J. Martens, D. Silvestro, N. S. Hatzakis, M. R. Wenk, T. R. Dafforn, C. E. Olsen, M. S. Motawia, B. Hamberger, B. L. Møller and J.-E. Bassard, *Science*, 2016, **354**, 890–893, DOI: 10.1126/science.aag2347.
181 Y. Wu, C. Liu, A. Koganitsky, F. L. Gong and S. Li, *Angew. Chem., Int. Ed.*, 2023, **62**, e202307995.
182 S.-L. Xu, R. Shrestha, S. S. Karunadasa and P.-Q. Xie, *Annu. Rev. Plant Biol.*, 2023, **74**, 285–312, DOI: 10.1146/annurev-arplant-070522-052132.
183 T. C. Branon, J. A. Bosch, A. D. Sanchez, N. D. Udeshi, T. Svinkina, S. A. Carr, J. L. Feldman, N. Perrimon and A. Y. Ting, *Nat. Biotechnol.*, 2018, **36**, 880–887.
184 J. Aravena-Calvo, S. Busck-Mellor and T. Laursen, *Front. Plant Sci.*, 2024, **15**, 1295750.
185 M. Altaf-Ul-Amin, F. M. Afendi, S. K. Kiboi and S. Kanaya, *BioMed Res. Int.*, 2014, **2014**, 428570.
186 One Thousand Plant Transcriptomes Initiative, *Nature*, 2019, **574**, 679–685.
187 J. K. Reinhardt, D. Craft and J.-K. Weng, *Trends Biochem. Sci.*, 2025, **50**, 311–321, https://www.cell.com/trends/biochemical-sciences/fulltext/S0968-0004(25)00025-8.
188 S. H. Swamidatta and B. R. Lichman, *Curr. Opin. Biotechnol.*, 2024, **88**, 103147.
189 M. Durand, S. Besseau, N. Papon and V. Courdavault, *Curr. Opin. Biotechnol.*, 2024, **87**, 103135.
190 M. Hesami, M. Alizadeh, A. M. P. Jones and D. Torkamaneh, *Appl. Microbiol. Biotechnol.*, 2022, **106**, 3507–3530.
191 I. Khan and B. K. Khare, *Plant Gene*, 2025, **44**, 100542.
192 N. Kühl, M. Schemmer, M. Goutier and G. Satzger, *Electron. Mark.*, 2022, **32**, 2235–2244.
193 L. Liao, M. Xie, X. Zheng, Z. Zhou, Z. Deng and J. Gao, *Nat. Prod. Rep.*, 2025, **42**, 911–936.

194 C. J. Vavricka, S. Takahashi, N. Watanabe, M. Takenaka, M. Matsuda, T. Yoshida, R. Suzuki, H. Kiyota, J. Li, H. Minami, J. Ishii, K. Tsuge, M. Araki, A. Kondo and T. Hasunuma, *Nat. Commun.*, 2022, **13**, 1405.
195 P. Wang, A. M. Schumacher and S. Shiu, *Curr. Opin. Plant Biol.*, 2022, **66**, 102171, https://www.sciencedirect.com/science/article/pii/S1369526621001734.
196 T. Kohonen, *Proc. IEEE*, 1990, **78**, 1464–1480.
197 T. Kohonen, *Biol. Cybern.*, 1982, **43**, 59–69.
198 T.-T. T. Dang, J. Franke, I. S. T. Carqueijeiro, C. Langley, V. Courdavault and S. E. O'Connor, *Nat. Chem. Biol.*, 2018, **14**, 760–763.
199 R. Lin, H. Li, Y. Xiao, Z. Wang, L. Liu, G. Saalbach, C. Martins, M. Furry, C. D. Vanderwal, C. Martin and E. C. Tatsis, *Plant Commun.*, 2025, **6**, 101286.
200 Y. Wang, A. L. Malaco Morotti, Y. Xiao, Z. Wang, S. Wu, J. Chen and E. C. Tatsis, *ACS Catal.*, 2022, **12**, 13630–13637.
201 J. K. Kim, T. Bamba, K. Harada, E. Fukusaki and A. Kobayashi, *J. Exp. Bot.*, 2007, **58**, 415–424.
202 S. E. Bamford, W. Gardner, D. A. Winkler, B. W. Muir, D. Alahakoon and P. J. Pigram, *J. Am. Soc. Mass Spectrom.*, 2024, **35**, 2516–2528.
203 D. Bustos-Korts, M. P. Boer, J. Layton, A. Gehringer, T. Tang, R. Wehrens, C. Messina, A. J. de La Vega and F. A. van Eeuwijk, *Theor. Appl. Genet.*, 2022, **135**, 2059–2082.
204 P. Chalise, D. Kwon, B. L. Fridley and Q. Mo, *Methods Mol. Biol.*, 2023, **2629**, 73–93.
205 M. Lovino, V. Randazzo, G. Ciravegna, P. Barbiero, E. Ficarra and G. Cirrincione, *Neurocomputing*, 2022, **488**, 494–508.
206 R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber and O. Stegle, *Mol. Syst. Biol.*, 2018, **14**, e8124.
207 A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt and K.-A. Lê Cao, *Bioinformatics*, 2019, **35**, 3055–3062.
208 R. Shen, A. B. Olshen and M. Ladanyi, *Bioinformatics*, 2009, **25**, 2906–2912.
209 T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding and K. Huang, *Nat. Commun.*, 2021, **12**, 3445.
210 B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nat. Methods*, 2014, **11**, 333–337.
211 S. Moon and H. Lee, *Bioinformatics*, 2022, **38**, 2287–2296.
212 T. Hernández-Lao, M. Tienda-Parrilla, M. Labella-Ortega, V. M. Guerrero-Sánchez, M.-D. Rey, J. V. Jorrín-Novo and M. Á. Castillejo-Sánchez, *Biomolecules*, 2024, **14**, 160.
213 O. Arriagada, C. Meneses, R. Pedreschi, G. Núñez-Lillo, C. Maureira, S. Reveco, V. Villarroel, Ú. Steinfort, F. Albornoz, P. Cabas-Lühmann, M. Silva, I. Matus and A. R. Schwember, *Front. Plant Sci.*, 2025, **16**, 1540179.
214 K. S. Singh, H. Suarez Duran, E. Del Pup, O. Zafra Delgado, S. C. M. van Wees, J. J. J. van der Hooft and M. H. Medema, *PLoS Biol.*, 2025, **23**, e3003307, DOI: 10.1371/journal.pbio.3003307.
215 H. Y. in Lam, X. E. Ong and M. Mutwil, *Trends Plant Sci.*, 2024, **29**, 1145–1155.
216 A. Hajikhani and C. Cole, *Quant. Sci. Stud.*, 2024, **5**, 736–756.
217 Z. P. Wang, P. Bhandary, Y. Wang and J. H. Moore, *BioData Min.*, 2024, **17**, 16.
218 X. Yang, J. Gao, W. Xue and E. Alexandersson, *PLLaMa: An Open-source Large Language Model for Plant Science*, 2024.
219 R. Knapp, B. Johnson and L. Busta, *Appl. Plant Sci.*, 2025, **13**, e70007.
220 N. Smith, X. Yuan, C. Melissinos and G. Moghe, *Bioinformatics*, 2024, **41**, btae756.
221 L. Busta and A. R. Oyler, *Quant. Plant Biol.*, 2025, **6**, e26.
222 V. Domazetoski, H. Kreft, H. Bestova, P. Wieder, R. Koynov, A. Zarei and P. Weigelt, *Appl. Plant Sci.*, 2025, **13**, e70011.
223 L. Busta, D. Hall, B. Johnson, M. Schaut, C. M. Hanson, A. Gupta, M. Gundrum, Y. Wang and H. A. Maeda, *Plant J.*, 2024, **120**, 406–419.
224 S. McGinnis and T. L. Madden, *Nucleic Acids Res.*, 2004, **32**, W20–W25.
225 D. M. Emms, Y. Liu, L. Belcher, J. Holmes and S. Kelly, *bioRxiv*, 2025, preprint arXiv:2025.07.15.664860, DOI: 10.1101/2025.07.15.664860v1.
226 M. Blum, A. Andreeva, L. C. Florentino, S. R. Chuguransky, T. Grego, E. Hobbs, B. L. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, G. A. Salazar, N. Bordin, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunic, F. Llinares-López, A. Marchler-Bauer, L. Meng-Papaxanthos, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, D. Piovesan, C. Rivoire, C. J. A. Sigrist, N. Thanki, F. Thibaud-Nissen, P. D. Thomas, S. C. E. Tosatto, C. H. Wu and A. Bateman, *Nucleic Acids Res.*, 2025, **53**, D444–D456.
227 M. Wang, Z. Nie, Y. He, A. V. Vasilakos, Q. Cheng and Z. Ren, *Eng. Appl. Artif. Intell.*, 2025, **155**, 110977.
228 R. S. Sunil, S. C. Lim, M. Itharajula and M. Mutwil, *Curr. Opin. Plant Biol.*, 2024, **82**, 102665.
229 S. Seo, M. Oh, Y. Park and S. Kim, *Bioinformatics*, 2018, **34**, i254–i262.
230 M. L. Bileschi, D. Belanger, D. H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M. A. DePristo and L. J. Colwell, *Nat. Biotechnol.*, 2022, **40**, 932–937.
231 M. Kulmanov and R. Hoehndorf, *Bioinformatics*, 2020, **36**, 422–429.
232 M. Li, P. Tan, X. Ma, B. Zhong, H. Yu, Z. Zhou, W. Ouyang, B. Zhou, L. Hong and Y. Tan, *bioRxiv*, 2024, preprint arXiv:2024.04.15.589672, DOI: 10.1101/2024.04.15.589672v3.
233 Y. Bhanushe, S. S. Tamrapalli, B. Mahanty, R. Mishra and R. K. Joshi, *J. Hortic. Sci. Biotechnol.*, 2025, **100**, 329–342.
234 S. Chen, T. Du, Z. Huang, K. He, M. Yang, S. Gao, T. Yu, H. Zhang, X. Li, S. Chen, C.-M. Liu and H. Li, *Plant Biotechnol. J.*, 2024, **22**, 2558–2574.
235 J. Hallgren, K. D. Tsirigos, M. D. Pedersen, J. J. Almagro Armenteros, P. Marcatili, H. Nielsen, A. Krogh and O. Winther, *bioRxiv*, 2022, preprint arXiv:2022.04.08.487609, DOI: 10.1101/2022.04.08.487609v1.

236 F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, M. H. Gíslason, S. I. Pihl, K. D. Tsirigos, O. Winther, S. Brunak, G. von Heijne and H. Nielsen, *Nat. Biotechnol.*, 2022, **40**, 1023–1025.
237 H. Nielsen, *Methods Mol. Biol.*, 2025, **2941**, 153–175.
238 Y. Jiang, D. Wang, Y. Yao, H. Eubel, P. Künzler, I. M. Møller and D. Xu, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 4825–4839.
239 D. Ma, H. Gordon, R. Nazir, J. E. Wulff and C. P. Constabel, *Plant Cell*, 2025, **37**, koaf241.
240 Y.-L. Zheng, Y. Xu, Y.-Q. Liu, Q.-W. Zhao and Y.-Q. Li, *ACS Synth. Biol.*, 2024, **13**, 3600–3608.
241 M. Xu, H. Li, H. Luo, J. Liu, K. Li, Q. Li, N. Yang and D. Xu, *Int. J. Mol. Sci.*, 2024, **25**, 13191.
242 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis and J. M. Jumper, *Nature*, 2024, **630**, 493–500.
243 Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido and A. Rives, *Science*, 2023, **379**, 1123–1130.
244 M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read and D. Baker, *Science*, 2021, **373**, 871–876.
245 B. Lamichhane, S.-E. Gélinas, N. Merindol, M. Koirala, K. C. G. Dos Santos, H. Germain and I. Desgagné-Penix, *Plant Biotechnol. J.*, 2025, **23**, 1988–2005.
246 S. Gupta, B. A. Akhoon, D. Sharma, D. Singh, S. Kaul and M. K. Dhar, *Steroids*, 2025, **214**, 109557.
247 P. Xu, J. Li, C. Chen, J. Chen, M. Yang, H. Deng, X. Jiang, K. Lou, X. Wu, R. Chen, Y. Hu, W. Liang and J. Pu, *BMC Genom.*, 2025, **26**, 561.
248 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. Jaakkola, *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*, 2022.
249 A. T. McNutt, Y. Li, R. Meli, R. Aggarwal and D. R. Koes, *J. Cheminf.*, 2025, **17**, 28.
250 H. Lai, L. Wang, R. Qian, J. Huang, P. Zhou, G. Ye, F. Wu, F. Wu, X. Zeng and W. Liu, *Nat. Commun.*, 2024, **15**, 10223.
251 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
252 A. Grosdidier, V. Zoete and O. Michielin, *Nucleic Acids Res.*, 2011, **39**, W270–W277.
253 R. Yang, C. Fu, X. Zhao, G. Xiang, W. Song, T. C. Baldwin, G. Chen, Y. Wang, Y. Zhao, Y. Xu, Y. Shu, S. Duan, B. Hao, S. He, G. Zhang, S. Yang and Y. Liang, *Plant Physiol. Biochem.*, 2025, **227**, 110166.
254 B. Chen, S. Zheng, H. Wang, R. Yang, Y. Xiang, Y. Huang, J. Pei, Y. Zhang and R. Fu, *Int. J. Biol. Macromol.*, 2025, **311**, 143924.
255 T. S. Waheeb, M. A. Abdulkader, D. A. Ghareeb and M. E. Moustafa, *Inflammopharmacology*, 2025, **33**, 2129–2150.
256 S. Aimjongjun, N. Khamto, V. Buangamdee, T. Sornda, J. Srivilai and N. Limpeanchob, *Int. J. Mol. Sci.*, 2025, **26**, 10158.
257 Y. Li, J. Yi, H. Li, K. Li, F. Kang, Y. Deng, C. Wu, X. Fu, D. Jiang and D. Cao, *Chem. Sci.*, 2025, **16**, 17374–17390.
258 S. Das, M. Shimshi, K. Raz, N. Nitoker Eliaz, A. R. Mhashal, T. Ansbacher and D. T. Major, *J. Chem. Theory Comput.*, 2019, **15**, 5116–5134.
259 R. Schwartz, S. Zev and D. T. Major, *Methods Enzymol.*, 2024, **699**, 265–292.
260 A. Kroll, Y. Rousset, T. Spitzlei and M. J. Lercher, *Nucleic Acids Res.*, 2025, **53**, W213–W218.
261 J. Wang, Z. Yang, C. Chen, G. Yao, X. Wan, S. Bao, J. Ding, L. Wang and H. Jiang, *Briefings Bioinf.*, 2024, **25**, bbae387.
262 J. Chang, X. Fan and B. Tian, *J. Chem. Inf. Model.*, 2024, **64**, 3149–3160.
263 M. Holmer, C. de Bruyn Kops, C. Stork and J. Kirchmair, *Molecules*, 2021, **26**, 4678.
264 K. Jung, I.-H. Jo, B. Y. Choi and J. Kim, *BioTech*, 2025, **14**, 73.
265 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, *bioRxiv*, 2021, preprint arXiv:2021.10.04.463034, DOI: 10.1101/2021.10.04.463034v2.
266 J. Pan, Z.-H. You, L.-P. Li, W.-Z. Huang, J.-X. Guo, C.-Q. Yu, L.-P. Wang and Z.-Y. Zhao, *Front. Bioeng. Biotechnol.*, 2022, **10**, 807522.
267 S. Murmu, H. Chaurasia, A. R. Rao, A. Rai, S. Jaiswal, A. Bharadwaj, R. Yadav and S. Archak, *J. Mol. Biol.*, 2025, **437**, 169093.
268 J. Zheng, X. Yang, Y. Huang, S. Yang, S. Wuchty and Z. Zhang, *Plant J.*, 2023, **114**, 984–994.
269 K. Zhou, C. Lei, J. Zheng, Y. Huang and Z. Zhang, *Plant Methods*, 2023, **19**, 141.
270 C. Liu, J. Han and S. Li, *bioRxiv*, 2024, preprint arXiv:2024.10.24.620109, DOI: 10.1101/2024.10.24.620109v1.
271 F. Homma, J. Huang and R. A. L. van der Hoorn, *Nat. Commun.*, 2023, **14**, 6040.
272 Y. Moriwaki, T. Shiraishi, Y. Katsuyama, K. Matsuda, T. Ose, A. Minami, H. Oikawa, T. Kuzuyama, R. Ishitani and T. Terada, *bioRxiv*, 2025, preprint

arXiv:2025.10.26.684697, DOI: 10.1101/2025.10.26.684697v1.

273 T. Kim, S. Lee, Y. Kwak, M.-S. Choi, J. Park, S. J. Hwang and S.-G. Kim, *New Phytol.*, 2024, **243**, 2512–2527.

274 J. Capela, J. Cheixo, D. de Ridder, O. Dias and M. Rocha, *J. Integr. Bioinform.*, 2025, **22**, 20240050.

275 B. M. Moore, P. Wang, P. Fan, B. Leong, C. A. Schenck, J. P. Lloyd, M. D. Lehti-Shiu, R. L. Last, E. Pichersky and S. Shiu, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 2344–2353.

276 W. Bai, C. Li, W. Li, H. Wang, X. Han, P. Wang and L. Wang, *BMC Genom.*, 2024, **25**, 418.

277 R. de Oliveira Almeida and G. T. Valente, *Plant Genome*, 2020, **13**, e20043.

278 WFO, *World Flora Online*, 2025, Published on the Internet, https://www.worldfloraonline.org, (accessed 5 December 2025).

279 H. L. Weil, K. Schneider, M. Tschöpe, J. Bauer, O. Maus, K. Frey, D. Brilhaus, C. Martins Rodrigues, G. Doniparthi, F. Wetzels, J. Lukasczyk, A. Kranz, B. Grüning, D. Zimmer, S. Deßloch, D. von Suchodoletz, B. Usadel, C. Garth and T. Mühlhaus, *Plant J.*, 2023, **116**, 974–988.

280 M. D. Wilkinson, M. Dumontier, I. J. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. Da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons, *Sci. Data*, 2016, **3**, 160018.

281 V. de Luca, V. Salim, D. Levac, S. M. Atsumi and F. Yu, *Methods Enzymol.*, 2012, **515**, 207–229.

282 D. K. Liscombe, A. R. Usera and S. E. O'Connor, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 18793–18798.

283 L.-A. Giddings, D. K. Liscombe, J. P. Hamilton, K. L. Childs, D. DellaPenna, C. R. Buell and S. E. O'Connor, *J. Biol. Chem.*, 2011, **286**, 16751–16757.

284 F. Kellner, J. Kim, B. J. Clavijo, J. P. Hamilton, K. L. Childs, B. Vaillancourt, J. Cepela, M. Habermann, B. Steuernagel, L. Clissold, K. McLay, C. R. Buell and S. E. O'Connor, *Plant J.*, 2015, **82**, 680–692.

285 E. A. Stander, B. Lehka, I. Carqueijeiro, C. Cuello, F. G. Hansson, H. J. Jansen, T. Dugé de Bernonville, C. Birer Williams, V. Vergès, E. Lezin, M. D. B. B. Lorensen, T.-T. Dang, A. Oudin, A. Lanoue, M. Durand, N. Giglioli-Guivarc'h, C. Janfelt, N. Papon, R. P. Dirks, S. E. O'connor, M. K. Jensen, S. Besseau and V. Courdavault, *Commun. Biol.*, 2023, **6**, 1197.

286 T. Dugé de Bernonville, E. Amor Stander, G. Dugé de Bernonville, S. Besseau and V. Courdavault, *Methods Mol. Biol.*, 2022, **2505**, 131–140.

