

Review Article

Open Access

Dynamic Changes in Protein and Oil Contents during Soybean Seed Development

Ling Jin ✉

Northwest A&F University, Xianyang, 712100, Shaanxi, China

✉ Corresponding author: 2024012253@nwfau.edu.cnBioscience Methods, 2026, Vol.17, No.5 doi: [10.5376/bm.2026.17.0025](https://doi.org/10.5376/bm.2026.17.0025)

Received: 18 Aug., 2026

Accepted: 22 Sep., 2026

Published: 07 Oct., 2026

Copyright © 2026 Jin, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:

Jin L., 2026, Dynamic changes in protein and oil contents during soybean seed development, Bioscience Methods, 17(5): 360-372 (doi: [10.5376/bm.2026.17.0025](https://doi.org/10.5376/bm.2026.17.0025))

Abstract Soybean (*Glycine max* L.) is one of the most important global crops, serving as a primary source of plant-based protein and vegetable oil for food, feed, and industrial applications. The accumulation of storage proteins and oils during seed development is a highly dynamic physiological process regulated by genetic programs, metabolic pathways, and environmental conditions. Understanding the temporal patterns and regulatory mechanisms of protein and oil deposition is essential for improving soybean nutritional quality and optimizing breeding strategies. This review systematically summarizes the dynamic changes in protein and oil contents throughout soybean seed development, with emphasis on developmental stages, physiological regulation, and molecular mechanisms underlying storage compound accumulation. During seed filling, protein accumulation is primarily controlled by nitrogen assimilation, amino acid transport, and storage protein biosynthesis, whereas oil accumulation depends on carbon metabolism, fatty acid synthesis, and triacylglycerol assembly. The interaction between carbon and nitrogen metabolism determines the balance between protein and oil contents, resulting in a complex trade-off relationship that influences soybean quality formation. Furthermore, this review discusses the effects of environmental factors, including temperature, water availability, and nutrient management, on seed composition dynamics. Recent advances in multi-omics technologies, high-throughput phenotyping, and machine learning approaches have provided new insights into the prediction and regulation of soybean seed quality. A case study is presented to illustrate how genotype-environment interactions influence protein and oil accumulation patterns under contrasting cultivation conditions. Finally, future perspectives are proposed, focusing on integrating molecular breeding, precision agriculture, and digital technologies to achieve coordinated improvement of soybean yield, protein content, and oil quality. This comprehensive understanding of dynamic storage compound accumulation provides a theoretical foundation for developing high-quality soybean varieties and sustainable production systems.

Keywords Soybean seed development; Protein accumulation; Oil biosynthesis; Carbon-nitrogen metabolism; Seed quality regulation

1 Introduction

Soybean (*Glycine max* L. Merr.) is one of the world's most important crops because it provides both edible oil and high-quality vegetable protein at a scale that strongly influences global food systems and agricultural markets (Messina, 2022). Its seeds typically contain about 18%-22% oil and 36%-42% protein, making soybean unusual among major crops in combining high energy density with substantial protein yield in a single commodity. This dual value underpins its broad use in human foods, animal feed, and industrial applications, including biodiesel and a wide range of processed products (Hamza et al., 2024). Soybean also occupies a strategic position in efforts to improve the sustainability of protein and oil supply, because demand for vegetable oils and plant-derived proteins continues to rise globally. For these reasons, improving soybean seed composition is not only a breeding objective but also an important challenge for food security, industrial raw material supply, and the design of more efficient crop systems (Duan et al., 2023).

The biological importance of soybean seed protein and oil accumulation lies in their central roles as the principal storage reserves that support germination, early seedling growth, and final seed quality. In mature soybean seed, protein and oil together account for almost 60% of total storage matter, and their composition is closely tied to seed size, nutritional value, and commercial worth (Duan et al., 2023). These reserves do not accumulate

randomly; rather, they reflect tightly coordinated developmental programs involving carbon and nitrogen allocation, fatty acid and triacylglycerol biosynthesis, and synthesis of major storage proteins such as β -conglycinin and glycinin. At the same time, oil and protein accumulation are linked by a persistent inverse relationship, indicating competition for shared metabolic precursors during seed development (Niu et al., 2025). Understanding when and how this trade-off is established is therefore essential for clarifying seed developmental physiology and for identifying routes to improve both traits simultaneously.

Research on soybean seed composition dynamics has shown that protein and oil contents change substantially over the course of development rather than appearing only as fixed mature-seed traits (Kambhampati et al., 2021). Early work established that oil percentage rises rapidly during mid-development, while large proportions of total mature protein and oil are synthesized during later filling stages even after percentage values begin to stabilize. Subsequent studies refined this developmental picture by showing that protein content often declines during the first weeks after flowering and then increases gradually, whereas oil accumulates rapidly at earlier stages, and sugars and oligosaccharides follow distinct temporal patterns toward maturity. Stage-based analyses further demonstrated that between reproductive stages R5 and R7, most mature dry weight is accumulated together with marked increases in protein, oil, and sugars, while after R7 moisture declines rapidly and stored-component composition changes relatively little. Collectively, these findings indicate that developmental timing is a major determinant of final seed composition and that the kinetics of reserve deposition must be considered explicitly in studies of soybean quality formation.

Despite substantial progress, important knowledge gaps remain in understanding the regulatory basis of dynamic protein and oil accumulation in soybean seeds. Recent genomic and transcriptomic studies have greatly expanded knowledge of the enzymes, transcription factors, and co-expression modules associated with reserve biosynthesis, especially during late developmental stages when divergence between high-oil and high-protein genotypes becomes most pronounced. Evidence now suggests that late maturation involves antagonistic activation of lipid-centered and nitrogen-centered pathways, providing a systems-level explanation for the oil-protein trade-off (Niu et al., 2025). However, many genetically mapped loci remain functionally unresolved, only a small number of candidate genes have been validated, and technical limitations in soybean transformation continue to slow mechanistic testing of proposed regulators (Duan et al., 2023). In addition, metabolic studies show that mature-seed composition can obscure important temporal shifts, including lipid decline and carbohydrate redistribution during maturation, meaning that endpoint measurements alone are insufficient to explain how seed composition is formed (Kambhampati et al., 2021). Accordingly, a focused analysis of dynamic changes in protein and oil contents during soybean seed development remains necessary to connect developmental stage, metabolic flux, and regulatory control, and to support breeding strategies aimed at improving seed quality without reinforcing the traditional trade-off between these two economically critical reserves.

2 Developmental Stages of Soybean Seeds and Physiological Regulation of Storage Compound Accumulation

2.1 Morphological and physiological characteristics during soybean seed development

Soybean seed development is commonly divided into lag, seed-filling, and maturation phases, and these phases can also be resolved morphologically from cotyledon formation through early maturity, mid-maturity, late maturity, and dry seed stages. Across these stages, seeds enlarge progressively, attain maximum size before desiccation, and then lose water rapidly as they approach physiological maturity, making morphology and water status reliable indicators of developmental progression. Physiologically, the major increase in seed mass and reserve deposition occurs mainly between R5 and R7, when moisture declines gradually but dry weight, protein, oil, and sugars all rise sharply. Cell division is largely completed by R4, whereas the pronounced increase in seed size from R5 to R6 is driven primarily by cell enlargement, during which carbon is partitioned simultaneously into protein, oil, and carbohydrate reserves (Islam et al., 2021).

2.2 Dynamic patterns of protein accumulation during seed development

Protein accumulation during soybean seed development is temporally regulated rather than continuous. Early stages contain relatively higher proportions of low-molecular-weight or nonstorage proteins, whereas the major storage proteins accumulate later as seeds enter active filling and maturation; correspondingly, storage-protein transcripts appear before visible protein deposition, indicating transcriptional control ahead of bulk reserve accumulation. The dominant storage fractions are 7S globulins and 11S glycinin, and their accumulation generally intensifies from mid seed filling toward late filling or maturation, although individual subunits differ in timing. More recent proteomic and compositional studies further show that 7S β -subunit and several 11S glycinin subunits increase steadily across filling stages, globulins remain higher than albumins throughout development, and globulin accumulation often peaks around R6 rather than rising linearly to maturity (Islam et al., 2021; Montanha et al., 2023).

2.3 Dynamic patterns of oil accumulation during seed development

Oil accumulation also shows a marked developmental pattern, beginning later than residual carbohydrate deposition and increasing rapidly during seed filling. A water-relations analysis indicated that residual accumulation starts first, followed by protein and then oil, while classical compositional studies showed that oil percentage rises sharply from about 24 to 40 days after flowering before stabilizing later in development. This rapid lipid deposition is supported by coordinated metabolic reprogramming during developing seed filling. Genes involved in carbon fixation, photosynthesis, glycolysis, and fatty acid biosynthesis are up-regulated during the critical period of oil accumulation, and later developmental transitions are associated with strong activation of carbohydrate degradation, triacylglycerol biosynthesis, and phospholipid signaling pathways, consistent with intensified flux toward storage lipid production in the late filling stages (Niu et al., 2025). Soybean seed development therefore involves a tightly staged transition from morphogenesis to reserve filling and finally desiccation, with protein and oil accumulation following distinct but overlapping temporal programs shaped by developmental state, carbon-nitrogen partitioning, and late-stage metabolic regulation.

3 Molecular Mechanisms Regulating Protein and Oil Biosynthesis in Developing Soybean Seeds

3.1 Genetic regulation of storage protein biosynthesis

Storage protein biosynthesis in developing soybean seeds is controlled by a hierarchical genetic program that integrates seed maturation regulators with structural genes encoding the major storage proteins. The principal storage proteins are β -conglycinin (7S) and glycinin (11S), and their accumulation is regulated not only by the expression of their own gene families but also by upstream AFL-type regulators, including LEC1, LEC2, FUS3, and ABI3, that coordinate the broader maturation state required for reserve deposition (Qi et al., 2026). Genetic mapping further shows that variation in storage protein composition is polygenic, with multiple loci affecting glycinin, β -conglycinin, total storage protein subunits, and the glycinin: β -conglycinin ratio, while polymorphism in the Gy1 promoter is specifically associated with 11S glycinin content (Zhang et al., 2021).

Recent transcriptomic and functional studies indicate that protein accumulation is also regulated through networks linking hormone signaling, nitrogen allocation, and storage protein processing. In contrasting soybean genotypes, candidate pathways associated with seed protein metabolism include photosynthesis, the TCA cycle, and starch and sucrose metabolism, and 40 days after flowering appears to be a critical stage for protein accumulation (Hu et al., 2025). At the gene level, *GmGASA12* acts as a molecular hub: its knockout increases water-soluble protein content, upregulates amino acid transporters and storage protein genes, and modulates the cooperative biosynthesis of β -conglycinin and glycinin through interaction with GmCG-6 (Yang et al., 2025).

3.2 Molecular regulation of oil biosynthesis and fatty acid metabolism

Oil biosynthesis in developing soybean seeds is governed by transcriptional circuits that channel carbon into fatty acid synthesis in plastids and triacylglycerol assembly in the endoplasmic reticulum. WRINKLED1 is a central regulator of this process, directly controlling numerous genes involved in lipid biosynthesis and participating in a positive feedback relationship with LEC1 that promotes the onset and stability of the fatty acid and TAG

accumulation program during seed maturation (Jo et al., 2024). Consistent with this framework, transcriptomic analyses of developing seeds show that rapid oil accumulation is accompanied by coordinated upregulation of carbon fixation, photosynthesis, glycolysis, and fatty acid biosynthesis pathways, indicating that lipid deposition depends on integrated activation of precursor supply and biosynthetic capacity.

Beyond the LEC1-WRI1 axis, soybean oil accumulation is fine-tuned by additional transcription factors and downstream metabolic enzymes. GmZF392 functions as a positive regulator of lipid production by activating lipid biosynthesis genes, acts synergistically with GmZF351, and is positioned downstream of GmNFYA within a three-factor regulatory module that enhances seed oil accumulation (Lu et al., 2021). A second layer of control involves protein-protein cooperation, as GmVOZ1A interacts with GmWRI1a, jointly upregulates GmACBP6a, and thereby promotes TAG accumulation, while broader network analyses in cultivated soybean also identify a BCCP2-SAD-FAD2-OBO/FA9 axis and a PLIP1-dependent pathway linked to enhanced oil biosynthesis (Yang et al., 2024; Niu et al., 2026).

3.3 Coordination between carbon and nitrogen metabolism

The balance between protein and oil accumulation in soybean seeds is shaped by competition for shared carbon skeletons and by developmental regulation of nitrogen assimilation. Sucrose imported into developing seeds is metabolized through glycolysis to produce intermediates such as acetyl-CoA and phosphoenolpyruvate, which support both fatty acid and amino acid biosynthesis, making carbon allocation a central determinant of storage compound partitioning (Qi et al., 2026). This shared metabolic dependence helps explain the widely observed inverse relationship between seed oil and protein, which reflects repartitioning among protein, lipid, and carbohydrate reserves rather than simple independent accumulation of each component (Qi et al., 2026).

Evidence from multi-omics, physiology, and transporter genetics shows that this trade-off is dynamically regulated late in seed development. High-oil and high-protein cultivars diverge mainly during later maturation, when lipid-centric pathways such as TAG synthesis and oil body biogenesis are antagonistically activated against nitrogen-centric pathways including nitrogen assimilation, amino acid metabolism, ABA signaling, and storage protein processing (Niu et al., 2025). Experimental manipulation of maternal carbon supply supports this model: reduced sugar delivery to embryos through disruption of *GmSWEET10a/b* or *GmSUT1* lowers expression of sucrose metabolism, fatty acid biosynthesis, and TAG assembly genes while inducing storage protein genes, shifting final seed composition toward lower oil and higher protein (Sun et al., 2025). In developing soybean seeds, protein and oil biosynthesis are therefore regulated by distinct but interconnected genetic programs, with storage protein accumulation tied closely to maturation and nitrogen-responsive networks, oil accumulation driven by LEC1/WRI1-centered lipid circuits, and the final balance between the two determined by developmental carbon-nitrogen partitioning.

4 Environmental and Agronomic Factors Affecting Protein and Oil Accumulation Dynamics

4.1 Effects of temperature and climate conditions

Temperature during seed filling strongly shifts the balance between soybean seed protein and oil accumulation, although the response is not strictly linear across environments. Meta-analysis indicates that high temperature generally increases final protein concentration, but this increase does not reflect greater absolute protein synthesis; instead, it arises because other seed fractions, especially oil, are reduced more strongly during stressful filling conditions. Controlled-environment studies are consistent with this pattern, showing that exposure to 35°C during seed fill increased seed protein by 4.0 percentage points and decreased oil by 2.6 percentage points relative to 29°C (Figure 1) (Sun et al., 2025).

The climatic control of composition is also modified by the broader thermal regime and by interactions among multiple weather variables. Large-scale sampling in China showed that crude protein was positively associated with accumulated and mean temperature, whereas crude oil showed the opposite trend overall, but oil increased with mean daily temperature when temperatures remained below 19.7°C, indicating a threshold response rather than a simple monotonic one. Developmental analyses likewise showed that high temperature accelerated maturity,

advanced the timing of maximum oil concentration, and increased the rate of late-stage protein accumulation, whereas low temperature prolonged seed growth but reduced growth rate, demonstrating that climate alters both the rate and timing of reserve deposition.

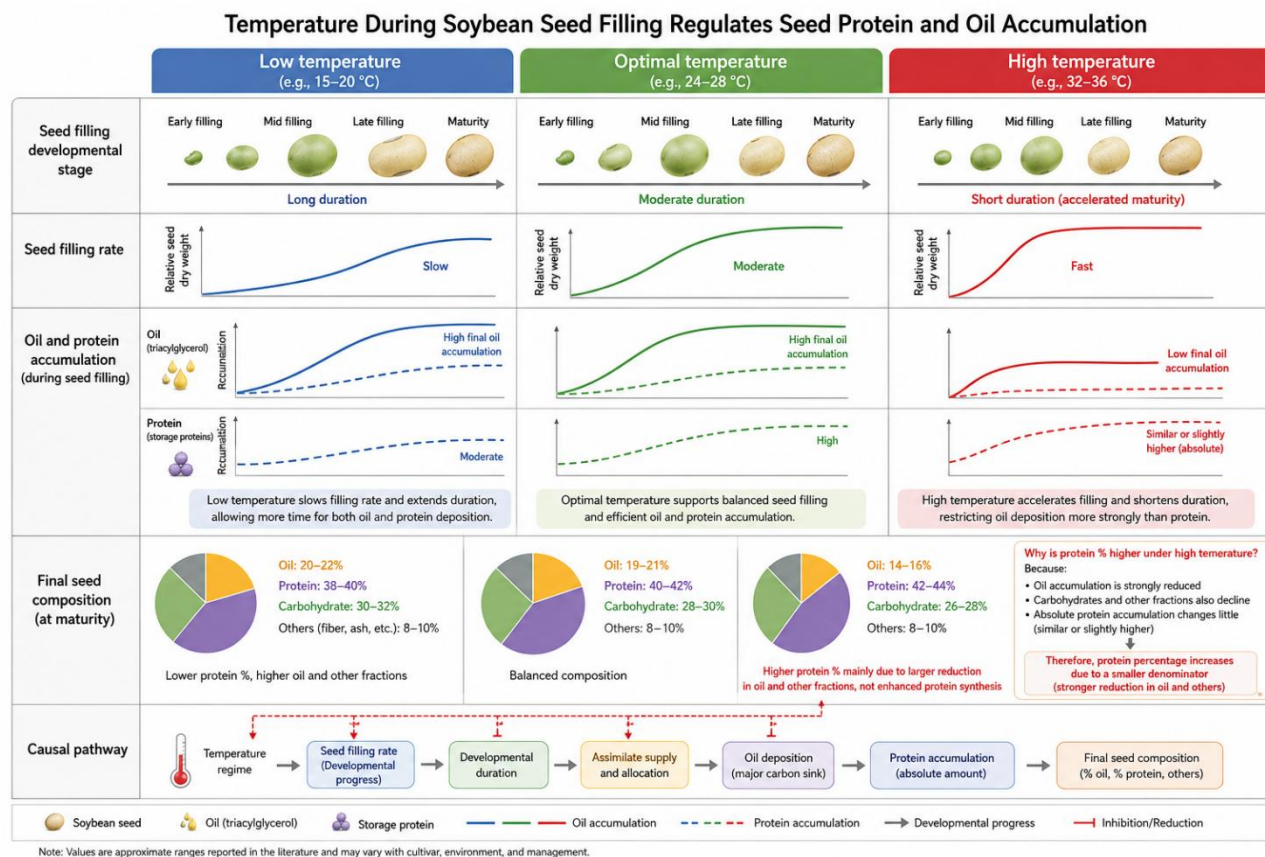


Figure 1 Conceptual framework illustrating the effects of temperature during soybean seed filling on protein and oil accumulation

4.2 Effects of water availability and drought stress

Water limitation during seed development alters both final composition and the temporal pattern of reserve accumulation, but its effect depends strongly on the developmental stage at which stress occurs. A broad meta-analysis found that water stress reduced the absolute content of protein, oil, and residual fractions per seed, yet protein accumulation was less affected than oil accumulation, leading to a higher final protein concentration on a dry-weight basis (Mo et al., 2024). Complementing this general result, mechanistic work showed that drought can change both the rates and durations of component accumulation while maintaining overall seed growth rate, with compensation typically favoring protein because late seed filling can rely more on remobilizable nitrogen whereas oil synthesis depends more heavily on current photosynthesis.

Field studies show that reproductive-stage drought is particularly important for soybean quality because it shifts partitioning away from oil and toward protein while simultaneously reducing yield. In Brazil, water deficit imposed during the reproductive stage consistently produced higher protein and lower oil contents across genotypes, whereas vegetative-stage stress had much weaker compositional effects. However, drought does not act independently of temperature: analyses from Argentine multi-environment trials showed that under stronger water deficit, oil concentration increased with increasing mean temperature whereas protein still increased with temperature but decreased with water deficit, underscoring that climate effects on seed composition should be interpreted as joint heat-water responses rather than isolated factors.

4.3 Effects of nutrient management and cultivation practices

Among agronomic factors, nutrient supply-especially nitrogen-has the most consistent documented effect on seed composition, because nitrogen availability during seed filling directly constrains protein deposition. Across a large

U.S. synthesis, low to moderate fertilizer N inputs increased both protein and oil concentration, although environmental variation still explained most of the total variation in composition. Field experiments in high-yield environments similarly showed that full-season nitrogen supply increased seed protein concentration without increasing oil concentration, while raising both protein and oil yields through higher seed production, indicating that N limitation can restrict both composition and total reserve output.

Management effects beyond nitrogen are real but less uniform, and they often interact with water regime, latitude, and cropping system. Recent field studies showed that late-season N application during seed filling increased protein concentration by 1.2-2.8% with little or no effect on oil concentration, supporting targeted N management as a practical strategy for improving meal quality (Khatri et al., 2026). By contrast, broader management syntheses indicate that delayed planting decreases oil concentration, corn-soybean rotation tends to improve composition and yield, and practices such as no-till, seed treatment, foliar nutrient application, and fungicide produce mixed responses, showing that cultivation practices mainly modify composition indirectly through their effects on crop growth environment and stress exposure.

5 Advanced Technologies for Monitoring and Predicting Protein and Oil Dynamics

5.1 Biochemical and omics approaches for understanding seed composition

Biochemical and omics approaches have become the main tools for dissecting how protein and oil contents change during soybean seed development because they capture regulatory variation across transcripts, proteins, and metabolites rather than only final composition. Integrated transcriptomic, proteomic, and metabolomic studies show that the pathways most consistently linked to protein and oil divergence involve glycolysis and carbon metabolism, while systems-level analyses further indicate that metabolic flux mapping during seed fill is especially useful for connecting developmental gene expression with the biochemical networks that determine mature seed composition (Mo et al., 2024). These datasets also reveal that different omics layers contribute distinct kinds of information, which is important for interpreting developmental dynamics. Joint transcriptome-proteome analysis identified generally poor correspondence between mRNA and protein abundance, while metabolomics studies in contrasting high-protein and high-oil lines detected coordinated shifts in the Calvin cycle, TCA cycle, and glycolysis that favor routing carbon into amino acid and fatty acid synthesis (Mo et al., 2024; Cui et al., 2025).

Proteomics has been particularly valuable for defining the temporal sequence of storage reserve accumulation during seed filling. High-resolution proteome mapping across 2 to 6 weeks after flowering showed a developmental decrease in metabolism-related proteins together with an increase in proteins associated with destination and storage, and more recent TMT-based proteomics further demonstrated that major 7S and 11S storage proteins accumulate steadily from early to late seed filling (Islam et al., 2021). At a broader scale, integrative omics has strengthened gene discovery for seed quality improvement by linking developmental expression patterns to inherited composition loci. Meta-analysis and time-course transcriptomics identified 11 shared meta-QTL hot regions for oil and protein and seven hub genes associated with storage accumulation, while multi-omics comparison among contrasting cultivars detected 22 candidate genes with potential simultaneous negative regulation of protein and oil content (Mo et al., 2024).

5.2 Imaging and phenotyping technologies for seed development monitoring

Non-destructive imaging technologies now allow soybean seed composition and developmental status to be monitored at the single-seed level, which is a major advance over destructive wet-chemistry assays. Near-infrared hyperspectral imaging predicted single-seed protein content with $R^2 = 0.92$ and RMSE of 1.08%, and it also generated chemical maps that visualized within-seed protein distribution for rapid screening of large numbers of samples (Aulia et al., 2022). Hyperspectral approaches are also effective for monitoring oil quality traits that contribute to overall lipid dynamics during development and selection. Reflective hyperspectral imaging classified oleic and linoleic acid contents of single seeds with validation accuracies of 90% and 93.3%, and the study emphasized that single-seed measurement is critical because bulk-seed spectroscopy cannot resolve the composition of individual seeds needed for precision breeding (Fu et al., 2021).

Imaging-based phenotyping is not limited to chemical traits, because seed architecture and visible phenotype can be quantified at scales that support genetic analysis and breeding decisions. High-throughput image analysis measured approximately 39 065 seeds from 400 lines for morphology and color traits, and the resulting quantitative phenotype data were proposed as useful inputs for GWAS and other gene-discovery efforts (Baek et al., 2020). More broadly, soybean seed monitoring increasingly draws on complementary sensor platforms rather than any single imaging modality. Reviews of soybean seed imaging identify radiography, magnetic resonance imaging, multispectral imaging, chlorophyll fluorescence imaging, infrared thermography, and computerized seedling analysis as viable tools for detecting incomplete maturation, structural injury, and other quality-related changes, while RGB image-based phenotyping has emerged as a practical, lower-cost route for capturing seed traits at phenome scale (Duc et al., 2023; França-Silva et al., 2023).

5.3 Machine learning and modeling approaches for predicting seed quality

Machine learning approaches now make it possible to predict soybean seed protein and oil before harvest by combining field observations with spectral or environmental data. In on-farm prediction using satellite imagery, XGBoost outperformed other algorithms and achieved absolute errors of 1.80% for protein and 1.04% for oil, with the best model timing occurring within one week after the peak of the green chlorophyll vegetation index (Hernández et al., 2023). Deep learning has also shown that time-series crop imagery can recover useful reproductive-stage signals for seed composition, although performance remains trait-dependent. Using PlanetScope imagery, recurrent neural network models identified later reproductive-stage vegetation indices as the most informative features, and the GRU model achieved the best reported performance in that study for protein and oil prediction from standing crops (Sarkar et al., 2023).

For breeding applications, prediction models are moving beyond field-level remote sensing toward integrated genomic, phenomic, and environmental frameworks. Comparative prediction studies found that phenomic prediction outperformed genomic prediction for seed yield, whereas genomic prediction performed better for seed protein and oil, and larger deep-learning frameworks using multi-decadal weather plus genotype and management information showed that pretrained temporal representations from yield data could be transferred to downstream oil and protein prediction tasks (Van Der Laan et al., 2024). An important recent trend is the emphasis on interpretability and efficient feature selection, because predictive accuracy alone is not sufficient for biological insight or breeding utility. In genomic prediction across 1 110 accessions, XGBoost or random forest outperformed deep learning for 13 of 14 prediction tasks and enabled up to 90% marker reduction without loss of performance, while Transformer-based models highlighted influential periods and variables during the growing season through self-attention, improving explainability for seed composition prediction (Gill et al., 2022; Ayanlade et al., 2026).

6 Case Study: Dynamic Changes in Protein and Oil Accumulation under Different Genotypes and Environmental Conditions

6.1 Comparative analysis of protein and oil accumulation among soybean varieties

Comparative studies consistently show that soybean genotypes differ not only in final seed composition but also in the trajectory of reserve deposition during development. In a direct developmental comparison, the high-protein cultivar NPS233 accumulated more protein and less oil than the high-oil cultivar NPS301 at all four sampled stages from 7 to 28 days after flowering, while a broader study of seven specialty genotypes found that although developmental trends were partly shared, substantial genotype-specific differences in protein and oil composition remained evident throughout seed maturation (Xu et al., 2022). This variation extends across diverse germplasm: in 320 resequenced accessions, protein ranged from 37.8% to 46.5% and oil from 16.7% to 22.6%, confirming that soybean varieties span a wide compositional spectrum that provides a practical basis for selecting contrasting developmental phenotypes (Jin et al., 2023).

Genotypic divergence also reflects distinct physiological and molecular allocation strategies rather than simple differences in endpoint composition. Transcriptomic analysis across two high-oil and two high-protein cultivars showed that transcriptional divergence was concentrated in late maturation, when high-oil genotypes activated

lipid-centered pathways and high-protein genotypes activated nitrogen-centered pathways, whereas proteomic comparison of two contrasting varieties indicated that compositional differences were driven mainly by the peripheral proteome, which changed strongly across developmental stages (Huang et al., 2025; Niu et al., 2025). Importantly, this trade-off is not absolute in all materials: under heat and drought stress, several high-protein lines maintained 46%-55% protein, and two lines showed no apparent negative relationship between seed protein and oil contents, indicating that some varieties partially escape the usual antagonism between these traits (Figure 2) (Kakati et al., 2024).

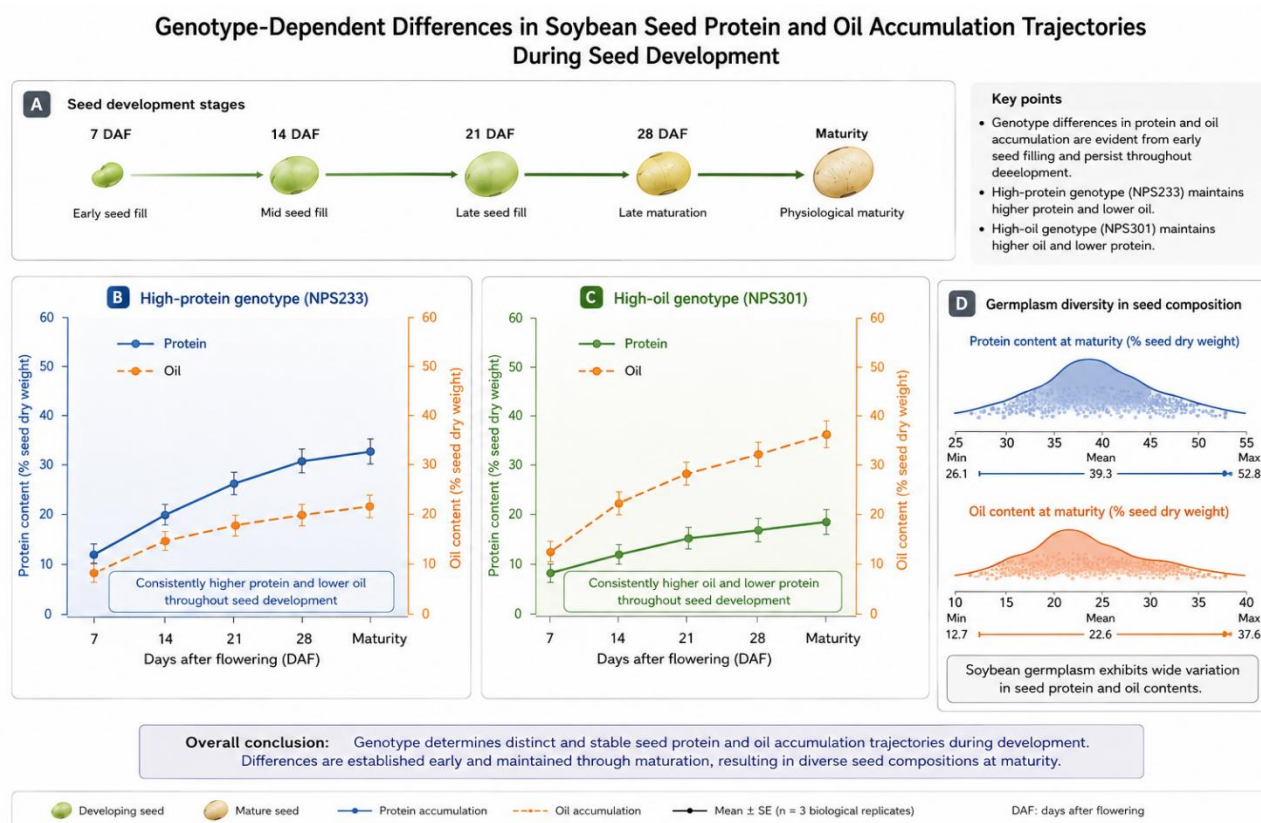


Figure 2 Developmental trajectories of seed protein and oil accumulation in contrasting soybean genotypes

6.2 Effects of environmental stress on seed composition dynamics: a case study

Environmental stress alters the rate and balance of storage compound accumulation, but the direction and magnitude of change depend on the stress scenario and genotype. A meta-analysis found that water stress reduced the absolute accumulation of protein, oil, and residual fractions per seed, yet protein was less affected than oil, which increased final protein concentration on a dry-weight basis; similarly, controlled experiments showed that severe drought increased protein by 4.4 percentage points and decreased oil by 2.9 percentage points during seed fill. Temperature acts in a comparable but not identical way, because high temperature also tends to raise protein concentration without increasing protein synthesis per se, indicating that stress often changes composition through differential suppression of reserve classes rather than through direct stimulation of protein deposition.

Case studies across multiple environments show that these responses are strongly modified by genotype × environment interaction. In Croatia, 32 elite genotypes grown across six dry and eight normal environments showed significant effects of genotype, environment, and their interaction for both traits; drought decreased mean protein by 4.5% and slightly increased oil by 1.2%, and the usual negative protein-oil correlation seen in normal environments disappeared under dry conditions. At larger scale, U.S. synthesis-analysis showed that environment accounted for more than 70% of the variation in seed composition, while management-induced environmental shifts such as delayed planting reduced oil concentration, reinforcing that stress effects on accumulation dynamics emerge from the combined action of weather, timing, and genotype rather than from a single factor alone.

6.3 Integrating multi-omics and phenotyping data to explain seed quality formation

Integrated omics studies now explain varietal differences in seed quality formation by linking developmental phenotypes to coordinated changes in transcripts, metabolites, and candidate regulators. In two contrasting cultivars, combined transcriptomics and metabolomics identified more than 12,000 differentially expressed genes and 315 differential metabolites across seed development, and the authors proposed that high-protein varieties differ from high-oil types through altered sensitivity to desiccation, photomorphogenesis, and senescence timing; in a separate nine-cultivar multi-omics study, more than 1,000 differential metabolites and 7 000 differentially expressed genes were used to define 10 core genes associated with oil content, with *GmADH1* and *GmCrRLK1L34* emerging as hub regulators (Xu et al., 2022). These integrative datasets also converge on central metabolism, because metabolomic analysis of extreme high- and low-protein/oil lines showed enhanced Calvin cycle, TCA cycle, and glycolytic activity that supports carbon entry into amino acid and fatty acid synthesis, helping explain how developmental resource partitioning generates contrasting seed compositions (Cui et al., 2025).

High-resolution phenotyping adds spatial and quantitative context that bulk omics alone cannot provide. In wild soybean, spatial transcriptomics, spatial metabolomics, and single-cell RNA sequencing at mid-maturity showed tissue-level separation between protein- and lipid-associated metabolism and identified *GsMAPK23-4* as a candidate regulator of seed quality, while FT-NIR phenotyping across 191 diverse accessions produced highly accurate protein and oil prediction models with R^2 values above 95%, enabling rapid compositional screening across maturity groups and seed phenotypes. Genetic integration strengthens this framework further: QTL mapping, BSA-seq, and RNA-seq identified 37 QTLs and 12 preliminarily validated candidate genes for protein and oil, showing that multi-omics and phenotyping together can move from descriptive developmental patterns to testable loci and breeding targets (Fang et al., 2025).

7 Breeding Strategies and Future Perspectives for Improving Soybean Protein and Oil Quality

7.1 Genetic improvement of soybean protein and oil traits

Genetic improvement of soybean protein and oil traits still depends on treating these characters as complex quantitative phenotypes shaped by many loci and strong genotype $\times$ environment interaction. QTL mapping, GWAS, and meta-QTL analysis have repeatedly identified stable genomic regions for both traits, and the narrower intervals produced by GWAS now make marker-assisted allele selection more practical for predictable compositional improvement (Kumar et al., 2021).

Breeding strategies are also shifting from locus discovery alone to predictive and design-based selection. Genomic selection achieved cross-validation accuracies of 0.68 for protein and 0.64 for oil, indicating that breeders can identify many top-quartile composition lines before extensive phenotyping (Miller et al., 2023). At the same time, integrative genetics shows that some quality traits can be improved together rather than traded off absolutely, because seed weight and oil content share a positive genetic correlation and candidate genes such as *GmRWOS1* have already been functionally validated for coordinated trait regulation (Yuan et al., 2024).

7.2 Optimizing agronomic management for enhanced seed quality

Agronomic management can partly buffer quality losses, but environment remains the dominant driver of soybean seed composition across production systems. In a synthesis of 13 574 U.S. data points, site-year explained more than 70% of the variation in protein, oil, and yield, which means management usually works by shifting how crops experience temperature, radiation, water, and nitrogen during seed filling. Within that constraint, delayed planting consistently reduced oil concentration, whereas corn-soybean rotation improved both composition and yield, showing that timing and system context can move quality in useful directions.

More targeted interventions can improve seed quality without necessarily imposing the usual yield penalty. Lower fertilizer N rates increased both oil and protein concentration in the U.S. synthesis, and irrigation in water-limited Nebraska fields increased yield and protein concentration simultaneously in about two-thirds of fields, especially where conditions were not favorable for oil synthesis (Carciochi et al., 2023). Regional field studies further show

that early planting tends to increase oil, late planting tends to increase protein, and improved nutrition at sowing can raise both yield and seed protein concentration, although the response depends on region and field productivity (Di Mauro et al., 2023).

7.3 Future research directions in dynamic regulation of seed composition

Future progress depends on resolving the developmental mechanisms that produce the protein-oil trade-off rather than only selecting around its final outcome. Current reviews argue that breeders need seed development-based studies using mutants, multi-omics, and metabolic flux analysis to explain how soybean seeds rebalance protein, oil, and sucrose during filling (Kumar et al., 2025). This need is reinforced by evidence that only a few soybean-specific oil and protein regulators have been functionally characterized so far, despite the likely existence of many crop-specific transporters, proteases, and transcription factors that are not predictable from Arabidopsis homologs alone.

The most promising future breeding framework combines multi-omics-guided target discovery with precise editing and broader use of untapped diversity. Spatial transcriptomics and metabolomics in wild soybean already show cell-type separation between protein- and lipid-associated metabolism and identify regulators such as *GsMAPK23-4* that alter amino acid and protein accumulation. At the translational end, gene editing now extends beyond natural alleles: DNA-free and AI-assisted editing pipelines are being developed for genotype-independent improvement, and AlphaFold-guided editing of *GmSWEET10a/b* has already increased oil or protein in multi-year, multi-site field trials without reducing yield (Wang et al., 2025; Kim et al., 2026).

8 Conclusions

Protein and oil accumulation in developing soybean seeds follow coordinated but often antagonistic temporal patterns that are shaped by shared metabolic resources and stage-specific regulation. Developmental studies show that the inverse association between mature seed protein and oil reflects cumulative changes across seed filling rather than a single static relationship, and late seed development is especially important because lipid content can decline during maturation while carbohydrates increase as maternal nutrient supply diminishes. Complementing this metabolic view, transcriptomic network analysis across contrasting high-oil and high-protein cultivars indicates that the major transcriptional divergence appears in later developmental stages, when lipid-centered and nitrogen-centered programs become antagonistically activated. These dynamic patterns also help explain why soybean seed composition is difficult to improve through endpoint-based selection alone. Comparative metabolomics of extreme phenotypes identified key intermediates such as glucose, citric acid, and α -ketoglutarate, and showed increased Calvin cycle, TCA cycle, and glycolytic activity in lines with strong protein or oil accumulation, supporting a model in which carbon rerouting underlies reserve deposition. Proteomic analysis further shows that differences in seed oil and protein content arise largely from the peripheral proteome, which fluctuates strongly across seed developmental stages rather than from a fixed constitutive protein set.

For soybean quality improvement, the central implication is that protein and oil should be treated as complex quantitative traits with partially shared but partly separable genetic control. Genome-wide studies have identified many loci affecting seed composition, including 87 chromosomal regions in one diverse panel and additional candidate genes involved in nitrogen fixation, amino acid biosynthesis, and fatty acid metabolism, providing a substantial marker base for breeding. More recent resequencing-based GWAS likewise detected 23 loci for protein and 29 for oil, including multiple new regions and nine candidate genes related to biosynthesis, transport, signaling, and development, reinforcing the feasibility of marker-assisted improvement. At the same time, breeding strategies must account for persistent trade-offs among composition, yield, and adaptation. Historical analyses indicate that selection for yield has reduced seed protein concentration by 0.06% per year while increasing residual low-energy fractions, and century-scale surveys in China found declining protein concentrations with largely stable oil content across decades of cultivar release. Conventional breeding therefore remains constrained by the negative correlation between protein and oil, but newer work suggests that simultaneous improvement is possible when breeders target independent loci, pleiotropic regulators, or specific genomic regions that affect one trait without penalizing the other.

Future research should move beyond descriptive composition analysis toward developmentally resolved, mechanism-based design of soybean seed quality. Reviews of soybean functional genomics emphasize that hundreds of QTLs have already been identified, but relatively few genes have been functionally validated, and progress now depends on integrating genomics, transcriptomics, proteomics, and transformation technologies more effectively. A parallel design-oriented synthesis argues that resolving the protein-oil trade-off will require seed development-focused mutant analysis, multi-omics integration, and isotope-based metabolic flux studies that can capture rebalancing among protein, oil, and sucrose during filling.

Sustainable soybean production will also require linking seed quality improvement with climate resilience and resource efficiency. Soybean already contributes to sustainable agriculture through biological nitrogen fixation, but future breeding must address the fact that abiotic stress disrupts seed filling, shifts the balance among protein, oil, and fatty acids, and reduces protein yield per hectare even when concentration rises under drought or heat. The most promising roadmap combines climate-resilient breeding, optimized source-sink balance and nitrogen fixation, wider use of wild and diverse germplasm, and predictive approaches that incorporate nonlinear environmental effects on seed composition across regions and maturity groups.

Acknowledgments

I extend my sincere gratitude to the anonymous reviewers for their valuable and insightful comments, which have greatly strengthened this paper.

Conflict of Interest Disclosure

The author affirms that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Aulia R., Kim Y., Amanah H.Z., Andi A.M.A., Kim H., Kim H., Lee W.H., Kim K.H., Baek J.H., and Cho B.K., 2022, Non-destructive prediction of protein contents of soybean seeds using near-infrared hyperspectral imaging, *Infrared Physics and Technology*, 127: 104365.
<https://doi.org/10.1016/j.infrared.2022.104365>
- Ayanlade T.T., Van der Laan L., Liu Q., Gangopadhyay T., Shook J., Singh A., Ganapathysubramanian B., Sarkar S., and Singh A.K., 2026, Transformer model to determine spatio-temporal relationships of variables, and interpretability for soybean seed yield, oil, and protein prediction, *Frontiers in Artificial Intelligence*, 9: 1750108.
<https://doi.org/10.3389/frai.2026.1750108>
- Baek J.H., Lee E., Kim N., Kim S.L., Choi I., Ji H., Chung Y.S., Choi M.S., Moon J.K., and Kim K.H., 2020, High throughput phenotyping for various traits on soybean seeds using image analysis, *Sensors*, 20(1): 248.
<https://doi.org/10.3390/s20010248>
- Bu M., Zhang Y., Xu W., Li Y., Yu H., Zhang Y., Yang S., Bhat J.A., and Feng X., 2026, Genome-wide detection of superior haplotypes for seed oil and protein content in Northeast China soybean (*Glycine max* L.) germplasm, *Frontiers in Plant Science*, 17: 1767299.
<https://doi.org/10.3389/fpls.2026.1767299>
- Carciochi W.D., Grassini P., Naeve S., Specht J.E., Mamo M., Seymour R., Nygren A., Mueller N., Sivits S., Proctor C., Rees J., Whitney T., and La Menza N.C., 2023, Irrigation increases on-farm soybean yields in water-limited environments without a trade-off in seed protein concentration, *Field Crops Research*, 304: 109163.
<https://doi.org/10.1016/j.fcr.2023.109163>
- Chiozza M.V., Shook J.M., Van der Laan L., Singh A.K., and Miguez F.E., 2025, Comprehensive assessment of soybean seed composition from field trials spanning 22 US states and 24 years: Predictive insights, *Crop Science*, 65(4): e70142.
<https://doi.org/10.1002/csc2.70142>
- Cui Y., Wang Z., Li M., Li X., Wang S., Liu C., Xin D., Qi Z., Chen Q., Yang M., and Zhao Y., 2026, Comparative metabolomics analysis of seed composition accumulation in soybean (*Glycine max* L.) differing in protein and oil content, *Plant, Cell and Environment*, 49(7): 3925-3941.
<https://doi.org/10.1111/pce.15448>
- Di Mauro G., Schwalbert R., Alvarez Prado S., Saks M.G., Ramirez H., Costanzi J., and Parra G., 2023, Exploring practical nutrition options for maximizing seed yield and protein concentration in soybean, *European Journal of Agronomy*, 146: 126794.
<https://doi.org/10.1016/j.eja.2023.126794>
- Duan Z., Li Q., Wang H., He X., and Zhang M., 2023, Genetic regulatory networks of soybean seed size, oil and protein contents, *Frontiers in Plant Science*, 14: 1160418.
<https://doi.org/10.3389/fpls.2023.1160418>
- Duc N.T., Ramlal A., Rajendran A., Raju D., Lal S.K., Kumar S., Sahoo R.N., and Chinnusamy V., 2023, Image-based phenotyping of seed architectural traits and prediction of seed weight using machine learning models in soybean, *Frontiers in Plant Science*, 14: 1206357.
<https://doi.org/10.3389/fpls.2023.1206357>

- Islam N., Krishnan H.B., and Natarajan S., 2022, Quantitative proteomic analyses reveal the dynamics of protein and amino acid accumulation during soybean seed development, *PROTEOMICS*, 22(7): 2100143.
<https://doi.org/10.1002/pmic.202100143>
- Jin H., Yang X., Zhao H., Song X., Tsvetkov Y., Wu Y., Gao Q., Zhang R., and Zhang J., 2023, Genetic analysis of protein content and oil content in soybean by genome-wide association study, *Frontiers in Plant Science*, 14: 1182771.
<https://doi.org/10.3389/fpls.2023.1182771>
- Jo L., Pelletier J., Goldberg R.B., and Harada J.J., 2024, Genome-wide profiling of soybean WRINKLED1 transcription factor binding sites provides insight into seed storage lipid biosynthesis, *Proceedings of the National Academy of Sciences of the United States of America*, 121(45): e2415224121.
<https://doi.org/10.1073/pnas.2415224121>
- Kakati J., Fallen B., Armstrong P.R., Yan S., Bridges W.C., and Narayanan S., 2024, High-protein soybean lines with stable seed protein content under heat and drought stresses, *Journal of Agriculture and Food Research*, 18: 101469.
<https://doi.org/10.1016/j.jafr.2024.101469>
- Kambhampati S., Aznar-Moreno J., Bailey S.R., Arp J.J., Chu K.L., Bilyeu K., Durrett T., and Allen D.K., 2021, Temporal changes in metabolism late in seed development affect biomass composition, *Plant Physiology*, 186(2): 874-890.
<https://doi.org/10.1093/plphys/kiab116>
- Khatri D., Magar L., Poudel S., Kc S., Gebremedhin M., Lucas S., and Chilwal A., 2026, Biochar and late-season nitrogen fertilization effects on soybean yield and seed quality, *Journal of Agriculture and Food Research*, 2026: 102941.
<https://doi.org/10.1016/j.jafr.2026.102941>
- Kim H., Chae J., Han S., Kim J.H., Chung Y.S., Karthik S., and Heo J.B., 2026, AI-guided DNA-free and genotype-independent genome editing for soybean improvement, *Plants*, 15(13): 2080.
<https://doi.org/10.3390/plants15132080>
- Kumar R., Mulkey S., Shelake R.M., Combs-Giroir R., Mukherjee T., Allen D.K., Clemente T., Stacey M., Lorenz A.J., and Stupar R.M., 2025, Targets and strategies to design soybean seed composition traits, *The Plant Genome*, 18(4): e70115.
<https://doi.org/10.1002/tpg2.70115>
- Kumar V., Vats S., Kumawat S., Bisht A., Bhatt V.D., Shivaraj S.M., Padalkar G., Goyal V., Zargar S., Gupta S., Kumawat G., Chandra S., Chalam V.C., Ratnaparkhe M., Gill B., Jean M., Patil G., Vuong T., Rajcan I., Sonah H., and collaborators, 2021, Omics advances and integrative approaches for the simultaneous improvement of seed oil and protein content in soybean (*Glycine max* L.), *Critical Reviews in Plant Sciences*, 40(5): 398-421.
<https://doi.org/10.1080/07352689.2021.1954778>
- Lu L., Wei W., Li Q.T., Bian X., Lu X., Hu Y., Cheng T., Wang Z., Jin M., Tao J.J., Yin C., He S.J., Man W., Li W., Lai Y.C., Zhang W.K., Chen S., and Zhang J., 2021, A transcriptional regulatory module controls lipid accumulation in soybean, *New Phytologist*, 231(2): 661-678.
<https://doi.org/10.1111/nph.17401>
- Messina M., 2022, Perspective: Soybeans can help address the caloric and protein needs of a growing global population, *Frontiers in Nutrition*, 9: 909464.
<https://doi.org/10.3389/fnut.2022.909464>
- Miller M.J., Song Q., and Li Z., 2023, Genomic selection of soybean (*Glycine max*) for genetic improvement of yield and seed composition in a breeding context, *The Plant Genome*, 16(4): e20384.
<https://doi.org/10.1002/tpg2.20384>
- Mo W., Wang P., Shi Q., Zhao X., Zheng X., Ji L., Zhang L., Geng M., Wang Y., Wang R., Bian M., Meng X., Zuo Z., and Yang Z., 2024, Uncovering key genes associated with protein and oil in soybeans based on transcriptomics and proteomics, *Industrial Crops and Products*, 222: 119981.
<https://doi.org/10.1016/j.indcrop.2024.119981>
- Montanha G., Mendes N.A.C., Perez L.C., Cunha M.L.O., Santos E., Pérez C.A., De Almeida E.L., Marques J.P.R., Umburanas R.C., Linhares F.S., Reis A.R.D., Sabatini S., and De Carvalho H.D., 2023, Unfolding the dynamics of mineral nutrients and major storage protein fractions during soybean seed development, *ACS Agricultural Science and Technology*, 3(8): 666-674.
<https://doi.org/10.1021/acsagscitech.3c00125>
- Nawaz M.A., Chung G., Pamirsky I.E., and Golokhvast K., 2026, Breeding climate-resilient soybeans for 2050 and beyond: Leveraging novel technologies to mitigate yield stagnation and climate change impacts, *Plants*, 15(8): 1201.
<https://doi.org/10.3390/plants15081201>
- Niu Y., Wang W., Wang X., Li H., Jin Y., Qi B., Zhao H., Huang Z., Yan F., Fan S., Zhang G., Mock H.P., Li J., Zhao Q., Huang Y., and Zhang F., 2026, Transcriptomic signatures of developing soybean seeds reveal the molecular mechanisms of oil accumulation during domestication, *Plant Biology*, 28(4): 1062-1076.
<https://doi.org/10.1111/plb.70194>
- Niu Y., Wu J., Li H., Wang H., Zhao H., Huang Z., Yan F., and Zhang G., 2025, Construction of regulatory networks related to oil and protein accumulation in developing soybean seeds, *Plant Growth Regulation*, 105(5): 1605-1621.
<https://doi.org/10.1007/s10725-025-01356-w>
- Patel J., Patel S., Cook L., Fallen B.D., and Koebernick J., 2025, Soybean genome-wide association study of seed weight, protein, and oil content in the southeastern USA, *Molecular Genetics and Genomics*, 300(1): 43.
<https://doi.org/10.1007/s00438-025-02228-8>

- Qi H., Han X., Huang J.-F., Wu X., and Han J., 2026, Integrated transcriptomic, proteomic, and metabolomic analysis of a chromosome segment substitution line reveals the regulatory mechanism governing fatty acids and storage proteins in soybean seeds, *Genes*, 17(4): 432.
<https://doi.org/10.3390/genes17040432>
- Sarkar S., Sagan V., Bhadra S., Rhodes K., Pokharel M., and Fritschi F.B., 2023, Soybean seed composition prediction from standing crops using PlanetScope satellite imagery and machine learning, *ISPRS Journal of Photogrammetry and Remote Sensing*, 204: 257-274.
<https://doi.org/10.1016/j.isprsjprs.2023.09.010>
- Sun J., Li W., Wei X., Shou H., Tran L.-S.P., Feng X., and Wang S., 2025, Mechanistic roles of *GmSWEET10a/b* and *GmSUT1* in the oil-protein balance in soybean mature seeds at transcriptional and metabolic levels, *The Plant Journal*, 123(4): e70435.
<https://doi.org/10.1111/tpj.70435>
- Tamagno S., Sadras V.O., Aznar-Moreno J., Durrett T.P., and Ciampitti I.A., 2022, Selection for yield shifted the proportion of oil and protein in favor of low-energy seed fractions in soybean, *Field Crops Research*, 279: 108446.
<https://doi.org/10.1016/j.fcr.2022.108446>
- Van Der Laan L., Parmley K.A., Saadati M., Pacin H.T., Panthulugiri S., Sarkar S., Ganapathysubramanian B., Lorenz A.J., and Singh A.K., 2025, Genomic and phenomic prediction for soybean seed yield, protein, and oil, *The Plant Genome*, 18(1): e70002.
<https://doi.org/10.1002/tpg2.70002>
- Wang J., Zhang L., Wang S., Wang X., Li S., Gong P., Bai M., Paul A., Tvedt N., Ren H., Yang M., Zhang Z., Zhou S., Sun J., Wu X., Kuang H., Du Z., Dong Y., Shi X., Guan Y., and collaborators, 2025, AlphaFold-guided bespoke gene editing enhances field-grown soybean oil contents, *Advanced Science*, 12(23): 2500290.
<https://doi.org/10.1002/adv.202500290>
- Xu W., Wang Q., Zhang W., Zhang H., Liu X., Song Q., Zhu Y., Cui X., Chen X., and Chen H., 2022, Using transcriptomic and metabolomic data to investigate the molecular mechanisms that determine protein and oil contents during seed development in soybean, *Frontiers in Plant Science*, 13: 1012394.
<https://doi.org/10.3389/fpls.2022.1012394>
- Yang M., Du C., Li M., Wang Y., Bao G., Huang J., Zhang Q., Zhang S., Xu P., Teng W., Li Q., Liu S., Song B., Yang Q., and Wang Z., 2024, The transcription factors GmVOZ1A and GmWRI1 synergistically regulate oil biosynthesis in soybean, *Plant Physiology*, 197(2): kiae485.
<https://doi.org/10.1093/plphys/kiae485>
- Yang Y., Zhang L., Zuo H., Yang Y., Hu D., Zhang S., Yuan W., Zhai X., He M., Xu M., Wang J., Lu W., Hu D., Yu D., Huang F., and Zhang D., 2025, *GmGASA12* coordinates hormonal dynamics to enhance soybean water-soluble protein accumulation and seed size, *Journal of Integrative Plant Biology*, 67(9): 2401-2415.
<https://doi.org/10.1111/jipb.13952>
- Yuan X., Jiang X., Zhang M., Wang L., Jiao W., Chen H., Mao J., Ye W., and Song Q., 2024, Integrative omics analysis elucidates the genetic basis underlying seed weight and oil content in soybean, *The Plant Cell*, 36(6): 2160-2175.
<https://doi.org/10.1093/plcell/koae062>
- Zhang M., Liu S., Wang Z., Yuan Y., Zhang Z., Liang Q., Yang X., Duan Z., Liu Y., Kong F., Liu B., Ren B., and Tian Z., 2022, Progress in soybean functional genomics over the past decade, *Plant Biotechnology Journal*, 20(2): 256-282.
<https://doi.org/10.1111/pbi.13682>
- Zhang S., Du H., Li H., Kan G., and Yu D., 2021, Linkage and association study discovered loci and candidate genes for glycinin and β -conglycinin in soybean (*Glycine max* L. Merr.), *Theoretical and Applied Genetics*, 134(4): 1201-1215.
<https://doi.org/10.1007/s00122-021-03766-6>



Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.