

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