# Identification of Sugar-Related Genes in Sugar Beet (Beta vulgaris) Through Comparative Genomics and Machine Learning Approaches

**Sara Behnamian**

sara.behnamian@sund.ku.dk

University of Copenhagen

**Method Article**

**Keywords:** Sugar metabolism, Beta vulgaris, transformer models, comparative genomics, Arabidopsis thaliana, functional annotation, Semantic analysis, Zero-shot classification, Orthology analysis

**Posted Date:** March 21st, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-7950380/v1



**Additional Declarations:** No competing interests reported.

# Abstract

Despite sugar beet's importance as a major sucrose source, comprehensive identification of genes underlying sugar metabolism remains incomplete. We developed an integrated approach combining comparative genomics with advanced machine learning to systematically catalog sugar-related genes in the Beta vulgaris genome. Analysis of the EL10 reference genome identified 18,223 high-quality protein-coding genes, of which 91.6% showed orthology to Arabidopsis thaliana proteins. Traditional keyword-based screening identified 310 sugar-related genes, of which 286 exhibited high-confidence orthology (E-value < $1\times10^{-50}$) to Arabidopsis proteins. To overcome limitations of keyword approaches, we implemented zero-shot classification using transformer-based Sentence-BERT embeddings to identify genes through semantic similarity to sugar-related concepts, independent of explicit nomenclature. This machine learning strategy identified 1,999 candidate genes, including 1,736 novel candidates absent from keyword results—an 85% expansion of the sugar gene catalog. Despite this novelty, 84.8% of keyword-identified genes were also detected by machine learning, validating the approach. Multiple high-confidence predictions corresponded to experimentally validated genes in published studies. This framework establishes transformer-based semantic analysis as a powerful complement to traditional annotation, with broad applicability for functional gene discovery in crop genomics.


# Introduction

Sugar beet (*Beta vulgaris* L.) represents one of the two primary sources of global sucrose production, contributing approximately 20% of the world's sugar supply and serving as a critical crop for temperate agricultural systems (McGrath et al., 2023). Unlike sugarcane, which stores sucrose in stem tissues, sugar beet accumulates exceptionally high concentrations of sucrose in its enlarged taproot, reaching up to 20% fresh weight under optimal conditions. This remarkable capacity for carbohydrate accumulation has made sugar beet an invaluable model system for studying sugar metabolism, transport, and storage mechanisms in dicotyledonous plants. Understanding the genetic basis of sugar-related processes in sugar beet is essential for crop improvement efforts. These efforts aim to enhance sucrose yield, optimize resource allocation, and develop climate-resilient cultivars capable of maintaining productivity under environmental stress.

Despite the agricultural and economic importance of sugar beet, comprehensive identification and characterization of genes involved in sugar metabolism and transport remain incomplete. The recent release of high-quality reference genome assemblies, including the *Beta vulgaris* cultivar EL10 genome (McGrath et al., 2023), has created unprecedented opportunities for systematic genome-wide analysis of sugar-related gene families. However, functional annotation of newly sequenced genomes typically relies on sequence similarity to well-characterized model organisms. While this enables annotation transfer, it may miss species-specific genes or misannotate divergent orthologs. *Arabidopsis thaliana*, as the most extensively studied dicot model organism with comprehensive functional annotations and curated metabolic pathway databases (The Arabidopsis Genome Initiative, 2000), provides an ideal reference system for comparative genomics approaches in sugar beet. Arabidopsis and sugar beet share a close

phylogenetic relationship within the eudicot clade. This relationship, combined with deep conservation of core metabolic pathways, enables reliable ortholog identification and functional inference through sequence similarity searches.

Traditional approaches to identifying genes involved in specific biological processes rely predominantly on keyword-based searching of functional annotations, which effectively captures genes with explicit nomenclature but may overlook functionally related genes lacking conventional terminology. Recent advances in natural language processing and machine learning offer complementary strategies for gene discovery through semantic analysis of textual annotations. Transformer-based language models, particularly those built upon BERT architecture (Devlin et al., 2019), have demonstrated remarkable capacity to capture contextual relationships and semantic similarity in biological text. The development of sentence embedding models such as Sentence-BERT (Reimers and Gurevych, 2019) enables efficient computation of semantic similarity between gene descriptions and concept definitions, facilitating zero-shot classification approaches that can identify functionally related genes based on contextual patterns rather than exact keyword matches. This machine learning paradigm has shown promise in various biological applications but remains underexplored for plant genome annotation and functional gene discovery.

In this study, we integrate comparative genomics and machine learning approaches to comprehensively identify and characterize sugar-related genes in the sugar beet genome. We employ BLASTP-based orthology analysis (Altschul et al., 1997) against *Arabidopsis thaliana* proteins to establish evolutionary relationships and transfer functional annotations, followed by systematic keyword-based identification of genes involved in sugar metabolism, transport, and related processes. Complementary to this traditional approach, we implement a zero-shot machine learning method using transformer-based semantic embeddings to identify sugar-related genes based on contextual similarity to predefined sugar concepts, enabling discovery of candidates that may lack explicit sugar-related terminology. Through this dual-strategy approach (Fig. 1), we aim to provide a comprehensive catalog of sugar-related genes in sugar beet, evaluate the complementary strengths of keyword-based and machine learning methods for functional gene identification, and establish a framework for semantic analysis-based gene discovery in crop genomics.

Schematic overview of the dual-strategy approach combining comparative genomics and machine learning. The pipeline begins with quality assessment of 30,391 annotated genes from the sugar beet EL10 genome, yielding 18,223 high-quality protein-coding sequences. BLASTP orthology analysis against Arabidopsis thaliana establishes evolutionary relationships through E-values and alignment scores. Two complementary identification strategies operate in parallel: keyword-based analysis using a comprehensive lexicon (Table 1) identifies 310 genes with explicit sugar-related terminology, while machine learning employs transformer-based semantic embeddings and 13 predefined concept prompts (Table 2) to identify 1,999 genes (lenient threshold) through contextual patterns. Integration reveals 263 genes (84.8%) identified by both methods, validating the approaches, while machine learning uniquely identifies 1,736 additional candidates. Unsupervised clustering identifies three sugar-enriched functional

groups containing 1,589 genes. The comprehensive catalog combines both strategies and is validated against experimentally characterized genes from published studies.

# Methods

The genomic analysis was conducted using the Beta vulgaris subsp. vulgaris cultivar EL10 reference genome assembly (EL10.2) (McGrath et al., 2023). This assembly was obtained from the National Center for Biotechnology Information (NCBI) RefSeq database (accession number GCF_026745355.1). This high-quality assembly, generated by the USDA Agricultural Research Service, comprises a total genome size of 568.8 Mb with nine chromosomes and two organellar genomes, assembled using PacBio RSII long-read sequencing technology and the FALCON assembler (v. 0.2.2). The assembly demonstrated exceptional contiguity with a scaffold N50 of 62 Mb and contig N50 of 1.3 Mb, representing 50× genome coverage. Quality assessment using BUSCO analysis (Manni et al., 2021) against the eudicots_odb10 database revealed 96.7% completeness (94.4% single-copy, 2.3% duplicated), with only 0.3% fragmented and 3.0% missing genes, confirming the assembly's high quality and suitability for comprehensive genomic analysis.

Genome annotation data were obtained from the NCBI RefSeq annotation pipeline (v. 10.1) dated June 7, 2023, which identified 30,391 total genes including 24,186 protein-coding sequences. The genomic feature file (GFF3 format) containing structural and functional annotations was processed computationally to create a local database for efficient data retrieval and analysis. The GFF3 file was converted to a SQLite database format using the gffutils Python library (Dale, 2024), employing a merge strategy to consolidate overlapping features while maintaining chronological order and attribute value sorting. This database structure facilitated systematic extraction and analysis of genomic features, including gene coordinates, functional annotations, and regulatory elements across the sugar beet genome.

Gene structure and coding sequence (CDS) analysis was performed by extracting genomic coordinates and sequence information for all annotated protein-coding genes from the SQLite database. Each gene's associated CDS features were identified through GeneID cross-references parsed from the JSON-formatted attribute fields in the annotation data. The corresponding DNA sequences were retrieved from the reference genome FASTA file (GCF_026745355.1_EL10.2_genomic.fna) based on chromosomal coordinates, with sequences on the negative strand subjected to reverse complementation using BioPython sequence manipulation functions (Cock et al., 2009). For genes containing multiple CDS segments, individual coding sequences were concatenated in genomic order to reconstruct the complete coding sequence for each gene.

Comprehensive quality assessment of coding sequences was conducted to evaluate annotation accuracy and gene structure integrity. Each reconstructed coding sequence was analyzed for reading frame conservation by verifying that the total length was divisible by three, and codon usage was examined through identification and validation of start codons (ATG) and stop codons (TAA, TAG, TGA).

Additional annotation quality metrics were extracted including partial gene predictions, hypothetical protein classifications, and pseudogene annotations. All analytical results, including gene coordinates, CDS counts per gene, sequence lengths, codon validation status, and annotation flags, were compiled into a comprehensive dataset for subsequent statistical analysis.

Internal stop codon analysis was performed to identify premature termination signals within coding sequences that could indicate annotation errors or pseudogenization events. Each reconstructed coding sequence was systematically examined for the presence of in-frame stop codons (TAA, TAG, TGA) occurring before the terminal stop codon, excluding the final three nucleotides which represent the legitimate termination signal. This analysis was conducted only on sequences longer than three nucleotides to ensure meaningful assessment of internal coding regions.

Gene classification was performed based on multiple quality criteria to distinguish between high-confidence protein-coding genes and potentially problematic annotations. Genes were categorized as clean coding sequences if they satisfied all of the following stringent criteria: presence of valid start codon (ATG), presence of valid stop codon (TAA, TAG, or TGA), absence of internal stop codons, complete gene annotation status (non-partial), and absence of partial CDS segments. Genes failing to meet any of these criteria were classified as problematic coding sequences, potentially representing pseudogenes, annotation artifacts, or incomplete gene models requiring further validation.

Functional gene annotations were retrieved from the NCBI Gene database using the Entrez Programming Utilities (E-utilities) to obtain detailed gene names and descriptions for all identified clean coding sequences. Gene identifiers were systematically queried against the NCBI Gene database using the Biopython Entrez module, with a 0.5-second delay implemented between requests to comply with NCBI rate limiting guidelines. For each gene, summary information including gene names and functional descriptions were extracted from the XML-formatted responses and integrated with the existing dataset. This annotation enrichment process provided comprehensive functional context for the high-quality protein-coding genes identified through the quality assessment pipeline.

Data preparation for downstream analysis involved partitioning the annotated clean coding sequences dataset into smaller, manageable subsets to facilitate computational processing. The complete dataset of high-quality protein-coding genes was systematically divided into chunks of 50 genes each, with each subset exported as a separate file to enable parallel processing and progress tracking. This chunking strategy was implemented to optimize computational efficiency and provide checkpoint capabilities for large-scale sequence analysis workflows, ensuring robust data management throughout subsequent comparative genomics analyses.

Comparative genomics analysis was conducted to identify functional orthologs between sugar beet and *Arabidopsis thaliana* using protein sequence similarity searches. For each high-quality protein-coding gene, the corresponding protein sequence was retrieved from the NCBI protein database using gene-specific searches constrained to *Beta vulgaris* entries. Protein sequences were then subjected to BLASTP analysis against the NCBI non-redundant protein database with searches restricted to

*Arabidopsis thaliana* to identify putative orthologous relationships. BLAST searches were performed with default parameters, retaining the top hit for each query sequence. Alignment results included protein names, functional annotations, E-values, alignment scores, percent identity, and query coverage as reported by BLAST. Query coverage and percent identity were not used as explicit filtering –criteria; instead, homology confidence was assessed using E-value–based stratification.For each sugar beet protein, only the top Arabidopsis BLASTP hit was retained, ranked by lowest E-value and highest alignment score. We did not explicitly distinguish between one-to-one, one-to-many, or many-to-one ortholog relationships, as the goal of this analysis was functional annotation rather than evolutionary reconstruction. This approach allows paralogous sugar beet genes arising from duplication events to be associated with conserved Arabidopsis functional annotations. For sequences that initially returned no significant matches or encountered search errors, BLAST analyses were repeated up to three times with extended delays to account for potential network connectivity issues or temporary database unavailability.

BLAST results from all gene chunks were consolidated into a unified dataset. Genes with failed or missing BLAST searches were identified and re-analyzed, with newly obtained results integrated into the final dataset by replacing empty entries based on gene identifier matching where successful. Statistical analysis categorized all genes by orthology confidence: high-confidence matches (E-value < $1\times10^{-50}$), moderate-confidence matches ($1\times10^{-50}$ ≤ E-value < $1\times10^{-5}$), low-confidence matches (E-value ≥ $1\times10^{-5}$), genes with no significant Arabidopsis matches, and genes lacking protein sequences.

# Sugar-related gene identification and functional categorization

Sugar-related genes were identified through keyword-based analysis of Arabidopsis ortholog annotations obtained from BLAST results. A comprehensive lexicon of sugar-related terms was compiled (Table 1), including carbohydrate types (sucrose, fructose, glucose, sugar, sweet, hexose, pentose, trehalose, maltose, galactose), sugar classes (monosaccharide, disaccharide, polysaccharide), related compounds (starch, glycogen, saccharide, glucan), transport terms (transporter, transport, translocation, carrier, permease, channel, pump, symporter, antiporter), metabolism terms (synthase, synthesis, biosynthesis, metabolism, metabolic, phosphatase, kinase, transferase), and root development terms (root, taproot, lateral root, root hair, root tip, root development, root growth, root elongation). Keywords were matched using case-insensitive substring matching applied to Arabidopsis protein function annotations. Genes were classified as sugar-related if their corresponding Arabidopsis protein function contained any sugar-related keyword. Additional functional categories were assigned based on co-occurrence with transport-related terms (transporter, carrier, channel, pump, permease, symporter, antiporter) or metabolism-related terms (synthase, biosynthesis, metabolism, phosphatase, kinase, transferase), enabling classification into sugar transport, sugar metabolism, or general sugar-related categories. Root development genes were similarly identified using root-specific keywords (root, taproot, lateral root, root hair). Only genes with valid BLAST results and E-values were included in the

analysis. Sugar-related genes were ranked by a dual-criterion system: primarily by E-value (ascending order, prioritizing evolutionary conservation) and secondarily by alignment score (descending order, prioritizing alignment quality) to identify the most reliable sugar-related orthologs. Genes with E-values less than $1\times10^{-50}$ were classified as high-confidence conserved orthologs, with further stratification into perfect matches ($E = 0$), exceptional matches ($0 < E < 1\times10^{-200}$), excellent matches ($1\times10^{-200} \leq E < 1\times10^{-100}$), and very strong matches ($1\times10^{-100} \leq E < 1\times10^{-50}$).

Table 1
Complete keyword lexicon for sugar-related gene identification

| Category | Keywords |
|---|---|
| Carbohydrate types | sucrose; fructose; glucose; sugar; sweet; hexose; pentose; trehalose; maltose; galactose |
| Sugar classes | monosaccharide; disaccharide; polysaccharide |
| Related compounds | starch; glycogen; saccharide; glucan |
| Transport | transporter; transport; translocation; carrier; permease; channel; pump; symporter; antiporter |
| Metabolism | synthase; synthesis; biosynthesis; metabolism; metabolic; phosphatase; kinase; transferase |
| Root development | root; taproot; lateral root; root hair; root tip; root development; root growth; root elongation |

Keywords were searched using case-insensitive substring matching (e.g., "glucan" matches "beta-glucan", "glucanase", "glucan synthase").

**Sugar-related annotation retention in Amaranthaceae homologs was assessed using the same comprehensive keyword lexicon defined in** Table 1, **ensuring methodological consistency across all comparative analyses.**

# Machine learning-based sugar gene identification

Complementary to the keyword-based approach, a zero-shot machine learning method was employed to identify sugar-related genes through semantic similarity analysis of functional annotations. All 18,223 genes with valid Arabidopsis ortholog descriptions were processed using transformer-based language models to capture contextual relationships beyond exact keyword matching. Gene descriptions were prepared by concatenating Arabidopsis protein functions (weighted twice to emphasize biological terminology), Arabidopsis protein names, sugar beet protein names, and detailed sugar beet gene names into unified text representations. Text embeddings were generated using the sentence-transformers/all-mpnet-base-v2 model (Reimers and Gurevych, 2019), a transformer-based encoder building upon BERT architecture (Devlin et al., 2019) and pre-trained on over 1 billion sentence pairs. The all-mpnet-base-v2

model was selected due to its strong performance on semantic similarity benchmarks and its ability to capture fine-grained contextual relationships in complex biological text. Semantic embedding quality was prioritized over computational efficiency, as accurate zero-shot functional inference was central to the objectives of this study. Thirteen sugar-related concept embeddings were constructed (Table 2) encompassing metabolic processes, transport functions, enzymatic activities, storage mechanisms, signaling pathways, and sugar beet-specific processes. Cosine similarity was calculated between each gene's embedding and all concept embeddings, with the maximum similarity score retained as the gene's sugar-relatedness metric. Adaptive thresholding based on statistical distribution properties was applied, with genes classified using moderate (mean + 1.5 standard deviations), lenient (mean + 1 standard deviation), and stringent (95th percentile) threshold strategies. Confidence levels were assigned as very high (> mean + 2σ), high (mean + 1.5σ to mean + 2σ), moderate (mean + σ to mean + 1.5σ), low (mean + 0.5σ to mean + σ), or very low (< mean + 0.5σ). Unsupervised clustering was performed on reduced-dimensional embeddings (50 principal components) using principal component analysis (Jolliffe and Cadima, 2016) and k-means clustering (MacQueen, 1967) with optimal cluster number selection via silhouette score analysis (Rousseeuw, 1987), identifying functionally related gene groups and sugar-enriched clusters containing > 10% sugar-related genes.

Table 2
Sugar-related concept prompts for semantic similarity analysis

| ID | Category | Exact prompt text |
|---|---|---|
| C1 | Metabolism | sucrose metabolism; biosynthesis; catabolism; sugar phosphate synthase; transferase |
| C2 | Metabolism | glucose; fructose; hexose phosphate; metabolism; glycolysis; gluconeogenesis |
| C3 | Metabolism | carbohydrate metabolism; polysaccharide biosynthesis; starch; cellulose; glucan |
| C4 | Transport | sugar transporter; sucrose transporter; glucose transporter; hexose carrier; symporter; antiporter |
| C5 | Transport | carbohydrate transport; translocation; phloem loading; unloading; sink; source |
| C6 | Enzymes | sucrose phosphate synthase; sucrose synthase; invertase; fructokinase; hexokinase; glucokinase |
| C7 | Enzymes | alpha-glucosidase; beta-glucosidase; amylase; glucanase; glycosyltransferase; glycosidase |
| C8 | Storage | sugar storage; accumulation; vacuolar storage; root storage; starch granule |
| C9 | Storage | osmotic regulation; turgor pressure; sugar alcohol; sorbitol; mannitol; trehalose |
| C10 | Signaling | sugar signaling; trehalose-6-phosphate; sucrose non-fermenting kinase; hexokinase sensor |
| C11 | Signaling | sugar-responsive gene expression; transcription factor; sugar sensing |
| C12 | Beet-specific | Beta vulgaris; sucrose accumulation; taproot storage; sink strength; sugar yield |
| C13 | Beet-specific | sugar beet; root development; storage root; sucrose content; brix |

Text embeddings were generated using the sentence-transformers/all-mpnet-base-v2 model with default parameters (768-dimensional embeddings, batch size 32). No text preprocessing was applied; gene descriptions were analyzed in original case without lowercasing, lemmatization, or stopword removal to preserve biological terminology. All stochastic procedures used random_state = 42 for reproducibility (PCA, k-means clustering). The weighting scheme for gene description construction assigned double emphasis to Arabidopsis protein functions by including them twice in concatenated text.

## Mathematical formulation of semantic similarity scoring

Let Ti denote the textual annotation associated with gene i, constructed by concatenating Arabidopsis protein functions (weighted twice), Arabidopsis protein names, sugar beet protein names, and sugar

beet gene names. A transformer-based sentence encoder f(·) maps each text into a fixed-dimensional embedding vector:

$$e_i = f(T_i) \in \mathbb{R}^d$$

where d = 768 for the all-mpnet-base-v2 model. Similarly, each predefined sugar-related concept prompt Cj (Table 2) is embedded as:

$$c_j = f(C_j) \in \mathbb{R}^d$$

Semantic similarity between a gene and a sugar-related concept is quantified using cosine similarity:

$$\mathrm{sim}(g_i, C_j) = \frac{e_i \cdot c_j}{\| e_i \| \| c_j \|}$$

For each gene, the maximum similarity across all K = 13 sugar-related concept prompts is retained as the gene's sugar-relatedness score:

$$S_i = \max_{j \in \{1,\ldots,K\}} \mathrm{sim}(g_i, C_j)$$

Genes are classified as sugar-related if their score Si exceeds an adaptive threshold τ, defined based on the empirical distribution of similarity scores across all N = 18,223 genes. Thresholds corresponding to lenient (μ + σ), moderate (μ + 1.5σ), and stringent (95th percentile) criteria were evaluated, where μ and σ denote the mean and standard deviation of {Si}:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} S_i$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (S_i - \mu)^2}$$

A gene is classified as sugar-related if Si ≥ τ.

Gene Ontology (GO) and KEGG pathway enrichment analysis was performed on ML-unique genes (identified by machine learning but absent from keyword results) to validate functional coherence of predictions. Arabidopsis thaliana gene identifiers were extracted from ortholog annotations using pattern matching for AT identifiers (e.g., AT1G12345) and gene symbols. Enrichment analysis was conducted using g:Profiler (Raudvere et al., 2019) with the following parameters: organism set to Arabidopsis thaliana, significance threshold of FDR < 0.05 using the Benjamini-Hochberg method, and functional categories including GO Biological Process, GO Molecular Function, GO Cellular Component, and KEGG pathways. All analyzed genes with extractable identifiers served as the statistical background to control for annotation biases.

# Results

Initial genome-wide analysis of the Beta vulgaris EL10 reference genome identified 30,391 total annotated genes from NCBI RefSeq, of which 29,386 protein-coding genes were subjected to quality assessment after excluding non-coding RNAs and pseudogenes. Comprehensive coding sequence validation revealed that 18,223 genes (62.0%) met stringent quality criteria and were classified as high-quality protein-coding sequences suitable for comparative genomics analysis. The remaining 11,163 genes (38.0%) exhibited various annotation quality issues that prevented reliable downstream analysis, with the primary concerns being absence of coding sequences (5,170 genes), presence of internal stop codons suggesting pseudogenization (5,862 genes), incomplete gene models (115 partial genes, 134 partial CDS), invalid start codons (100 genes), and invalid stop codons (43 genes). Only the validated high-quality coding sequences were subjected to comparative genomics analysis.

The comparative genomics analysis successfully identified putative orthologous relationships for the majority of high-quality sugar beet protein-coding genes through BLASTP searches against Arabidopsis thaliana proteins. Of the 18,223 genes analyzed, 16,698 genes (91.6%) showed detectable sequence similarity to known Arabidopsis proteins. Within the valid BLAST hits, 11,059 genes (60.7%) demonstrated high confidence orthologous relationships with E-values less than $1\times10^{-50}$, while an additional 5,072 genes (27.8%) showed moderate confidence matches with E-values between $1\times10^{-50}$ and $1\times10^{-5}$. An additional 567 genes (3.1%) displayed low confidence matches with E-values greater than or equal to $1\times10^{-5}$, suggesting more distant evolutionary relationships. For genes that initially failed to return BLAST results or encountered search errors, repeated BLAST analyses were performed on 4,318 genes (23.7% of the dataset), with successful results integrated into the final dataset where obtained. A total of 1,088 genes (6.0%) returned no significant matches to Arabidopsis proteins, and 437 genes (2.4%) could not be analyzed due to absent protein sequences, potentially representing species-specific adaptations, divergent evolutionary lineages, or annotation gaps unique to sugar beet.

Detailed characterization of the orthology dataset revealed a gradient of sequence conservation across the 18,223 validated genes. E-value distributions showed that 3,872 genes (21.2%) exhibited perfect sequence identity ($E = 0$), while 3,484 genes (19.1%) demonstrated excellent conservation ($E < 1\times10^{-100}$). Strong to very strong orthologous relationships were observed in 8,345 genes (45.8%), comprising 3,703 very strong matches ($1\times10^{-100} \leq E < 1\times10^{-50}$) and 4,642 strong matches ($1\times10^{-50} \leq E < 1\times10^{-10}$). Lower confidence assignments included 430 moderate (2.4%), 150 weak (0.8%), and 417 very weak matches (2.3%). BLAST alignment scores ranged from 58 to 16,468 (median: 671, mean: 933.2), with E-values spanning $1.04\times10^{-180}$ to 9.95 (median: $1.42\times10^{-58}$). A subset of 1,525 genes (8.4%) lacked E-values due to missing protein sequences or BLAST search failures. The orthology assignments mapped sugar beet genes to 6,193 unique Arabidopsis proteins across 39 functional categories. Predominant enzyme classes included kinases (116), dehydrogenases (69), and transcriptases (45), while functional classification identified oxidoreductases (137), transcription factors (136), transferases (104), and transporters (88) as the most abundant groups. Significant representation was observed for organellar

proteins (chloroplast: 78, mitochondrial: 54), stress response systems (heat shock: 62, oxidative stress: 32), and regulatory elements (receptors: 33, transcriptional regulators: 31).

Keyword-based functional annotation identified 310 sugar-related genes (1.7% of the validated gene set) based on Arabidopsis ortholog annotations, with 286 genes (92.3%) classified as high-confidence orthologs exhibiting E-values less than $1\times10^{-50}$. Among these conserved sugar-related genes, 169 (59.1%) demonstrated perfect sequence matches (E = 0) to Arabidopsis orthologs, while 87 genes (30.4%) showed excellent conservation ($1\times10^{-200} \leq E < 1\times10^{-100}$), and 30 genes (10.5%) exhibited very strong orthology ($1\times10^{-100} \leq E < 1\times10^{-50}$), with E-values for non-perfect matches ranging from $2.28\times10^{-179}$ to $2.99\times10^{-51}$ (median: $2.59\times10^{-133}$). Functional categorization revealed 91 sugar metabolism genes (31.8%), including 54 perfect matches and 32 excellent matches, with top-ranked genes encoding callose synthases, cellulose synthases, sucrose-phosphate synthases (SBSPS1), and sucrose synthases (SBSS1). The sugar transport category comprised 58 genes (20.3%), including 36 perfect matches and 10 excellent matches, with the highest-ranked transporters being monosaccharide-sensing proteins, sugar transport proteins, and sugar carrier proteins. An additional 137 genes (47.9%) were classified as general sugar-related functions, including 79 perfect matches and 45 excellent matches, with notable genes encoding protein SWEETIE, glucan phosphorylases, and xylosidases. The most frequently occurring sugar-related terms in ortholog annotations were glucan (243 occurrences), sugar (188 occurrences), glucose (134 occurrences), and galactose (62 occurrences), followed by sucrose (49 occurrences) and fructose (39 occurrences). The highest-ranked sugar-related genes exhibited perfect sequence matches with exceptionally high alignment scores ranging from 3,034 to 8,644, underscoring the fundamental conservation of carbohydrate biosynthesis machinery between sugar beet and Arabidopsis.

To further validate sugar-related gene assignments using phylogenetically closer species, cross-species BLASTP analysis was performed against two Amaranthaceae members, *Chenopodium quinoa* and *Spinacia oleracea*. All 100 high-confidence sugar-related genes (Table S6) showed identifiable homologs in both species, with E-values = 0 and mean sequence identity exceeding 87%. Using the same sugar-related keyword lexicon defined in Table 1, 78% of quinoa homologs and 74% of spinach homologs retained sugar-related functional annotations (Tables S7–S8).

Machine learning-based semantic analysis identified sugar-related genes through contextual similarity to predefined sugar concepts, complementing the keyword-based approach (Fig. 2). Using adaptive thresholding based on similarity score distribution (mean = 0.399, SD = 0.083, range: 0.108–0.738), three detection strategies were employed: moderate threshold (mean + 1.5σ = 0.523) identifying 1,002 genes (Table S1), lenient threshold (mean + 1σ = 0.482) identifying 1,999 genes (Table S2), and stringent 95th percentile identifying 912 genes (Table S3) (Fig. 3). To statistically validate threshold selection, precision-recall (PR) and receiver operating characteristic (ROC) analyses were performed using the 310 keyword-identified genes as ground truth positives (Fig. 4, Table 3). The ROC analysis demonstrated excellent discriminative performance (AUC = 0.971), confirming that the machine learning approach effectively ranks sugar-related genes higher than non-sugar genes. The PR analysis yielded an AUC of

0.515, which reflects the inherent class imbalance (1.7% positive rate; random baseline ≈ 0.017) rather than poor model performance, representing a 30-fold improvement over random classification. At the lenient threshold (0.482), the model achieved 92.0% recall (263 of 286 keyword genes detected) with 13.4% precision. The moderate threshold (0.525) achieved 83.2% recall with 24.5% precision. The F1-optimal threshold (0.599) yielded the best balance between precision (57.7%) and recall (48.3%), with an F1 score of 0.526. These metrics demonstrate that the selected thresholds appropriately balance discovery potential against prediction confidence, with lower thresholds prioritizing recall for exploratory gene discovery. The top-ranked genes included sucrose synthase LOC104900089 (similarity: 0.738), sucrose-phosphate synthases SBSPS1 and LOC104909043 (0.729), glucan endo-1,3-beta-glucosidases (0.716–0.728), sugar transport protein 10 LOC104902554 (0.716), and beta-amylase LOC104903010 (0.714) (Table S4). Confidence level assignment across all 18,223 genes revealed 492 very high (2.7%), 510 high (2.8%), 997 moderate (5.5%), 2,181 low (12.0%), and 14,043 very low confidence genes (77.1%) (Fig. 5). Comparison with keyword-based results showed 84.8% overlap (263 of 310 genes), while machine learning uniquely identified 1,736 candidates, including 887 genes entirely absent from keyword detection (Fig. 8, Table S2). To validate that ML-unique genes represent functionally coherent sugar-related candidates, GO and KEGG enrichment analysis was performed using 615 Arabidopsis identifiers extracted from the 1,736 ML-unique genes, with 4,275 identifiers from all analyzed genes as background. The analysis revealed 116 significantly enriched terms (FDR < 0.05), strongly supporting the functional relevance of ML predictions (Table 4, Supplementary Table S5). For GO Biological Process, the most enriched terms included transmembrane transport ($P = 1.38\times10^{-15}$, 79 genes) and carbohydrate metabolic process ($P = 1.09\times10^{-6}$, 56 genes). GO Molecular Function analysis showed highly significant enrichment for transmembrane transporter activity ($P = 1.66\times10^{-15}$, 74 genes), glycosyltransferase activity ($P = 7.55\times10^{-14}$, 50 genes), UDP-glycosyltransferase activity ($P = 5.79\times10^{-10}$, 28 genes), and hexosyltransferase activity ($P = 4.03\times10^{-9}$, 33 genes). GO Cellular Component analysis revealed enrichment for Golgi apparatus ($P = 2.78\times10^{-5}$, 59 genes) and vacuole ($P = 3.77\times10^{-4}$, 57 genes), consistent with subcellular sites of glycosylation and sugar storage. KEGG pathway analysis identified enrichment for metabolic pathways ($P = 9.13\times10^{-9}$, 111 genes) and carbon metabolism ($P = 6.13\times10^{-3}$, 18 genes). These results demonstrate that ML-unique genes are significantly enriched for carbohydrate metabolism, sugar transport, and glycosyltransferase functions, validating that the machine learning approach identifies functionally coherent candidates despite their absence from keyword-based detection. Functional analysis of the lenient set revealed 370 transport-related genes, 99 glucan-processing genes, 88 synthase genes, and 49 glucose-related genes. Unsupervised clustering via k-means identified 19 functionally distinct clusters (silhouette score: 0.210), with clusters 6, 8, and 16 designated as sugar-enriched, collectively containing 1,589 genes (79.5% of lenient predictions), predominantly cluster 16 (898 genes), cluster 8 (466 genes), and cluster 6 (225 genes) (Fig. 6). Similarity score distributions demonstrated distinct right-shifted patterns for sugar genes (scores 0.48–0.74, average 0.54) compared to genome-wide background (Fig. 7), validating the semantic approach's capacity to capture functional relationships independent of explicit keyword presence. Given the overlapping nature of biological functions and semantic embedding spaces, k-means clustering is used

here as an exploratory tool to identify functionally enriched gene groups rather than as a strict classification of sugar-related genes.

Table 3
Performance metrics for machine learning classification thresholds

| Threshold | Performance Metrics |
|---|---|
| Lenient (mean+1σ) | Value: 0.482; Precision: 13.4%; Recall: 92.0%; F1: 0.233; TP: 263; FP: 1,706; Genes predicted: 1,969 |
| Moderate (mean + 1.5σ) | Value: 0.525; Precision: 24.5%; Recall: 83.2%; F1: 0.379; TP: 238; FP: 733; Genes predicted: 971 |
| Stringent (95th pct) | Value: 0.529; Precision: 25.4%; Recall: 81.1%; F1: 0.387; TP: 232; FP: 680; Genes predicted: 912 |
| F1-optimal | Value: 0.599; Precision: 57.7%; Recall: 48.3%; F1: 0.526; TP: 138; FP: 101; Genes predicted: 239 |

*Performance metrics calculated using 286 keyword-identified sugar genes as ground truth positives. TP = true positives; FP = false positives; Precision = TP/(TP + FP); Recall = TP/(TP + FN); F1 = 2×Precision×Recall/(Precision+Recall).*

Table 4
GO and KEGG enrichment analysis of ML-unique genes

| Category | ID | Term Name | Adj. P-value | Gene Count |
|---|---|---|---|---|
| GO:BP | GO:0055085 | Transmembrane transport | $1.38\times10^{-15}$ | 79 |
| GO:BP | GO:0005975 | Carbohydrate metabolic process | $1.09\times10^{-6}$ | 56 |
| GO:BP | GO:0044281 | Small molecule metabolic process | $1.09\times10^{-6}$ | 81 |
| GO:MF | GO:0022857 | Transmembrane transporter activity | $1.66\times10^{-15}$ | 74 |
| GO:MF | GO:0016757 | Glycosyltransferase activity | $7.55\times10^{-14}$ | 50 |
| GO:MF | GO:0008194 | UDP-glycosyltransferase activity | $5.79\times10^{-10}$ | 28 |
| GO:MF | GO:0016758 | Hexosyltransferase activity | $4.03\times10^{-9}$ | 33 |
| GO:CC | GO:0005794 | Golgi apparatus | $2.78\times10^{-5}$ | 59 |
| GO:CC | GO:0005773 | Vacuole | $3.77\times10^{-4}$ | 57 |
| KEGG | KEGG:01100 | Metabolic pathways | $9.13\times10^{-9}$ | 111 |
| KEGG | KEGG:01200 | Carbon metabolism | $6.13\times10^{-3}$ | 18 |

*Top enriched terms from GO Biological Process (BP), Molecular Function (MF), Cellular Component (CC), and KEGG pathways for 1,736 ML-unique genes. Analysis performed using g:Profiler with 615 Arabidopsis identifiers as query and 4,275 identifiers as background. Full results in Supplementary Table S5.*

UpSet plot comparing keyword-based identification (310 genes) with machine learning lenient threshold (1,999 genes). The intersection reveals 263 genes (84.8% of keyword genes) identified by both methods, validating the machine learning approach. Machine learning uniquely identified 1,736 additional candidates, demonstrating substantial discovery potential beyond traditional keyword matching. Twenty-three genes were found exclusively by the keyword method, indicating complementary strengths of both approaches.

# Discussion

This study presents a comprehensive genome-wide identification of sugar-related genes in sugar beet through an integrated approach combining traditional comparative genomics with advanced machine learning methodologies. Our dual-strategy framework successfully identified 310 high-confidence sugar-related genes through keyword-based analysis while simultaneously discovering over 1,700 additional candidates through semantic similarity analysis, demonstrating the complementary strengths of rule-based and machine learning approaches for functional gene annotation in crop genomes.

The comparative genomics analysis revealed exceptional conservation of sugar-related genes between sugar beet and *Arabidopsis thaliana*, with the vast majority of identified sugar genes exhibiting high-confidence orthologous relationships and over half showing perfect sequence matches. This remarkable conservation underscores the fundamental nature of core carbohydrate metabolism pathways across eudicots and validates the use of *Arabidopsis* as a reference system for functional annotation in sugar beet despite their evolutionary divergence. The identification of highly conserved sucrose synthases (SBSS1), sucrose-phosphate synthases (SBSPS1), and sugar transporters confirms the retention of essential enzymatic machinery for sucrose biosynthesis and accumulation, consistent with strong purifying selection on sugar metabolism genes given their critical roles in plant primary metabolism and the agronomically valuable trait of taproot sucrose accumulation in sugar beet.

The machine learning approach based on transformer-derived semantic embeddings demonstrated remarkable capacity to identify sugar-related genes through contextual similarity patterns independent of explicit keyword presence in annotations. The substantial overlap between machine learning and keyword-based methods validates the semantic approach's ability to capture known sugar-related genes, while the identification of over 1,700 unique candidates reveals substantial discovery potential beyond traditional annotation methods. These ML-unique candidates likely represent genes involved in sugar-related processes but annotated with non-standard terminology, regulatory genes affecting sugar metabolism without explicit carbohydrate-related descriptions, or novel functional relationships not captured by conventional keyword lexicons. The discovery of hundreds of transport-related, glucan-processing, and synthase genes within the lenient ML set suggests extensive functional diversity in sugar-related processes beyond core biosynthetic pathways.

The unsupervised clustering analysis revealed functionally coherent gene groups, with three sugar-enriched clusters collectively containing nearly 80% of lenient-threshold predictions. The dominant

cluster with hundreds of genes suggests a major functional module related to carbohydrate metabolism and transport, while the presence of multiple sugar-enriched clusters indicates functional specialization within sugar-related processes, potentially reflecting distinct subcellular localizations, tissue-specific expression patterns, or involvement in different metabolic pathways. The adaptive thresholding strategy enables flexible prioritization based on research objectives. The stringent threshold provides high-confidence candidates for experimental validation with minimal false positives, suitable for resource-intensive functional studies. The moderate threshold balances discovery breadth with prediction confidence, appropriate for large-scale transcriptomic or proteomic analyses. The lenient threshold maximizes discovery potential, capturing genes with subtle sugar-related functions or indirect involvement in carbohydrate processes, ideal for systems biology approaches or network analysis. This multi-threshold framework allows researchers to select appropriate stringency based on downstream applications and available resources.

The complementary nature of keyword-based and machine learning approaches suggests optimal annotation strategies should integrate both methodologies. Keyword methods excel at identifying genes with explicit, well-established nomenclature and provide interpretable, easily validated results based on direct textual matches. Machine learning methods capture contextual relationships, identify genes with non-standard terminology, and reveal functional associations not evident from keywords alone. Future work could explore ensemble methods combining keyword confidence scores with ML similarity metrics to create unified classification systems leveraging strengths of both approaches.

Several limitations warrant consideration. First, reliance on *Arabidopsis* orthologs for functional inference may miss sugar beet-specific adaptations or fail to capture functional divergence following gene duplication events. Second, the machine learning approach depends on annotation quality and completeness; genes with minimal or uninformative descriptions cannot be effectively classified regardless of their true function. Third, the zero-shot classification framework, while powerful for discovery, requires experimental validation to confirm predicted sugar-related functions. Fourth, the study focuses on protein-coding genes and does not address regulatory RNAs or cis-regulatory elements that may play important roles in sugar metabolism regulation.

Future research directions include experimental validation of ML-identified candidates through transcriptomic analysis across developmental stages and tissues with varying sucrose content, functional characterization of high-confidence novel candidates through reverse genetics approaches, integration of co-expression network analysis to identify regulatory relationships among sugar-related genes, and extension of the semantic analysis framework to identify genes involved in other agronomically important traits. The framework established here is broadly applicable to other crops and biological processes, providing a template for semantic analysis-based gene discovery in plant genomics.

In conclusion, this integrated comparative genomics and machine learning approach provides the most comprehensive catalog of sugar-related genes in sugar beet to date, identifying both highly conserved

core metabolic genes and novel candidates with putative sugar-related functions. The demonstration that transformer-based semantic analysis effectively complements traditional keyword methods establishes a new paradigm for functional genome annotation in crops, with implications extending beyond sugar metabolism to any biological process amenable to textual description. These findings provide a foundation for systems-level understanding of sugar accumulation in sugar beet and offer candidate genes for crop improvement strategies aimed at enhancing sucrose yield and quality.

# Validation against experimentally characterized sugar-related gene families

Validation in this study was performed at the level of experimentally characterized gene families rather than individual genes. Because no comprehensive, curated benchmark set of sugar-related genes exists for *Beta vulgaris*, validation focused on assessing whether gene families with established experimental evidence in sugar beet are systematically recovered by the proposed keyword-based and machine learning approaches. Collectively, these experimentally studied families—including the SWEET sugar transporter family (16 members; La et al., 2022), the sucrose transporter (SUT) family (9 members; Sun et al., 2023), tonoplast sugar transporters involved in vacuolar sucrose accumulation (e.g., BvTST2.1; Jung et al., 2015), and core sucrose metabolism enzymes such as sucrose synthases (SUS), sucrose-phosphate synthases (SPS), and invertases (Hesse et al., 1995)—comprise several dozen experimentally validated sugar-related genes. All of these families were recovered by at least one of the two identification strategies, with the majority detected by both keyword-based and machine learning approaches.

Across these families, our combined approaches recovered all 16 reported BvSWEET members and all 9 BvSUT members, demonstrating comprehensive coverage of experimentally characterized sugar transport machinery in sugar beet.

The sucrose synthase gene SBSS1, which our analysis identified with perfect sequence conservation (E-value = 0) and achieved high machine learning similarity scores, has been previously characterized as a 2762 bp cDNA encoding an 822 amino acid polypeptide expressed predominantly in sugar beet taproot tissue (Hesse et al., 1995). Experimental studies have demonstrated that both sucrose synthase and sucrose-phosphate synthase activities increase with plant age in sugar beet tissues, with maximum sucrose synthase activity reaching 33.5 μmol/mg protein·h in beet tissue (Godt and Roitsch, 1997), confirming the functional significance of these key biosynthetic enzymes we identified. Our identification of multiple SWEET family sugar transporters aligns with recent genome-wide surveys that identified 16 BvSWEET family members involved in sucrose translocation and storage in sugar beet (La et al., 2022). Additionally, the tonoplast sugar transporter BvTST2.1, which mediates vacuolar sucrose uptake in sugar beet taproots, has been experimentally validated through proteomic and electrophysiological analyses (Jung et al., 2015), providing independent confirmation for sugar transport genes identified in our analysis. The sucrose transporter (SUT) gene family members identified in our study are consistent with recent characterizations of nine BvSUT genes in sugar beet, with expression patterns showing

significantly elevated levels during tuber enlargement stages (Sun et al., 2023). This concordance between our computational predictions and experimentally validated genes across multiple independent studies establishes the reliability of both keyword-based and machine learning approaches for functional gene annotation. Importantly, our machine learning approach uniquely identified over 1,700 additional candidates not captured by traditional keyword methods, representing promising targets for future experimental validation and functional characterization.

# Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The datasets generated and/or analyzed during the current study are available as follows: (1) Beta vulgaris EL10 reference genome assembly (GCF_026745355.1) and annotation data are publicly available from NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_026745355.1/). (2) Analysis code and computational pipeline are available at https://github.com/sarabehnamian/BetaVulgaris-SugarGenes-ML. (3) Complete gene lists are provided in Supplementary Tables S1-S4, with GO/KEGG enrichment results in Supplementary Table S5 and high-confidence candidate genes in Supplementary Table S6.

### Competing interests

The author declares that there are no competing interests.

### Funding

No funding was received for this research.

### Authors' contributions

S.B. conceived the study, performed all analyses, and wrote the manuscript. The author read and approved the final manuscript.

### Acknowledgements

Not applicable.

# References


1. McGrath JM, Funk A, Galewski P, Ou S, Townsend B, Davenport K, Daligault H, Johnson S, Lee J, Hastie A, et al. A contiguous de novo genome assembly of sugar beet EL10 (*Beta vulgaris* L). DNA Res. 2023;30:dsac033.
2. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408:796–815.
3. Devlin J, Chang M-W, Lee K, Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.
4. Reimers N, Gurevych I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992.
5. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
6. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54.
7. Dale RK. (2024). *gffutils* (Version 0.13). GFF/GTF manipulation in Python. Available at: https://github.com/daler/gffutils
8. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
9. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical Trans Royal Soc A. 2016;374:20150202.
10. MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 281–297.
11. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
12. Godt DE, Roitsch T. Regulation and tissue-specific distribution of mRNAs for three extracellular invertase isoenzymes of tomato suggests an important function in establishing and maintaining sink metabolism. Plant Physiol. 1997;115:273–82.
13. Hesse H, Sonnewald U, Willmitzer L. Cloning and expression analysis of sucrose-phosphate synthase from sugar beet (Beta vulgaris L). Mol Gen Genet. 1995;247:515–20.

14. Jung B, Ludewig F, Schulz A, Meißner G, Wöstefeld N, Flügge UI, Pommerrenig B, Wirsching P, Sauer N, Koch W, et al. Identification of the transporter responsible for sucrose accumulation in sugar beet taproots. Nat Plants. 2015;1:14001.
15. La HV, Chu HD, Ha QT, Tran TTH, Tong HV, Tran TV, Le QTN, Bui HTT, Cao PB. SWEET Gene Family in Sugar Beet (Beta vulgaris): Genome-Wide Survey, Phylogeny and Expression Analysis. Pak J Biol Sci. 2022;25:387–95.
16. Sun F, Dong X, Li S, Sha H, Weishi G, Bai X, Li-ming Z, Yang H. Genome-wide identification and expression analysis of SUT gene family members in sugar beet (Beta vulgaris L). Gene. 2023;857:147171.
17. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47:W191–8.

# Figures

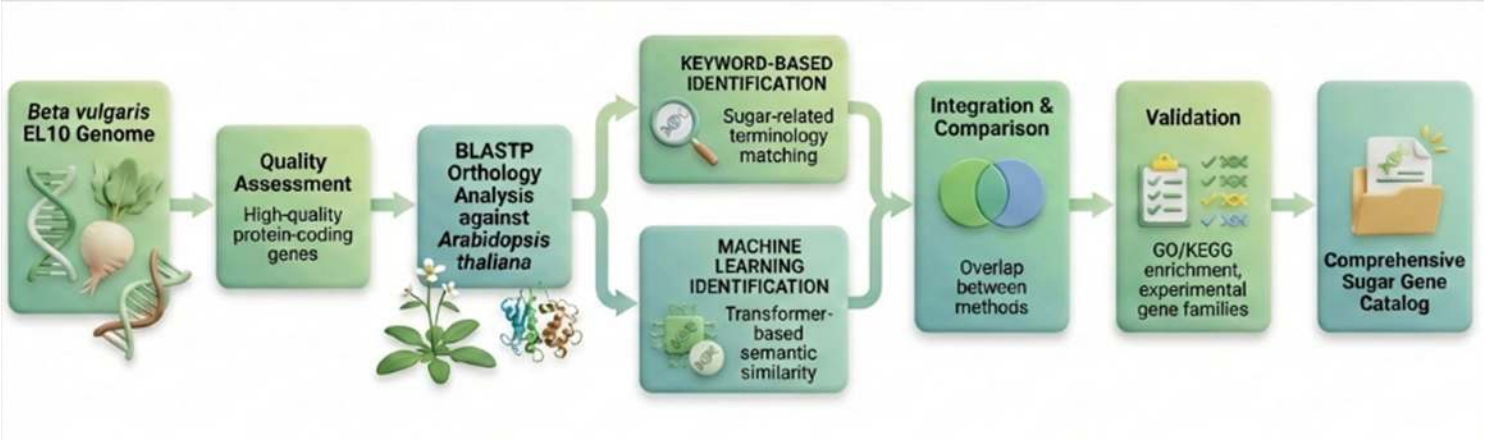



**Figure 1**

**Integrated workflow for sugar-related gene identification in sugar beet.**

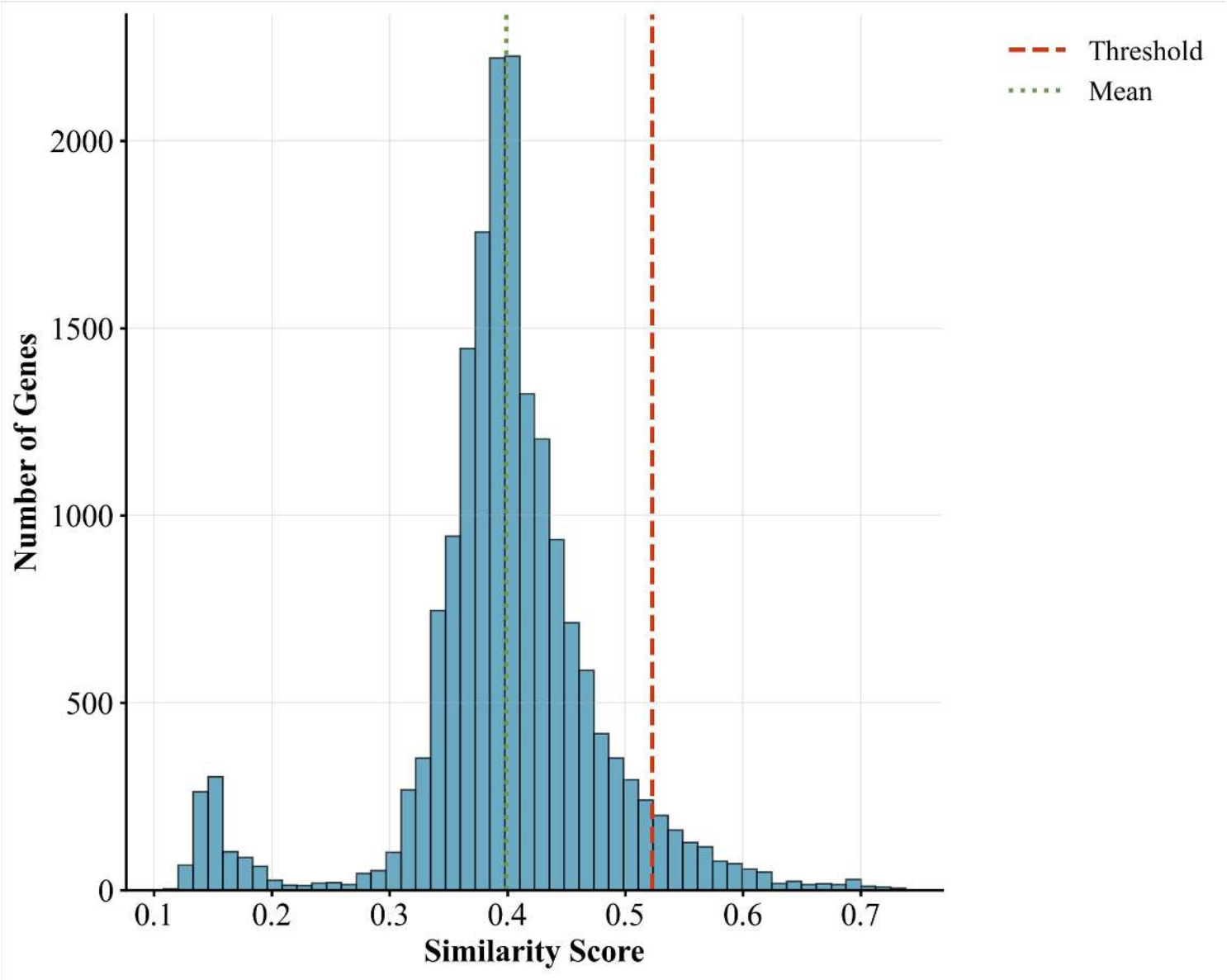



**Figure 2**

**Machine learning similarity score distribution for sugar gene identification.** Distribution of cosine similarity scores between gene functional annotations and predefined sugar-related concepts across all 18,223 validated genes. Green dotted line: mean score (0.399); red dashed line: moderate classification threshold (0.523). Genes above the threshold (n = 1,002) were classified as sugar-related.

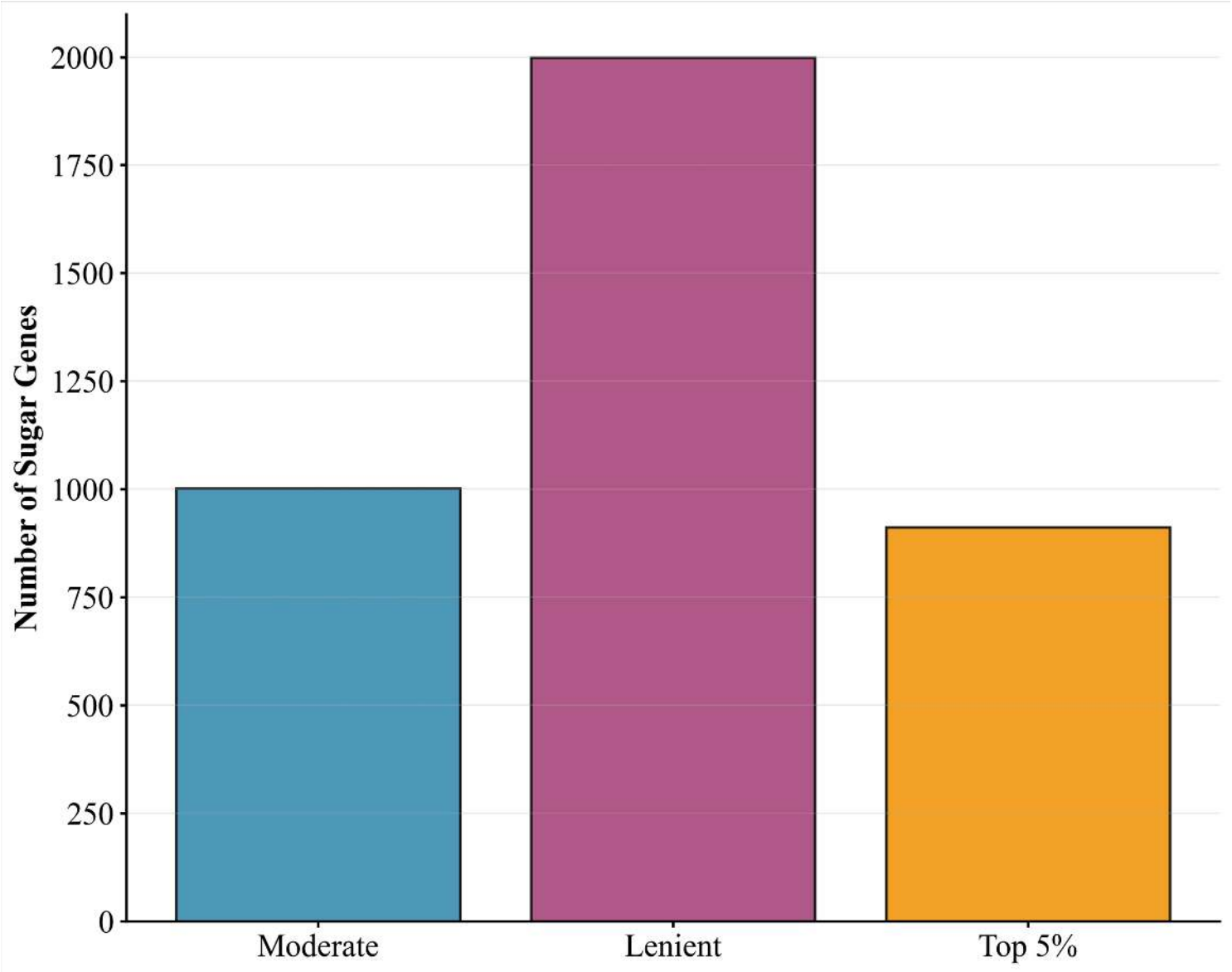
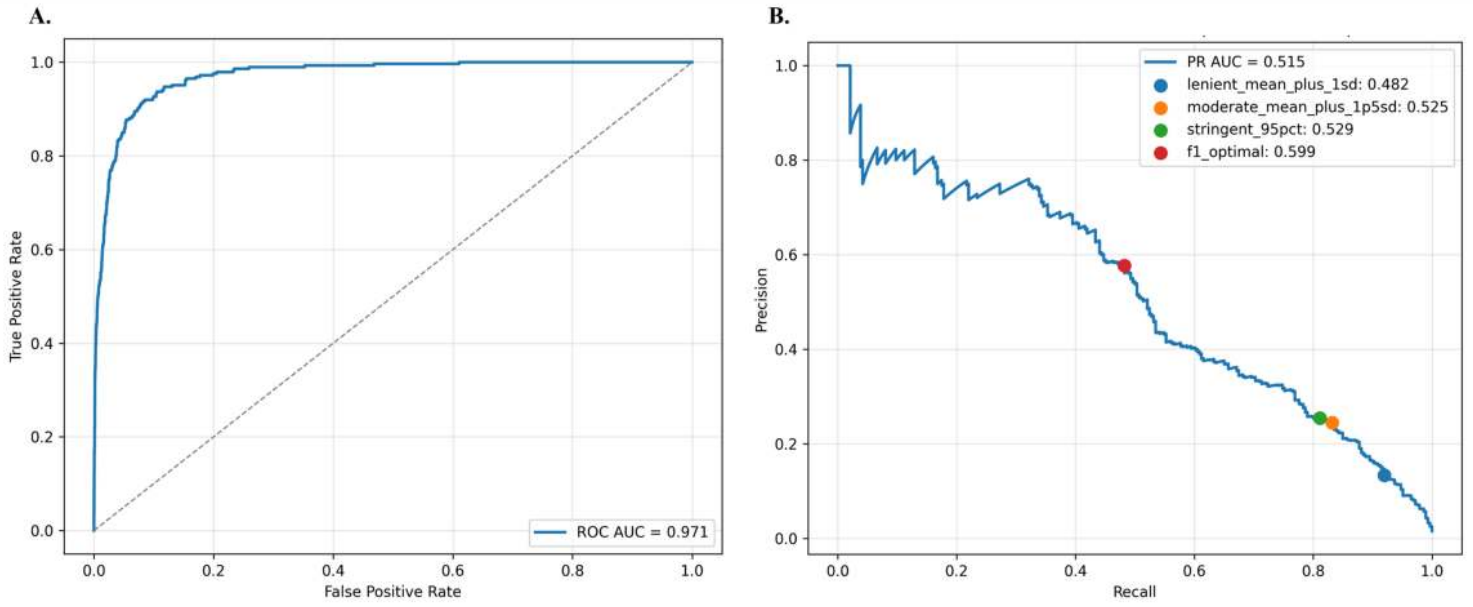



**Figure 3**

**Sugar gene identification across threshold strategies.** Number of genes classified as sugar-related using moderate (1,002 genes), lenient (1,999 genes), and top 5% stringent (912 genes) classification thresholds. Different thresholds enable flexible prioritization based on research objectives, balancing discovery breadth versus prediction confidence.




## Figure 4

**Statistical validation of machine learning classification thresholds using keyword-identified genes as ground truth.**(A) Receiver operating characteristic (ROC) curve demonstrating excellent discriminative performance (AUC = 0.971). The curve shows the trade-off between true positive rate and false positive rate across all similarity thresholds, with the dashed diagonal line representing random classification. (B) Precision-recall curve (AUC = 0.515) with threshold markers indicating performance at each classification threshold: lenient (blue; mean+1σ = 0.482), moderate (orange; mean+1.5σ = 0.525), stringent (green; 95th percentile = 0.529), and F1-optimal (red; 0.599). The moderate PR-AUC reflects the inherent class imbalance (1.7% positive rate) rather than poor model performance.

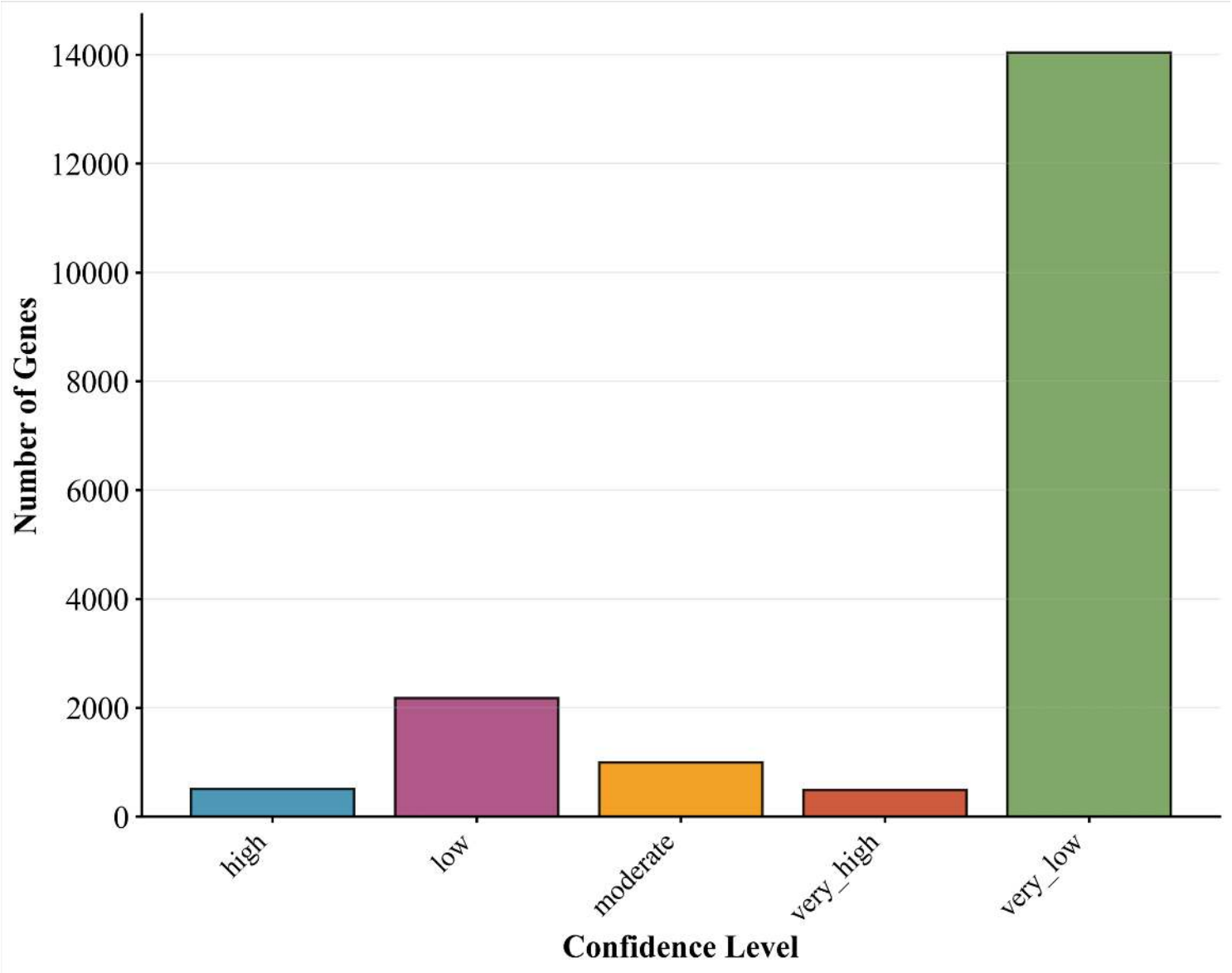



**Figure 5**

**Confidence level distribution across the sugar beet genome.** Distribution of 18,223 genes across five confidence categories: very high (492), high (510), moderate (997), low (2,181), and very low (14,043). The majority of genes show low similarity to sugar-related concepts, while 1,002 genes (5.5%) achieved high or very high confidence, representing strong candidates for sugar metabolism and transport functions.

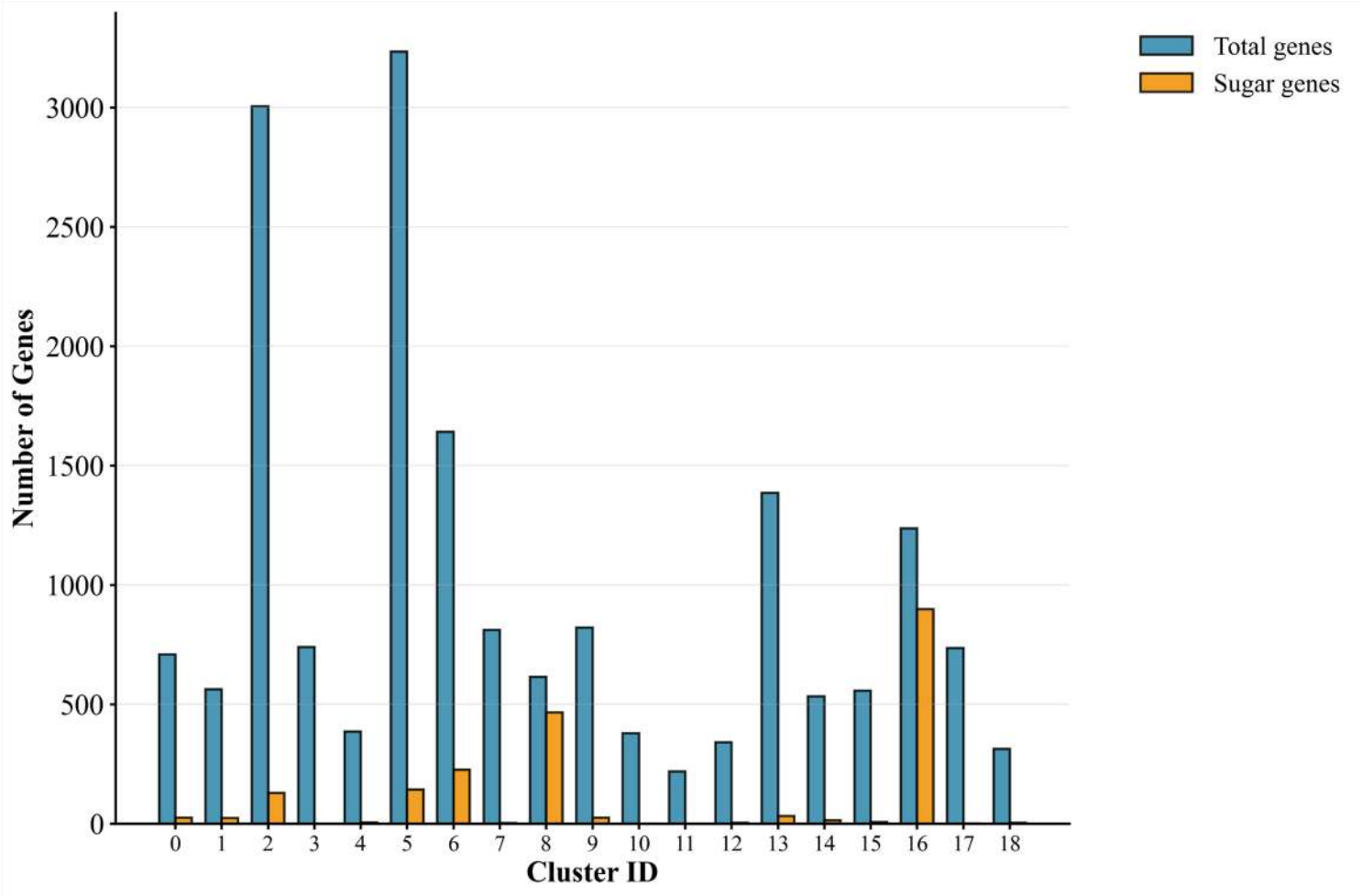



**Figure 6**

**Functional clustering reveals sugar-enriched gene groups in the sugar beet genome.** Bar chart showing the distribution of total genes (blue) and sugar-related genes (orange) across 19 k-means clusters (silhouette score: 0.210) derived from 50-dimensional principal component embeddings. Three clusters (6, 8, and 16) were designated as sugar-enriched based on containing >10% sugar-related genes. Cluster 16 is the dominant sugar-enriched cluster with approximately 1,250 genes including 900 sugar candidates (72%), followed by cluster 8 (625 genes, 450 sugar candidates, 72%) and cluster 6 (1,650 genes, 225 sugar candidates, 14%). These sugar-enriched clusters collectively contain approximately 1,589 genes, representing 79.5% of lenient threshold predictions, suggesting functionally coherent groups involved in carbohydrate metabolism, transport, and storage.

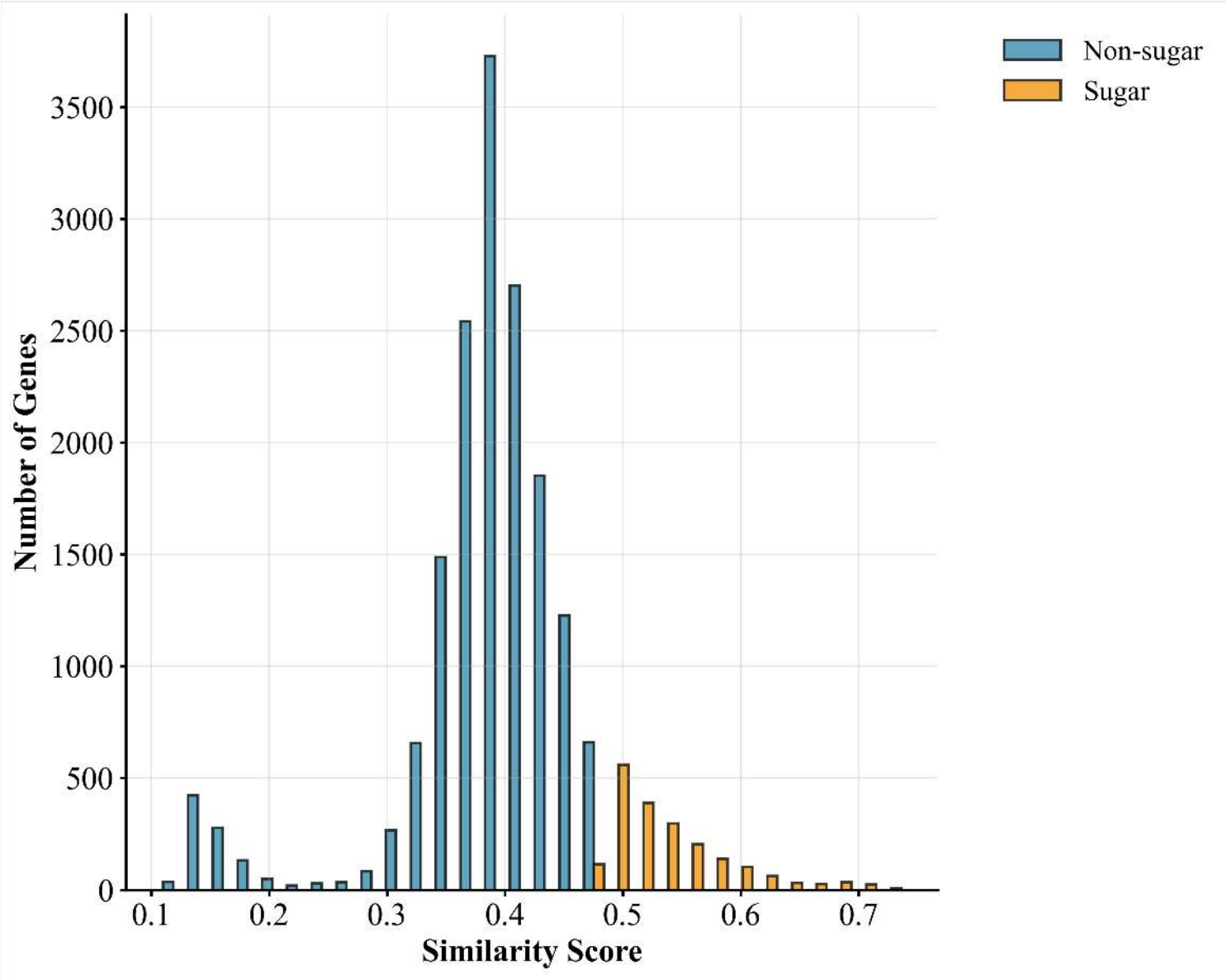



**Figure 7**

**Similarity score distributions reveal distinct patterns between sugar-related and non-sugar genes.**Overlapping histograms comparing the distribution of machine learning similarity scores for genes classified as sugar-related (orange, lenient threshold ≥0.482, n = 1,999) versus non-sugar genes (blue, n = 16,224). Sugar-related genes exhibit a right-shifted distribution with scores ranging from 0.48 to 0.74 (mean: 0.54), demonstrating consistently higher semantic similarity to predefined sugar concepts compared to the genome-wide background (scores 0.11–0.48, centered around 0.40). The clear separation between distributions validates the adaptive thresholding approach and confirms that machine learning classification captures biologically meaningful functional distinctions independent of explicit keyword presence in gene annotations.

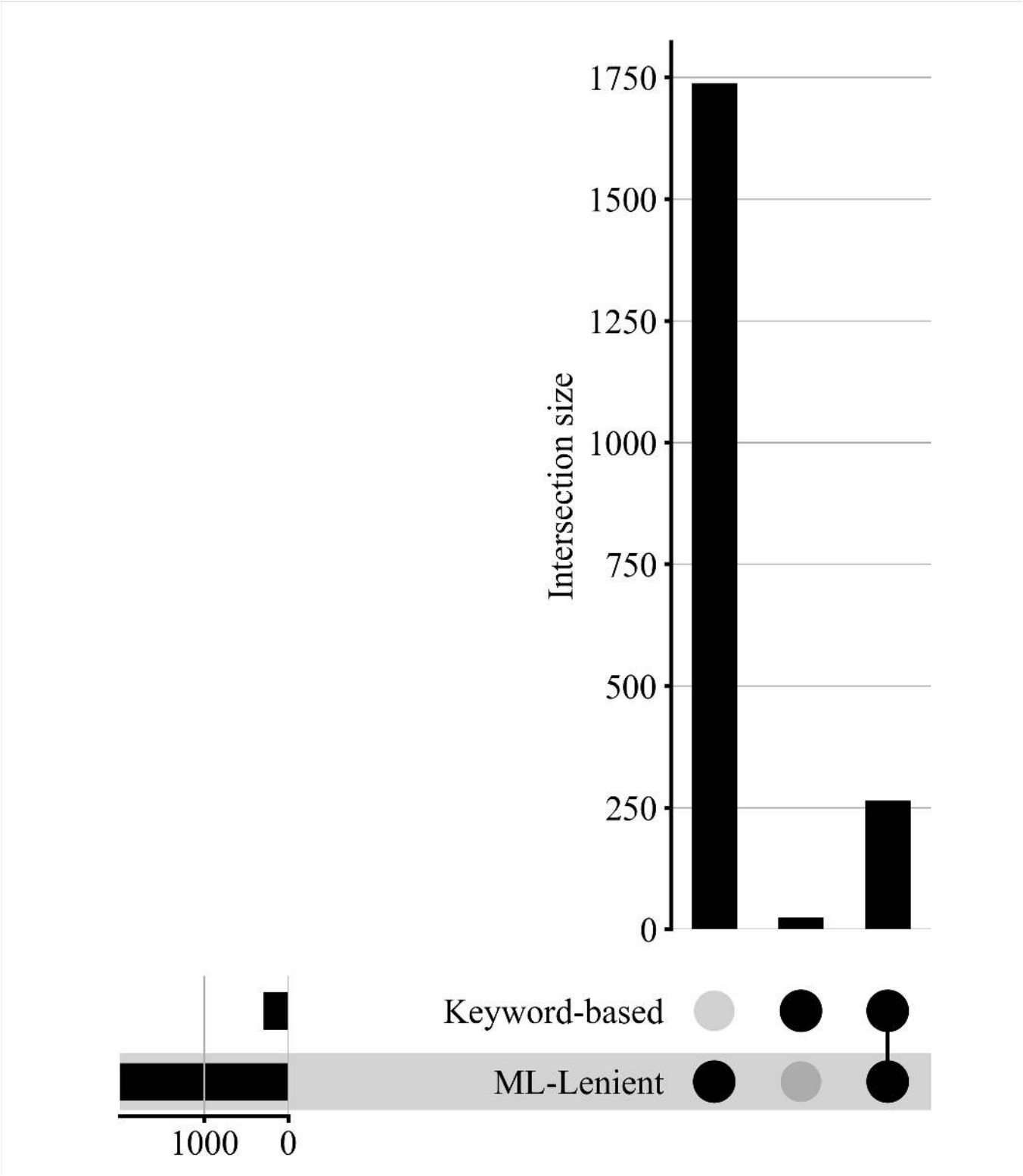



**Figure 8**

**Intersection analysis of keyword-based and machine learning sugar gene identification.**

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarytables20260130.zip