Bangladesh J. Plant Taxon. **33**(1): 1-19, 2026 (June) DOI: https://doi.org/10.3329/bjpt.v33i1.91030


# CHLOROPLAST PHYLOGENOMICS OF *SALACIA CHINENSIS* L. (CELASTRACEAE) WITH MACHINE LEARNING-ASSISTED INSIGHTS INTO ANTICANCER DRUG DISCOVERY

SHEIKH SUNZID AHMED AND M. OLIUR RAHMAN*

*Department of Botany, University of Dhaka, Dhaka 1000, Bangladesh*

*Keywords:* Comparative plastomics; Simple sequence repeats; Phylogenetics; LightGBM; AKT-Scan AI; Tanimoto similarity; Molecular docking.

**Abstract**

The present study reports the first complete chloroplast (Cp) genome of *Salacia chinensis* L. (Celastraceae), an important medicinal shrub native to Bangladesh, alongside a machine learning-driven exploration of its therapeutic potential. The circular plastome spans 157,454 bp, comprising a large single-copy of 85,757 bp, a small single-copy of 18,451 bp, and two inverted repeats of 26,623 bp each. The Cp genome encodes 127 genes, including 83 protein-coding genes, 36 tRNAs, and eight rRNAs. Comparative plastome analysis indicated a conserved genomic organization with no major structural rearrangements among the closely related members. A total of 95 simple sequence repeats were identified, predominantly mononucleotide motifs (69), suggesting potential markers for genetic diversity studies. Phylogenomic reconstruction confirmed the systematic placement of *S. chinensis* within Celastraceae. Complementing the genomic insights, a machine learning-guided anticancer drug discovery framework was employed targeting the AKT1 protein (RAC-alpha serine/threonine-protein kinase). A supervised LightGBM model achieved 90.4% accuracy with an AUC of 0.950, enabling the identification of two promising phytochemical leads, Regeol A and Carnaubadiol, exhibiting predicted bioactivities of 51.4% and 68.1%, respectively. Molecular docking analysis demonstrated strong binding affinities of –8.8 kcal/mol and –8.7 kcal/mol for Regeol A and Carnaubadiol, respectively, surpassing the reference drug (–8.0 kcal/mol), while ADMET profiling supported favorable pharmacokinetic properties with minimal toxicity concerns. In addition, we developed AKT-Scan AI (https://aktscanai.streamlit.app), a high-throughput machine learning platform for predicting AKT1-targeted bioactivity and assessing drug-likeness properties. Collectively, this integrative study enriches the genomic understanding of *S. chinensis* (GenBank Accession: PZ250435.1) and underscores its potential as a promising source of bioactive compounds for targeted therapeutic applications.

## Introduction

*Salacia chinensis* L. (Celastraceae), commonly known as lolly berry, is an important medicinal plant widely recognized for its therapeutic value in Ayurveda and other ethnomedicinal practices (Deokate and Khadabadi, 2012; Haque, 2024). The genus *Salacia* comprises a diverse group of nearly 131 species distributed across tropical and subtropical regions, including North Africa, South America, and East Asia, with notable representation in Bangladesh, India, China, Sri Lanka, Thailand, Indonesia, and Brazil (Kamat *et al.*, 2020). Within Bangladesh, *S. chinensis* occurs predominantly along the seashore, sandy river bank, and lowland primary forests, where it grows as an evergreen scandent shrub reaching up to 4 m in height. Morphologically, the species is characterized by elliptic to narrowly ovate leaves with dentate margins, axillary fascicles of 3-6 small yellowish-green flowers, and smooth, thin-pericarp fruits that turn from green to red upon ripening. The fruits are typically one-seeded, containing brown, rounded seeds (Kamat *et al.*, 2020; Haque, 2024).

*S. chinensis* has long been valued in traditional medicine, where different plant parts, including roots, and bark are used in the management of diabetes, hyperlipidemia, inflammation,

*Corresponding author. Email: <oliur.bot@du.ac.bd>

and related metabolic disorders. Phytochemical investigations have identified a wide range of bioactive constituents, including triterpenoids, flavonoids, glycosides, phenolics, and anthocyanidins, which contribute to its diverse pharmacological activities. In addition to its antidiabetic and anti-inflammatory potential, the plant has also been traditionally used as a tonic, blood purifier, and food preservative (Nikule *et al*., 2024). The broad spectrum of ethnomedicinal applications, together with growing phytochemical and pharmacological evidence, highlights *S. chinensis* as a promising source of bioactive compounds with considerable therapeutic potential, warranting further scientific exploration.

The chloroplast (Cp) genome is a crucial resource in plant molecular systematics because of its conserved genomic architecture, uniparental mode of inheritance, and comparatively slow evolutionary rate, which enable robust phylogenetic and taxonomic inference. In Celastraceae, commonly used Cp-derived barcoding regions such as *ndhF*, *rbcL*, *rpoC1*, and *matK* may provide useful but often limited resolution for closely related taxa (Simmons *et al*., 2001; Zhang and Simmons, 2006). In this context, the complete chloroplast genome has emerged as a "super-barcode," offering substantially higher discriminatory power by integrating information from the entire plastome (Chen *et al.*, 2018; Zhang *et al.*, 2019). Accordingly, whole plastome analysis of *S. chinensis* provides a comprehensive dataset encompassing both coding and non-coding regions, thereby improving species identification and phylogenetic resolution. Advances in next-generation sequencing (NGS) along with the increasing availability of public genomic data have further accelerated plastome research by enabling accurate assembly and annotation of plastomes without the need for new sequencing efforts. The reuse of publicly available datasets not only reduces costs and technical constraints but also enhances reproducibility and supports large-scale comparative, phylogenetic, and evolutionary analyses across diverse plant lineages (Park *et al.*, 2020; Ahmed and Rahman, 2025a). Collectively, these approaches provide a robust molecular framework for elucidating the evolutionary history and taxonomic placement of *S. chinensis* within the Celastraceae family.

While chloroplast genomic data provide a robust foundation for species identification and evolutionary insight, the pharmacological relevance of *S. chinensis* is largely attributed to its diverse repertoire of phytochemicals. In this context, modern computational strategies, particularly the integration of machine learning (ML) with structure-based drug design (SBDD), offer a powerful framework for translating phytochemical diversity into therapeutic applications (Islam *et al*., 2024; Murmu *et al*., 2025). Among potential molecular targets, AKT (RAC-alpha serine/threonine-protein kinase) plays a central role in cancer progression and cell survival pathways, making it a promising target for anticancer drug development (Aswathanarayan *et al*., 2026). ML-based approaches enable efficient prioritization of bioactive compounds, which can be further evaluated through SBDD techniques, including molecular docking and ADMET profiling, to assess binding interactions and pharmacokinetic properties. Thus, integrating genomic insights with computational drug discovery approaches provides a unified strategy to explore both the evolutionary significance and therapeutic potential of *S. chinensis*.

To date, the complete chloroplast genome of *S. chinensis* has not been reported, and no integrated ML-SBDD framework has been applied to identify potential AKT inhibitors from its phytocompounds. Therefore, the objectives of the present study are twofold: (i) to assemble and characterize the complete chloroplast genome of *S. chinensis* and to reconstruct its phylogenetic relationships within Celastraceae; (ii) to apply an integrated ML- and SBDD-based framework to identify and evaluate potential anticancer compounds from *S. chinensis*. This combined strategy is expected to generate valuable genomic resources while uncovering promising therapeutic lead compounds.

## Materials and Methods

### *Chloroplast genome endeavor*

High-throughput sequencing data of *Salacia chinensis* were retrieved from the NCBI SRA database (Accession ID SRX22362685). Sequencing was performed using the Illumina HiSeq X platform, generating approximately 11.4 million paired-end reads. Prior to downstream analyses, the quality and integrity of the raw reads were assessed using FastQC (Hejazi *et al.*, 2025) to ensure suitability for chloroplast (Cp) genome analysis.

### *Assembly and annotation*

The plastome was assembled using GetOrganelle v.1.7.7.0 (Jin *et al.*, 2020). Functional annotation was performed using the GeSeq platform (Tillich *et al.*, 2017) and the CPGView server (Liu *et al.*, 2023). A circular map of the plastome was generated using the OGDRAW server (Greiner *et al.*, 2019). Sequencing depth of the assembled plastome was examined using UGENE v.52.1 (Okonechnikov *et al.*, 2012). The finalized and annotated plastome sequence was deposited to the NCBI Nucleotide Database under accession number PZ250435.1.

### *SSR identification and IR boundary analysis*

Simple sequence repeats (SSRs) were identified using the MISA-Web platform (Beier *et al.*, 2017). In addition, the organization and potential shifts at the junctions of the inverted repeat (IR) regions were analyzed employing the IRscope server (Amiryousefi *et al.*, 2018).

### *Comparative plastomics and phylogenetics*

The plastome of *S. chinensis* was compared with related taxa to evaluate genome-wide structural conservation using Mauve v.20150226 (Darling *et al.*, 2011). Phylogenetic position of *S. chinensis* within Celastraceae was assessed using MEGA v.11 (Tamura *et al.*, 2021). A Neighbor-Joining (NJ) tree was constructed from aligned plastome sequences with 1,000 bootstrap replicates to evaluate branch support.

### *Machine learning-guided drug design*

Bioactivity data targeting AKT1 (RAC-alpha serine/threonine-protein kinase) were retrieved from the ChEMBL database using the identifier “CHEMBL4282”. The dataset was curated by removing entries with missing values and retaining only $IC_{50}$ measurements for consistency. Molecular structures were standardized through canonical SMILES generation using RDKit (Scalfani *et al.*, 2022). Compounds were labeled as active ($IC_{50} \leq 100$ nM) or inactive ($IC_{50} \geq 1000$ nM), while intermediate values were excluded. The final dataset was randomly divided into training and test sets using an 80:20 ratio (Niharika *et al.*, 2025). Additionally, an independent external validation set was constructed from previously unused ChEMBL4282 compounds to further assess model generalizability.

### *Chemical space exploration, diversity, and physicochemical analysis*

Chemical space analysis was conducted using RDKit-generated Morgan fingerprints to assess structural diversity. Feature selection was performed employing the “SelectKBest” method based on mutual information, and the selected descriptors were used for subsequent analyses. Principal Component Analysis (PCA) was applied to reveal clustering patterns between active and inactive compounds in both the training and test datasets (Alshehri, 2023). Structural diversity was assessed using pairwise Tanimoto similarity based on Morgan fingerprints, while physicochemical properties were evaluated by plotting molecular weight (MolWt) against lipophilicity (LogP) for active and inactive compounds (Samad *et al.*, 2023).

*ML model development and training*

A supervised machine learning framework was used to classify bioactive compounds targeting AKT1. The LightGBM (Light Gradient Boosting Machine) algorithm was selected to develop the ML model due to its strong performance on high-dimensional data (Ke *et al.*, 2017). Molecular features were generated using RDKit Morgan fingerprints. Feature selection was applied using mutual information (SelectKBest), and Synthetic Minority Oversampling Technique (SMOTE) was applied to address class imbalance. Model training and hyperparameter optimization were performed using an integrated pipeline with randomized search and cross-validation (Alshehri, 2023).

*ML Model validation and virtual screening*

Model performance was evaluated using stratified five-fold cross-validation. Evaluation metrics included accuracy, precision, recall, F1-score, MCC, and ROC-AUC to ensure robust performance assessment (Alshehri, 2023). The trained model was further validated using the external dataset derived from previously unseen ChEMBL compounds to assess its generalizability (Murmu *et al.*, 2025). The optimized LightGBM model was used to screen a set of 63 phytocompounds derived from *S. chinensis*, with AT-7867 as a control drug (Arthur *et al.*, 2021). The phytochemical dataset was compiled based on reported constituents (Nikule *et al.*, 2024).

*Development and deployment of web application*

To make the trained model accessible through a user-friendly web interface, AKT-Scan AI was developed as a Streamlit-based software for rapid AKT1 bioactivity prediction from SMILES input. The app supports both single-compound and batch screening, allowing users to paste SMILES or upload CSV files. It also provides drug-likeness profiling, QED estimation, applicability-domain assessment, and scaffold similarity comparison with selected AKT1 reference drugs. Predicted active compounds can be stored, with contributor consent, in a PostgreSQL-backed database, where users can search, view, and download saved entries. The source code is available on GitHub (https://github.com/ORSA-DUBot-PTax/AKTScan-AI).

*Molecular docking and ADMET analyses*

Molecular docking was performed to evaluate the binding interactions between selected ligands and the target protein AKT1. The ligands were retrieved in SDF format from the PubChem database, and the target protein (PDB ID: 4EKL) was retrieved from the Protein Data Bank. Docking simulations were carried out using the CB-Dock2 server (Liu *et al.*, 2022) under default settings. The pharmacokinetic and toxicity profiles of the top-ranked compounds were evaluated using SwissADME (Daina *et al.*, 2017) and StopTox servers (Borba *et al.*, 2022), enabling assessment of drug-likeness and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties.

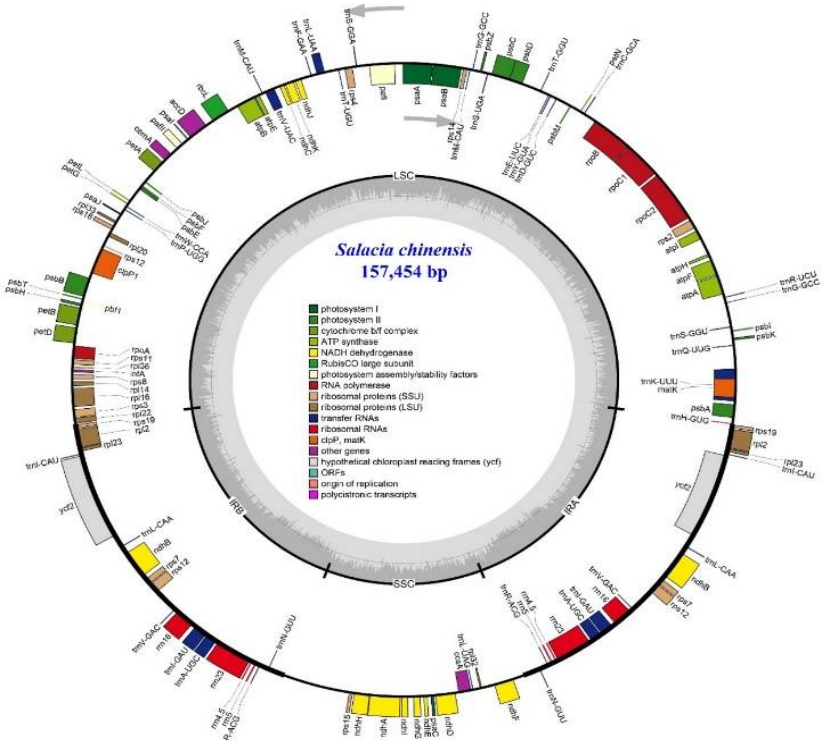
## Results and Discussion

*Quality assessment of the NGS reads*

Quality assessment of the raw NGS reads using the FastQC tool indicated that both forward and reverse reads passed all quality control metrics, with no sequences flagged as low quality. A total of 11,422,081 reads were obtained for each direction, generating approximately 1.7 Gbp of data with a read length of 150 bp and a GC content of 40%. Per-base sequence quality scores were consistently high across most read positions, with mean Phred scores exceeding Q30, indicating high base-calling accuracy. Although a slight decline in quality was observed toward the 3′ end of the reads, particularly in the reverse reads, the scores remained within acceptable limits.

*Cp genome assembly and annotation*

The complete Cp genome of *S. chinensis* was assembled as a circular molecule with a total length of 157,454 bp. It exhibited the typical quadripartite architecture, consisting of a large single-copy (LSC) region of 85,757 bp, a small single-copy (SSC) region of 18,451 bp, and a pair of inverted repeat regions (IRa and IRb), each measuring 26,623 bp (Fig. 1).




| Categories | Genes |
|---|---|
| Subunits of ATP synthase | *atpA, atpB, atpE, atpF, atpH, atpI* |
| Subunits of NADH-dehydrogenase | *ndhA, ndhB (×2), ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK* |
| Subunits of cytochrome b/f complex | *petA, petB, petD, petG, petL, petN* |
| Subunits of photosystem I | *psaA, psaB, psaC, psaI, psaJ* |
| Subunits of photosystem II | *psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbM, psbT, psbZ* |
| Large subunit of ribosome | *rpl14, rpl16, rpl2 (×2), rpl20, rpl22, rpl23 (×2), rpl32, rpl33, rpl36* |
| Small subunit of ribosome | *rps11, rps12 (×2), rps14, rps15, rps18, rps19 (×2), rps2, rps3, rps4, rps7 (×2), rps8* |
| DNA dependent RNA polymerase | *rpoA, rpoB, rpoC1, rpoC2* |
| Subunit of rubisco | *rbcL* |
| c-type cytochrome synthesis gene | *ccsA* |
| Envelop membrane protein | *cemA* |
| Maturase | *matK* |
| Subunit of acetyl-CoA-carboxylase | *accD* |
| Translational initiation factor | *infA* |
| Conserved open reading frames | *ycf2 (×2)* |
| Other genes | *clpP1, pafI, pafII, pbf1* |

Fig. 1. Plastome map showing circularized chloroplast genome of *Salacia chinensis* with its gene contents.

The overall nucleotide composition of the plastome showed a bias toward A/T bases, with an AT content of 62.46% and a GC content of 37.54%. Specifically, the genome comprised 30.86% adenine (A), 31.60% thymine (T), 18.90% cytosine (C), and 18.64% guanine (G).

Genome annotation revealed a total of 127 genes, of which 107 were unique (Fig. 1). These comprised 83 protein-coding genes (76 unique), 36 transfer RNA (tRNA) genes (27 unique), and 8 ribosomal RNA (rRNA) genes (4 unique). Gene duplication within inverted repeat regions resulted in multiple copies of several genes. Functionally, the annotated genes were classified into distinct categories based on their roles in chloroplast metabolism and gene expression. Photosynthesis-related genes included those encoding subunits of photosystem I (*psa* genes), photosystem II (*psb* genes), ATP synthase (*atp* genes), and the cytochrome b/f complex (*pet* genes). Additionally, genes associated with NADH dehydrogenase activity (*ndh* genes) and carbon fixation (*rbcL*) were identified. The plastome also harbored genes involved in transcription and translation, including RNA polymerase subunits (*rpoA*, *rpoB*, *rpoC1*, *rpoC2*), ribosomal proteins of both large (*rpl*) and small (*rps*) subunits, and multiple tRNA and rRNA genes. Furthermore, several genes with specialized or regulatory functions were detected, such as *matK* (maturase), *accD* (acetyl-CoA carboxylase subunit), *infA* (translation initiation factor), *ccsA* (cytochrome c synthesis), and *cemA* (envelope membrane protein). Conserved open reading frames, including *ycf2*, were also present, along with additional genes such as *clpP1*, *pafI*, *pafII*, and *pbf1*.

The plastome of *S. chinensis* showed a high degree of structural similarity to that of *S. amplifolia* (Lin *et al.*, 2019), reflecting the conserved organization of plastomes within the genus. However, the genome size of *S. chinensis* (157,454 bp) was slightly smaller than that of *S. amplifolia* (163,255 bp), primarily due to differences in the length of the inverted repeat (IR) regions (26,623 bp in *S. chinensis* versus 28,932 bp in *S. amplifolia*). Despite this variation, both species retained the characteristic quadripartite structure comprising LSC, SSC, and IR regions. The overall gene content of *S. chinensis* was largely comparable to that of *S. amplifolia*, with minor discrepancies likely attributable to variation in gene duplication or annotation. Additionally, both plastomes exhibited a similar A/T-rich nucleotide composition, with *S. chinensis* (62.46%) closely aligning with *S. amplifolia* (62.50%). These findings underscore the conserved nature of plastome architecture within *Salacia*, with modest differences in genome size and IR regions contributing to interspecific variation.

The plastome of *S. chinensis* harbored several intron-bearing genes undergoing cis-splicing, including *atpF*, *rpoC1*, *pafI*, *clpP1*, *petB*, *petD*, *rpl16*, *rpl2*, *ndhA*, and *ndhB*. Most of these genes contained a single intron, whereas *pafI* and *clpP1* contained two introns (Fig. 2). In addition, *rps12* was identified as a trans-spliced gene, comprising three exons distributed across different regions of the plastome, a characteristic feature commonly observed in angiosperm plastomes (Chen *et al.*, 2026).

*Coverage depth analysis*

Coverage depth analysis showed a high and uniform sequencing depth throughout the *S. chinensis* Cp genome, supporting the accuracy of assembly and base calling. The average sequencing depth reached 1,972.95X, indicating substantial read support throughout the plastome (Fig. 3). The highest coverage was observed at position 30,621, reaching 5,615X, whereas the lowest coverage occurred at position 349, with a depth of 757X. Despite minor positional variation, the overall depth remained consistently high, ensuring reliable downstream analyses. In comparison, the plastome of *S. chinensis* exhibited slightly higher average coverage than *Fraxinus griffithii* (1,824.8X), indicating a similarly robust sequencing effort (Ahmed and Rahman, 2025a). Notably, the minimum coverage in *S. chinensis* (757X) substantially exceeded that reported for *F.*

*griffithii* (302X) and *Scaphium scaphigerum* (14X), indicating improved coverage uniformity and a reduced risk of low-confidence regions (Ahmed and Rahman, 2025b). Collectively, these findings confirm the robustness and reliability of the assembled plastome of *S. chinensis*.

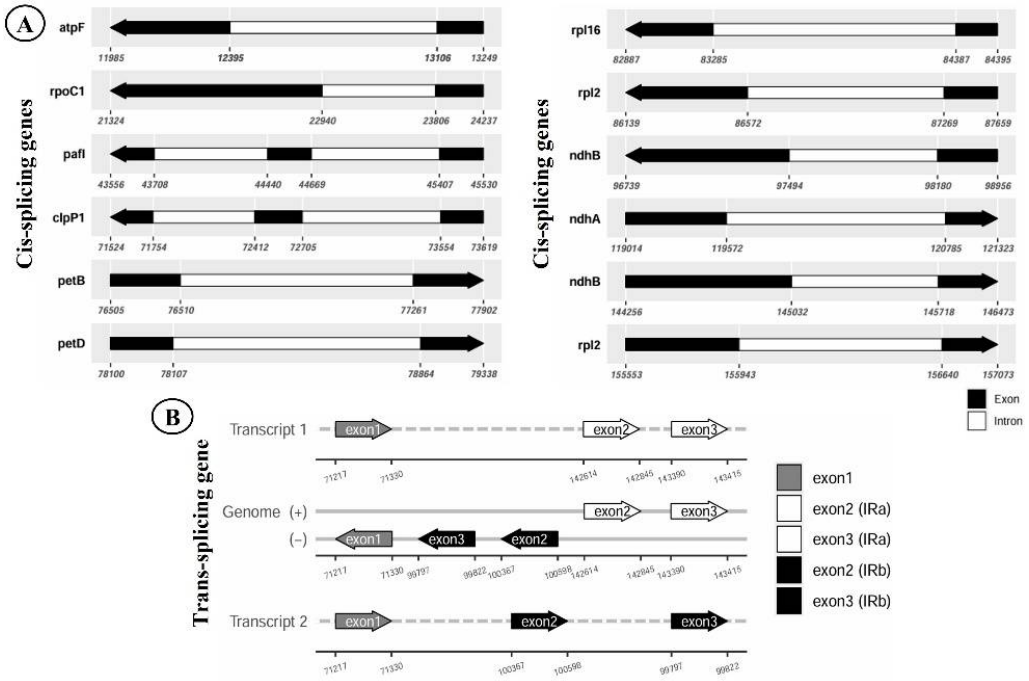



Fig. 2. Schematic representation of the cis- and trans-splicing genes in *Salacia chinensis* plastome. A. Cis-splicing genes, B. Trans-splicing gene.

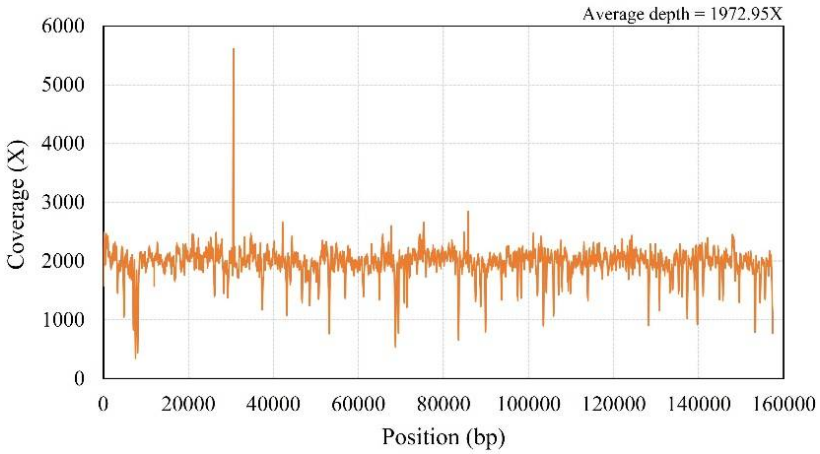



Fig. 3. Genome-wide coverage depth analysis of NGS reads aligned to the *Salacia chinensis* chloroplast genome.

*SSR and IR junction sites*

The chloroplast genome of *S. chinensis* contained a total of 95 simple sequence repeats (SSRs), which is higher than those reported for *S. menglaensis* (56), *S. amplifolia* (73), and *S. obovatilimba* (71) (Fig. 4). Among the SSR types in *S. chinensis*, mononucleotide repeats were predominant (69), followed by dinucleotide (10), trinucleotide (8), tetranucleotide (6), and hexanucleotide repeats (2), while pentanucleotide repeats were absent. This predominance of mononucleotide SSRs was also observed across other *Salacia* species, indicating a conserved genomic feature within the genus. However, variation in the total number and distribution of repeat types among species suggests potential differences in plastome evolution and genome stability (Nie *et al.*, 2025).

Comparative analysis of IR boundary regions among *Salacia* species revealed structural variation associated with expansion and contraction events. Across the examined taxa, the LSC region ranged from 85,757 to 86,922 bp, the SSC region from 18,451 to 18,553 bp, and each IR region from 26,623 to 28,953 bp (Fig. 5). In *S. chinensis*, the gene *rpl22* was located near the junction between the LSC and IRb (JLB), while *rps19* was positioned within the IR regions and occurred in duplicated copies, in IRa and IRb. Notably, the IR regions of *S. chinensis* were relatively contracted compared to closely related species such as *S. obovatilimba*, *S. amplifolia*, and *S. menglaensis*. This contraction was accompanied by a corresponding expansion of the SSC region, resulting in a comparatively larger SSC in *S. chinensis*. These findings highlight structural variation at IR boundaries within the genus, reflecting dynamic plastome evolution among *Salacia* (Ahmed and Rahman, 2025a).

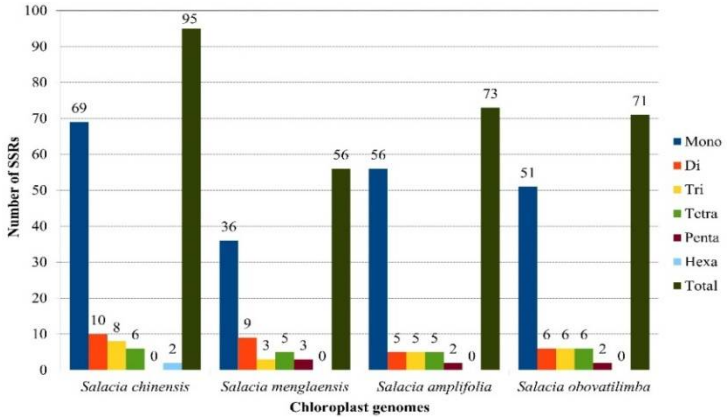



Fig. 4. Distribution of simple sequence repeats in the plastome of *Salacia chinensis* and closely related taxa.

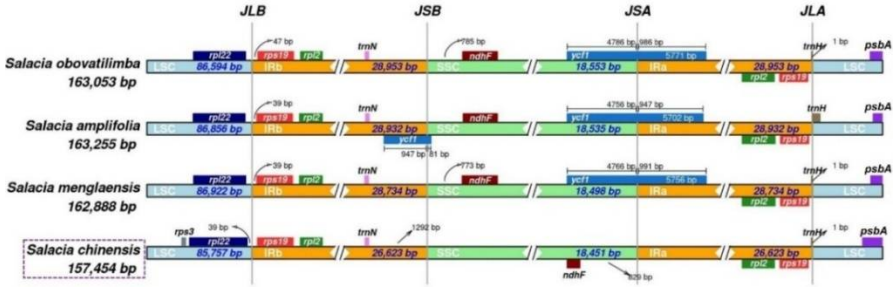



Fig. 5. Contraction of inverted repeats within the plastome of *Salacia chinensis.*

*Comparative plastomics and phylogenetics*

Comparative plastome analysis using Mauve alignment showed a high degree of structural conservation between *S. chinensis* and closely allied species, including *S. S. obovatilimba*, *S. amplifolia*, and *S. menglaensis* (Fig. 6). The local collinear block (LCB) patterns were largely conserved across all species, indicating strong synteny and minimal genomic rearrangements. These results suggest that the plastome organization within the genus *Salacia* is highly stable. Moreover, the observed alignment patterns are consistent with previously reported studies, further supporting the conserved nature of chloroplast genome architecture in related taxa (Xu *et al*., 2023; Rahman *et al*., 2025a). The major genomic features of the analyzed species are summarized in Table 1.

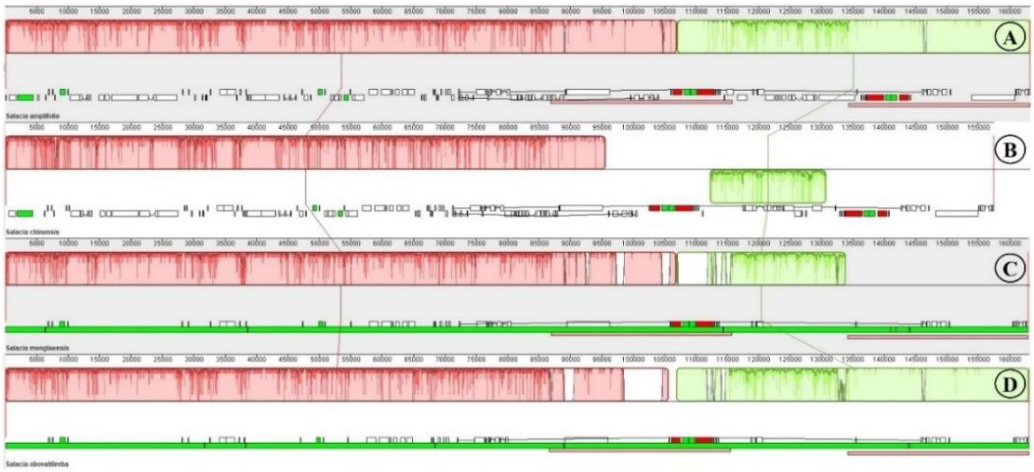



Fig. 6. Progressive alignment using Mauve elucidating local collinear blocks in the plastome of *Salacia chinensis*.

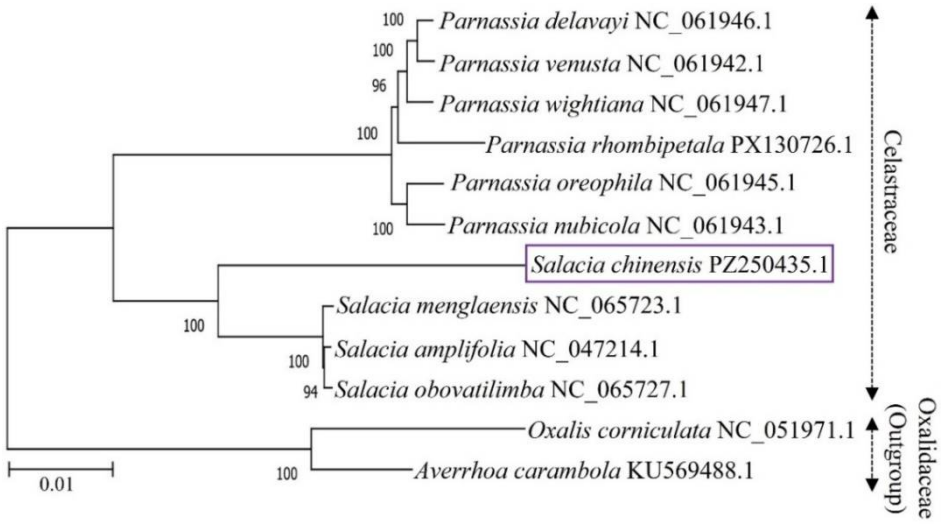



Fig. 7. Phylogenetic tree showing systematic position of *Salacia chinensis* within Celastraceae.

Phylogenetic reconstruction based on the Neighbor-Joining (NJ) method demonstrated that the genus *Salacia* formed a well-supported monophyletic clade. Within this clade, *S. chinensis* exhibited a close evolutionary affinity with *S. menglaensis*, supported by a bootstrap value of 100%, indicating strong phylogenetic confidence (Fig. 7). In addition, the *Parnassia* clade was clearly resolved as a distinct monophyletic group, reinforcing the overall tree topology. The placement of *S. chinensis* within Celastraceae was well supported, corroborating its taxonomic position and confirming the reliability of the assembled plastome. Moreover, the phylogenetic relationships were consistent with previous studies, thereby validating the robustness of the present analysis (Lin *et al*., 2019).

**Table 1. Comparative overview of the plastome characteristics of Celastraceae members analyzed in the present investigation.**

| Taxa | Accessions | Plastome (bp) | GC (%) | PCGs | tRNAs | rRNAs | Total genes |
|---|---|---|---|---|---|---|---|
| *Parnassia delavayi* | NC_061946.1 | 151,640 | 37.17 | 90 | 38 | 8 | 136 |
| *P. nubicola* | NC_061943.1 | 153,668 | 36.99 | 90 | 38 | 8 | 136 |
| *P. oreophila* | NC_061945.1 | 154,755 | 37.00 | 90 | 38 | 8 | 136 |
| *P. rhombipetala* | PX130726.1 | 152,110 | 36.96 | 87 | 36 | 8 | 131 |
| *P. venusta* | NC_061942.1 | 152,099 | 37.03 | 90 | 38 | 8 | 136 |
| *P. wightiana* | NC_061947.1 | 151,704 | 37.05 | 90 | 38 | 8 | 136 |
| *Salacia amplifolia* | NC_047214.1 | 163,255 | 37.55 | 86 | 37 | 8 | 130 |
| ***S. chinensis*** | PZ250435.1 | 157,454 | 37.54 | 83 | 36 | 8 | 127 |
| *S. menglaensis* | NC_065723.1 | 162,888 | 37.50 | 89 | 38 | 8 | 135 |
| *S. obovatilimba* | NC_065727.1 | 163,053 | 37.56 | 89 | 38 | 8 | 135 |

*Machine Learning (ML) dataset characteristics*

The curated dataset comprised 1,401 phytocompounds, all of which were valid and suitable for analysis, with no invalid SMILES detected. The dataset was partitioned into a training set of 1,120 compounds and a test set of 281 compounds, maintaining an 80:20 distribution. For external validation, an independent dataset consisting of 500 compounds (250 active and 250 inactive) derived from the ChEMBL4282 database was utilized. In total, 1,901 compounds were considered across all analyses. The inclusion of an external dataset enabled a rigorous assessment of model performance on unseen chemical space, thereby enhancing the reliability and generalizability of the predictive models (Murmu *et al.*, 2025).

*Chemical space exploration, diversity, and physicochemical analysis*

Pairwise Tanimoto similarity analysis revealed a high degree of structural diversity within the dataset. The majority of compound pairs (92.34%) exhibited very low similarity (0–0.2), while 5.04% and 1.71% were classified as low (0.2–0.4) and moderate (0.4–0.6) similarity, respectively. Compounds with high (0.6–0.8) and very high similarity (0.8–1.0) collectively accounted for less than 1% of all comparisons, confirming minimal redundancy and a broadly diverse chemical space (Fig. 8A). Physicochemical analysis based on molecular weight (MolWt) and lipophilicity (LogP) demonstrated consistent distribution patterns across both training and test datasets. In the training set, inactive compounds showed a wider MolWt range (132.16–1204.6) than active compounds (287.1–695.88), although their mean values were relatively comparable (458.17 and 485.6,

respectively) (Fig. 8B). A similar trend was observed in the test set, where inactive compounds exhibited greater variability in MolWt (196.23–1092.36) than active compounds (282.3–679.83), with mean values remaining close (454.43 and 485.09, respectively) (Fig. 8C). For LogP, inactive compounds displayed a wider range, particularly in the training set (−15.72 to 12.66), whereas active compounds were more constrained (0.71 to 6.82). The active compounds exhibited slightly higher mean LogP values across both datasets, indicating moderately increased lipophilicity.

Principal Component Analysis (PCA) further supported the structural diversity of the dataset, revealing a broad dispersion of compounds in reduced dimensional space (Fig. 8D). The first two principal components (PC1 and PC2) accounted for 21.02% and 8.20% of the total variance, respectively, with a cumulative contribution of 29.22%. The distribution pattern indicated no significant clustering, reflecting heterogeneous structural characteristics among the compounds. These findings are consistent with previous studies demonstrating high structural diversity, broad physicochemical distribution, and well-dispersed chemical space in bioactive compound datasets (Alshehri, 2023; Samad *et al.*, 2023).

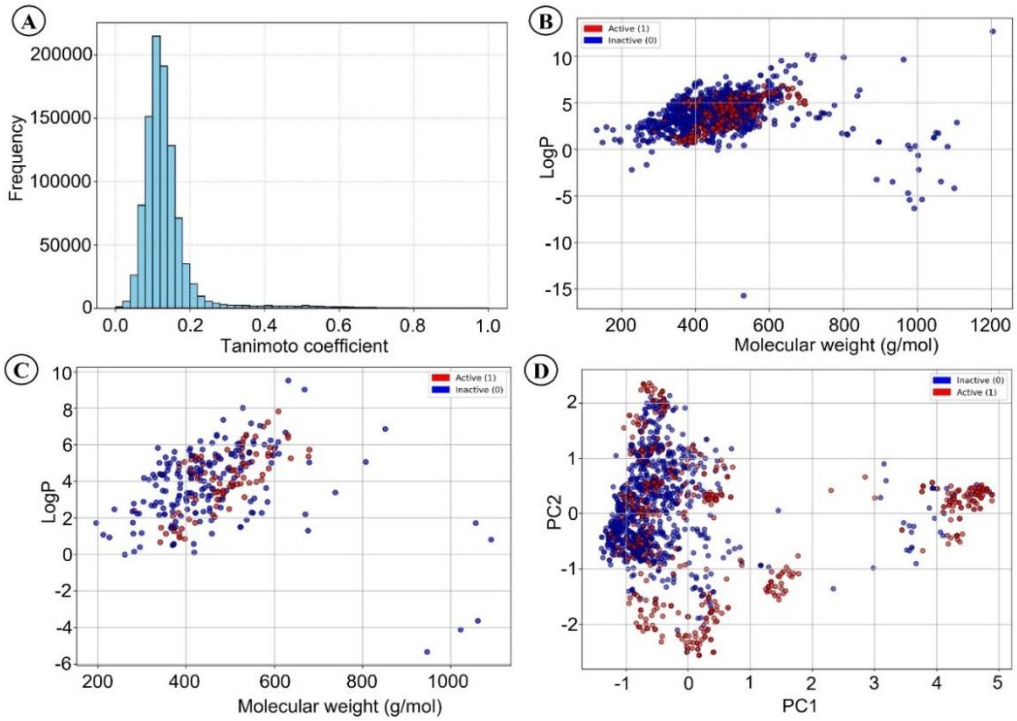



Fig. 8. Chemical space, diversity, and PCA analyses of the dataset used for machine learning model development. A. Tanimoto similarity distribution plot, B. Molecular weight versus lipophilicity for the training dataset, C. Molecular weight versus lipophilicity for the test dataset, D. PCA scatter plot.

*ML model performance: Five-fold cross-validation and external validation*

The performance of the LightGBM model was evaluated using five-fold cross-validation and an independent external validation dataset. During cross-validation, the model demonstrated strong predictive capability, achieving an accuracy of 0.904, F1 score of 0.861, area under the receiver operating characteristic curve (AUC) of 0.950, and Matthews correlation coefficient

(MCC) of 0.789 (Table 2). The sensitivity and specificity were 0.890 and 0.911, respectively, indicating a well-balanced classification performance. The confusion matrices across all five folds showed consistent classification patterns characterized by a high number of true positives and true negatives and relatively low misclassification rates, thereby confirming the stability and robustness of the model (Fig. 9).

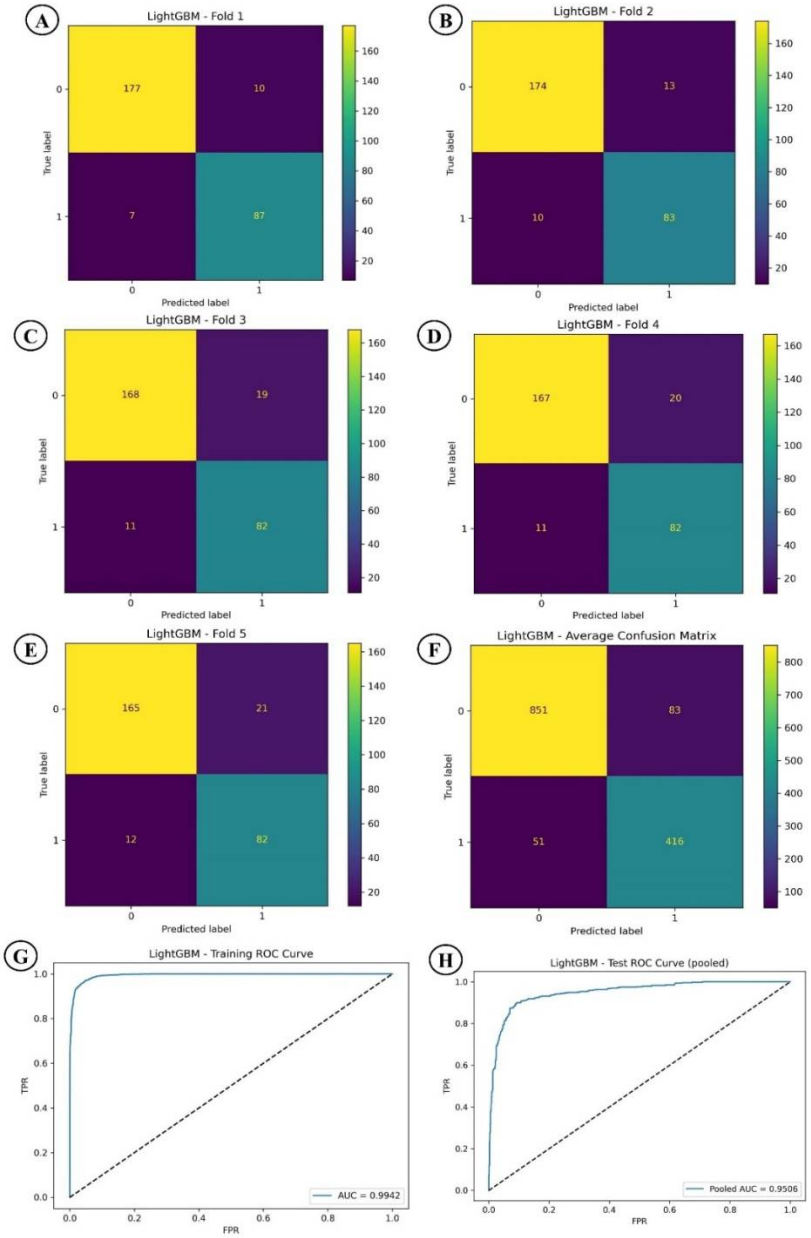



Fig. 9. Performance evaluation of the LightGBM model. A-F. Foldwise and average confusion matrix, G. ROC-AUC plot for training dataset, H. ROC-AUC plot for test dataset.

**Table 2. Performance of the machine learning model (LightGBM) evaluated using five-fold cross-validation and external validation datasets across multiple metrics**.

| Metrices | Five-fold Cross Validation | External Validation |
|---|---|---|
| Accuracy | 0.904 | 0.740 |
| F1 Score | 0.861 | 0.657 |
| AUC | 0.950 | 0.818 |
| MCC | 0.789 | 0.547 |
| Sensitivity | 0.890 | 0.500 |
| Specificity | 0.911 | 0.980 |

In external validation, the model retained satisfactory predictive performance, with an accuracy of 0.740, F1 score of 0.657, AUC of 0.818, and MCC of 0.547. Notably, the model exhibited high specificity (0.980), indicating a strong ability to correctly identify inactive compounds. However, sensitivity decreased to 0.500, suggesting a reduced capacity to detect active compounds in the external dataset (Table 2). Overall, the LightGBM model demonstrated good generalizability, with a moderate decline in performance on independent data. The findings are consistent with previous studies (Alshehri, 2023; Samad *et al.*, 2023).

*ML model-guided bioactivity prediction*

The LightGBM model predicted the bioactivity of 63 phytocompounds derived from *S. chinensis*. Among these, four compounds were classified as active, while the remaining compounds were predicted to be inactive. The predicted active compounds included Regeol A, Carnaubadiol, Foliasalacins A1, and Ethyl gallate, with probability scores ranging from 0.514 to 0.681 (Table 3). In comparison, the reference compound AT-7867 exhibited a higher predicted probability of 0.791. These results indicate that only a small subset of phytochemicals demonstrated potential bioactivity according to the machine learning model, highlighting their relevance for further evaluation. Such a relatively low proportion of predicted active compounds is consistent with earlier reports, where similar proportions have been observed despite the use of larger ligand libraries (Alshehri, 2023; Samad *et al.*, 2023).

**Table 3. Machine learning bioactivity prediction and binding affinity of the top scoring compounds.**

| Phytochemicals | PubChem CID | Chemical formula | Molecular weight (g/mol) | LightGBM model prediction (probability) | Binding affinity (kcal/mol) |
|---|---|---|---|---|---|
| Regeol A | 10694409 | $C_{28}H_{40}O_4$ | 440.6 | Active (0.514) | –8.8 |
| Carnaubadiol | 101289745 | $C_{31}H_{54}O_2$ | 458.8 | Active (0.681) | –8.7 |
| Foliasalacins A1 | 101863230 | $C_{31}H_{54}O_2$ | 458.8 | Active (0.681) | –7.8 |
| Ethyl gallate | 13250 | $C_9H_{10}O_5$ | 198.1 | Active (0.601) | –6.1 |
| AT-7867 (control) | 11175137 | $C_{20}H_{20}C_lN_3$ | 337.8 | Active (0.791) | –8.0 |

*Application of AKT-Scan AI*

The functionality of AKT-Scan AI was evaluated using both single-compound prediction and batch-screening modes. In single-compound mode, Uprosertib was classified as active, with a predicted active probability of 0.8369, high prediction confidence, no detected PAINS/Brenk alerts, and placement within the reference-scaffold applicability domain. The application further

reported relevant drug-likeness parameters, including QED (0.6277), LogP (3.5649), Lipinski rule compliance, and molecular weight (429.2540 Da). In batch-screening mode, the Supabase-backed AKT1 Active Library contained 15,780 unique predicted active compounds, comprising 2,233 compounds from CMAUP, 159 from MeFSAT, and 13,388 from LOTUS database. These findings demonstrate the applicability of AKT-Scan AI for AKT1-focused virtual screening, compound prioritization, and active-compound library construction, while emphasizing that all predicted hits require experimental validation. The web interface of the AKT-Scan AI has been visualized in the Figure 10.

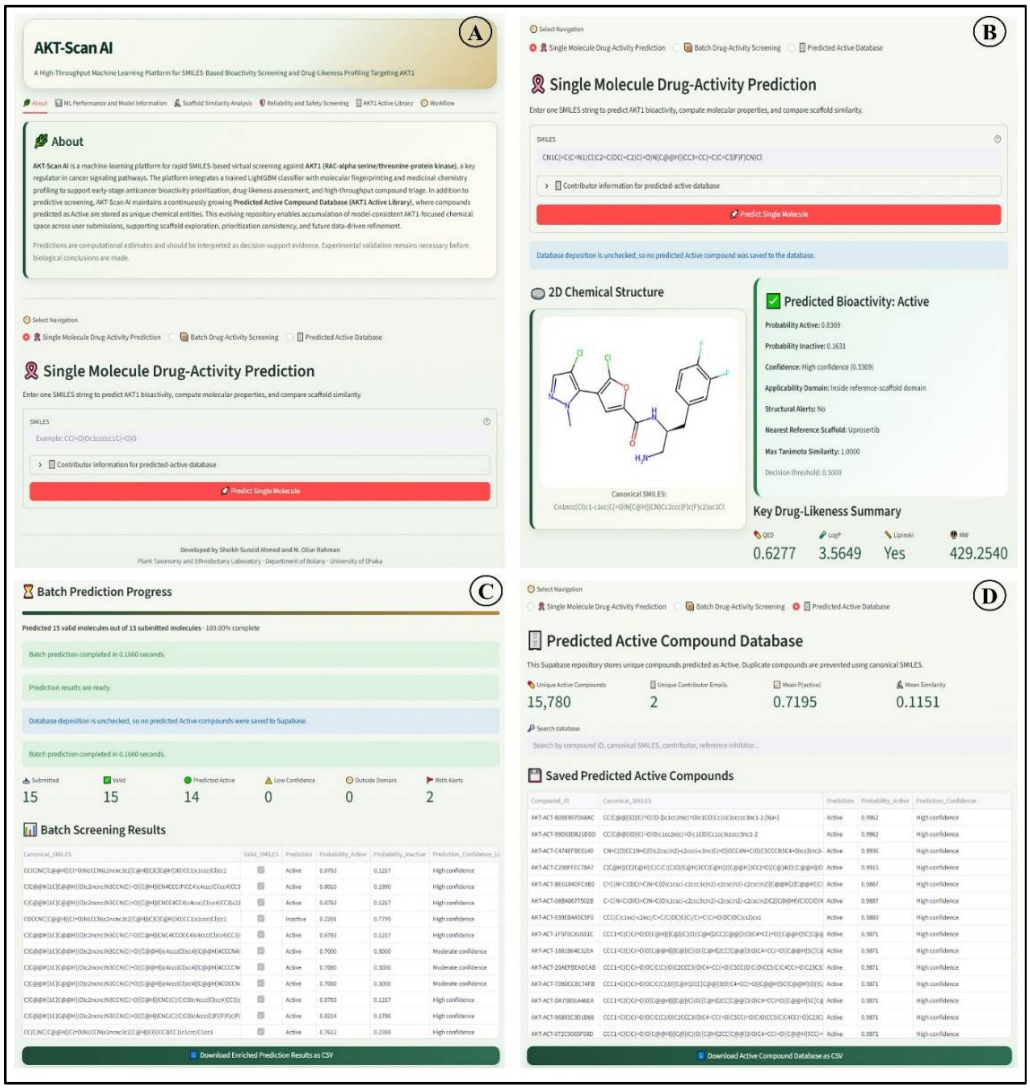



Fig. 10. AKT-Scan AI web application interface. A. Home page and single-molecule input interface. B. Uprosertib prediction output showing active classification, probability, molecular structure, and drug-likeness summary. C. Batch-screening results and downloadable prediction table. D. Supabase-backed AKT1 Active Library showing stored predicted active compounds.

*Molecular docking and ADMET analyses*

Molecular docking targeting the active site of AKT1 protein revealed that several compounds exhibited favorable binding affinities, ranging from –6.1 to –8.8 kcal/mol (Table 3). Among the screened compounds, Regeol A and Carnaubadiol demonstrated the strongest binding affinities of –8.8 kcal/mol and –8.7 kcal/mol, respectively, exceeding that of the reference inhibitor AT-7867 (–8.0 kcal/mol). These results highlight their potential as promising lead candidates for AKT1 inhibition. In contrast, Foliasalacins A1 and Ethyl gallate showed comparatively lower binding affinities of –7.8 and –6.1 kcal/mol, respectively. Interaction analysis further revealed that both Regeol A and Carnaubadiol formed stable hydrophobic interactions within the AKT1 binding pocket (Fig. 11). Regeol A interacted with key residues, including Phe161, Leu295, Asp292, and Pro313, whereas Carnaubadiol established hydrophobic contacts with Met281, His194, Val164, and Phe161. In comparison, the control compound AT-7867 interacted with multiple active-site residues, including Thr291, Glu234, Val164, Met281, Phe438, Met227, Ala177, Ala230, and Glu228, with Glu228 forming a conventional hydrogen bond while the remaining interactions were predominantly hydrophobic. These interaction patterns are consistent with previously reported site-specific docking studies (Hanson *et al*., 2024).

**Table 4. Drug-likeness and toxicity properties of the lead compounds and control drug.**

| Criteria | Ligands | Regeol A | Carnaubadiol | AT-7867 |
|---|---|---|---|---|
| Physicochemical properties | Formula | $C_{28}H_{40}O_4$ | $C_{31}H_{54}O_2$ | $C_{20}H_{20}ClN_3$ |
| | Molecular weight (g/mol) | 440.61 | 458.76 | 337.85 |
| | H-bond acceptors | 4 | 2 | 2 |
| | H-bond donors | 3 | 2 | 2 |
| | Molar refractivity | 128.73 | 143.52 | 102.20 |
| | TPSA (Å$^2$) | 77.76 | 40.46 | 40.71 |
| Lipophilicity ($logP_{0/w}$) | iLOGP | 3.15 | 4.96 | 2.39 |
| | XLOGP3 | 6.58 | 9.00 | 4.25 |
| | WLOGP | 5.42 | 7.78 | 4.02 |
| | MLOGP | 4.12 | 6.19 | 3.69 |
| | Silicos-IT Log P | 5.37 | 7.07 | 5.36 |
| | Consensus Log P | 4.93 | 7.00 | 3.97 |
| Pharmacokinetics | GI absorption | High | Low | High |
| | BBB permeant | No | No | Yes |
| | Pgp substrate | Yes | No | Yes |
| | CYP1A2 | No | No | Yes |
| | CYP2C19 | No | No | No |
| | Log Kp (cm/s) | -4.32 | -2.71 | -5.34 |
| Water solubility (ESOL) | Log S | -6.86 | -8.02 | -4.94 |
| | Solubility (mg/ml) | 6.14E-05 | 4.34E-06 | 3.89E-03 |
| | Solubility (mol/l) | 1.39E-07 | 9.46E-09 | 1.15E-05 |
| | Class | Poorly soluble | Poorly soluble | Moderately soluble |
| Drug likeness | Lipinski (violations) | 0 | 1 | 0 |
| | Bioavailability score | 0.55 | 0.55 | 0.55 |
| Medicinal chemistry | PAINS (alerts) | 1 | 0 | 0 |
| | Synthetic accessibility | 4.85 | 5.88 | 2.36 |
| Toxicity | Acute inhalation toxicity | No | No | Yes |
| | Acute oral toxicity | No | Yes | Yes |
| | Acute dermal toxicity | No | No | No |

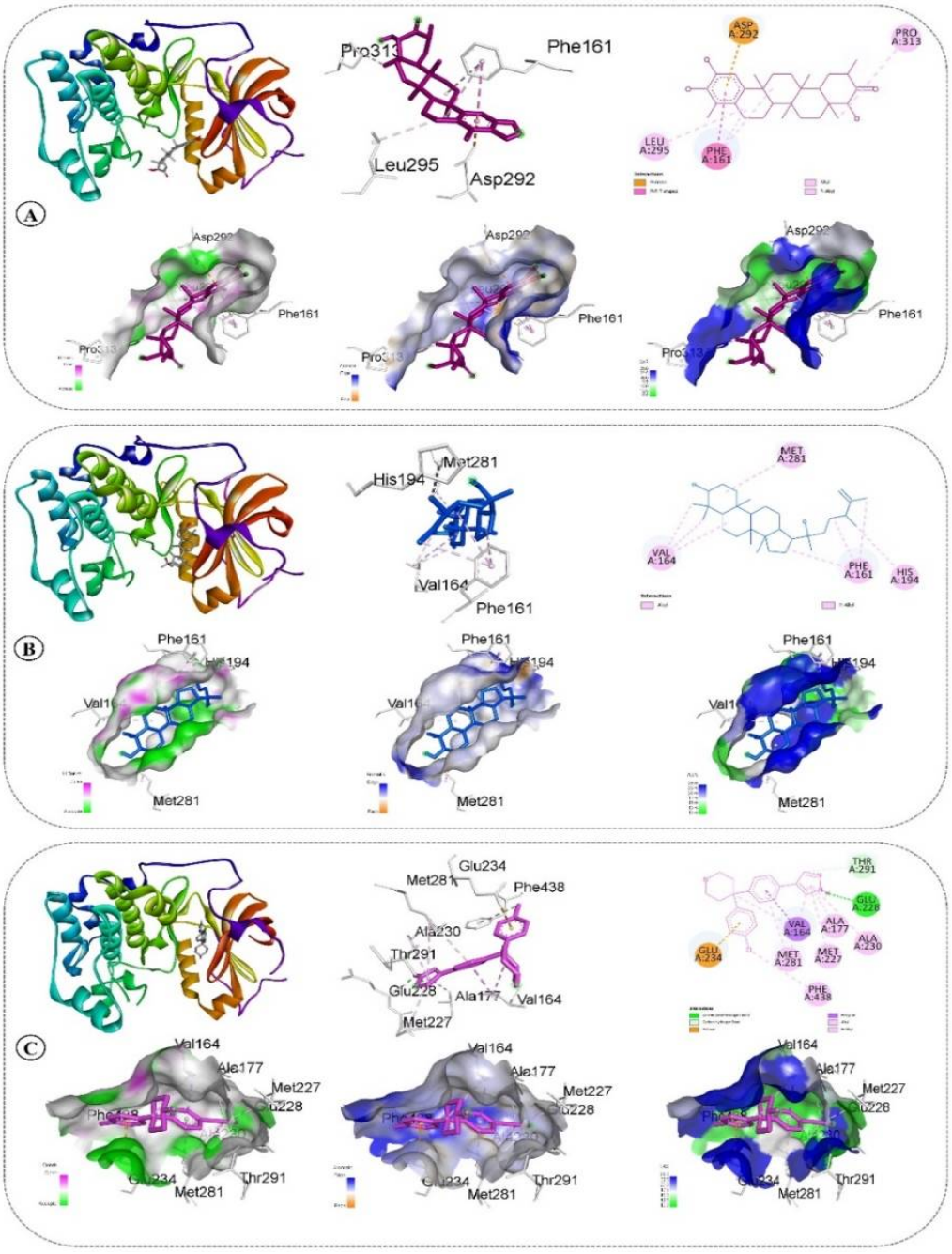



Fig. 11. Molecular docking analysis of selected lead compounds from *Salacia chinensis* and the reference inhibitor. Each compound is presented with (from left to right) the docked complex, 3D interaction view, and 2D interaction diagram in the upper panel, and hydrogen bond surface, aromatic interaction surface, and solvent-accessible surface (SAS) representation in the lower panel. A. Regeol A, B. Carnaubadiol, C. AT-7867 (control).

The predicted ADMET profiles indicated that both Regeol A and Carnaubadiol possess favorable drug-like properties compared to the reference compound AT-7867 (Table 4). Regeol A fully complied with Lipinski's rule of five, while Carnaubadiol showed only a single violation, suggesting acceptable oral drug-likeness for both compounds. In terms of pharmacokinetics, Regeol A demonstrated high gastrointestinal (GI) absorption comparable to the control, whereas Carnaubadiol exhibited lower absorption but was not predicted to be a P-glycoprotein substrate, which may reduce efflux-related limitations

Both lead compounds were predicted to be non-BBB permeant, potentially minimizing central nervous system-related side effects, in contrast to AT-7867. Additionally, neither Regeol A nor Carnaubadiol showed inhibitory effects against key cytochrome P450 enzymes, suggesting a lower risk of metabolic interactions. Although both leads exhibited lower water solubility than the control, their physicochemical properties remained within acceptable ranges for drug development. Importantly, toxicity predictions suggested an improved safety profile for the lead compounds. Regeol A showed no acute toxicity alerts across inhalation, oral, and dermal routes, while Carnaubadiol exhibited limited toxicity concerns. In contrast, AT-7867 displayed multiple toxicity risks, including inhalation toxicity and oral toxicity. These findings collectively indicate that Regeol A and Carnaubadiol possess enhanced safety and drug-likeness profiles relative to the control drug. The ADMET results are consistent with previous reports (Hanson *et al*., 2024; Rahman *et al*., 2025b; Banu *et al*., 2026).

In conclusion, this study provides the first comprehensive characterization of the complete chloroplast genome of *Salacia chinensis*, including detailed structural characterization and annotation. The assembled plastome (GenBank accession: PZ250435.1) exhibited the typical quadripartite organization and provided valuable insights into gene composition, repeat architecture, and phylogenetic placement of *S. chinensis* within the family Celastraceae. In parallel, a machine learning-guided drug discovery framework was implemented using a supervised LightGBM model to predict bioactivity against AKT1. This integrative approach identified Regeol A and Carnaubadiol as promising lead compounds, further supported by favorable molecular docking interactions and ADMET profiles. Collectively, this combined genomic and computational strategy establishes a robust platform for linking molecular systematics with rational drug discovery, while advancing our understanding of plastome evolution and genomic architecture within Celastraceae.

## References

Ahmed, S.S. and Rahman, M.O. 2025a. Complete chloroplast genome of *Fraxinus griffithii* CB Clarke (Oleaceae): Insights into genome structure and molecular phylogenetics. Bangladesh J. Plant Taxon. **32**(1): 27–44.

Ahmed, S.S. and Rahman, M.O. 2025b. Comparative genomics and phylogenetic analysis of complete chloroplast genome of *Scaphium scaphigerum* (Wall. ex G. Don) G. Planch. Dhaka Univ. J. Biol. Sci. **34**(1): 119–143.

Alshehri, F.F. 2023. Integrated virtual screening, molecular modeling and machine learning approaches revealed potential natural inhibitors for epilepsy. Saudi Pharm. J. **31**(12): 101835.

Amiryousefi, A., Hyvönen, J. and Poczai, P. 2018. IRscope: An online program to visualize the junction sites of chloroplast genomes. Bioinform. **34**(17): 3030–3031.

Arthur, D.E., Akoji, J.N., Sahnoun, R., Okafor, G.C., Abdullahi, K.L., Abdullahi, S.A. and Mgbemena, C. 2021. Computational design of novel AKT inhibitors. Netw. Model. Anal. Health Inform. Bioinform. **10**(1): 18.

Aswathanarayan, J.B., Naaz, R., Doreswamy, S.H., Karnik, M., Kumar, S., Sreenivasan, A., Sharma, A. and Madhunapantula, S.V. 2026. Advances in the understanding of AKT signaling in cancers and the potential of inhibiting AKT-driven tumors using small molecule inhibitors: An overview. Cancers **18**(4): 578.

Banu, M., Ahmed, S.S., Begum, M. and Rahman, M.O. 2026. Computational identification of epifriedelanol and derived analogs from *Mikania cordata* as potential HMG-CoA reductase inhibitors. PLoS One **21**(1): e0340573.

Beier, S., Thiel, T., Münch, T., Scholz, U. and Mascher, M. 2017. MISA-web: A web server for microsatellite prediction. Bioinform. **33**(16): 2583–2585.

Borba, J.V., Alves, V.M., Braga, R.C., Korn, D.R., Overdahl, K., Silva, A.C., Hall, S.U.S., Overdahl, E., Kleinstreuer, N., Strickland, J., Allen, D., Andrade, C.H., Muratov, E.N. and Tropsha, A. 2022. STopTox: An in silico alternative to animal testing for acute systemic and topical toxicity. Environ. Health Perspect. **130**(2): 027012.

Chen, Q., Gao, R., Liu, C. and Yao, Y. 2026. The first complete chloroplast genome of *Cosmos sulphureus* Cav. 1791 (Asteraceae) and its phylogenetic analysis. Mitochondrial DNA B **11**(5): 572–576.

Chen, X., Zhou, J., Cui, Y., Wang, Y., Duan, B. and Yao, H. 2018. Identification of *Ligularia* herbs using the complete chloroplast genome as a super-barcode. Front. Pharmacol. **9**: 695.

Daina, A., Michielin, O. and Zoete, V. 2017. SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci. Rep. **7**(1): 42717.

Darling, A.E., Tritt, A., Eisen, J.A. and Facciotti, M.T. 2011. Mauve assembly metrics. Bioinform. **27**(19): 2756–2757.

Deokate, U.A. and Khadabadi, S.S. 2012. Phytopharmacological aspects of *Salacia chinensis*. Pharmacophore **3**(3): 156–163.

Greiner, S., Lehwark, P. and Bock, R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. **47**(W1): W59–W64.

Hanson, G., Adams, J., Kepgang, D.I., Zondagh, L.S., Tem Bueh, L., Asante, A., Shirolkar, S.A., Kisaakye, M., Bondarwad, H. and Awe, O.I. 2024. Machine learning and molecular docking prediction of potential inhibitors against dengue virus. Front. Chem. **12**: 1510029.

Haque, A.K.M.K. 2024. *Salacia chinensis*. In*:* Khan, S.A. (Ed.). 2024. *Plant Red List of Bangladesh Volume 1*. Bangladesh National Herbarium, Forest Department, Ministry of Environment, Forest and Climate Change and IUCN, International Union for Conservation of Nature and Natural Resources. 446 pp.

Hejazi, F.A., Mohammadi, P. and Soorni, A. 2025. Comparative chloroplast genomics of *Teucrium* species reveals genome evolution, phylogenetic relationships and candidate molecular markers. Sci. Rep. **15**(1): 44318.

Islam, M.T., Aktaruzzaman, M., Saif, A., Hasan, A.R., Sourov, M.M.H., Sikdar, B., Rehman, S., Tabassum, A., Abeed-Ul-Haque, S., Sakib, M.H., Muhib, M.M.A., Setu, M.A.A., Tasnim, F., Rayhan, R., Abdel-Daim, M.M. and Raihan, M.O. 2024. Identification of acetylcholinesterase inhibitors from traditional medicinal plants for Alzheimer's disease using *in silico* and machine learning approaches. RSC Adv. **14**(47): 34620–34636.

Jin, J.J., Yu, W.B., Yang, J.B., Song, Y., DePamphilis, C.W., Yi, T.S. and Li, D.Z. 2020. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. **21**(1): 241.

Kamat, S.G., Vasudeva, R. and Patil, C.G. 2020. Taxonomic identity and occurrence of six species of *Salacia* and first report on chromosome numbers of *Salacia chinensis* L. and *Salacia oblonga* Wall. ex Wight and Arn. var. from Western Ghats of Karnataka (India). Genet. Resour. Crop Evol. **67**(1): 241–255.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. 2017. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. **30**: 1–9.

Lin, R.Y., Zhao, K.K., Wang, H.X., Zhu, Z.X. and Wang, H.F. 2019. Complete plastome sequence of *Salacia amplifolia* (Celastraceae): An endemic shrub in Hainan, China. Mitochondrial DNA B **4**(1): 1977–1978.

Liu, S., Ni, Y., Li, J., Zhang, X., Yang, H., Chen, H. and Liu, C. 2023. CPGView: A package for visualizing detailed chloroplast genome structures. Mol. Ecol. Resour. **23**(3): 694–704.

Liu, Y., Yang, X., Gan, J., Chen, S., Xiao, Z.X. and Cao, Y. 2022. CB-Dock2: Improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting. Nucleic Acids Res. **50**(W1): W159–W164.

Murmu, S., Aravinthkumar, A., Singh, M.K., Sharma, S., Das, R., Jha, G.K., Prakash, G., Rana, V.S., Kaushik, P. and Farooqi, M.S. 2025. Identification of potent phytochemicals against *Magnaporthe oryzae* through machine learning aided virtual screening and molecular dynamics simulation approach. Comput. Biol. Med. **188**: 109862.

Nie, Z., Ma, J., Wang, C., Tang, M., Jia, T., Liao, G. and Zhang, L. 2025. Comparative analysis of chloroplast genomes of Meliaceae species: Insights into evolution and species identification. Front. Plant Sci. **16**: 1536313.

Niharika, D.G., Salaria, P. and M, A.R. 2025. Unraveling potent *Glycyrrhiza glabra* flavonoids as AKT1 inhibitors using network pharmacology and machine learning-assisted QSAR. Mol. Divers. **29**(4): 3607–3635.

Nikule, H.A., Nikam, T.D., Borde, M.Y., Pawar, S.D., Shelke, D.B. and Nitnaware, K.M. 2024. Phytochemical and pharmacological insights into *Salacia chinensis* L. (Saptarangi): An underexplored important medicinal plant. Discov. Plants **1**(1): 67.

Okonechnikov, K., Golosova, O., Fursov, M. and UGENE Team. 2012. Unipro UGENE: A unified bioinformatics toolkit. Bioinform. **28**(8): 1166–1168.

Park, J., Xi, H. and Oh, S.H. 2020. Comparative chloroplast genomics and phylogenetic analysis of the *Viburnum dilatatum* complex (Adoxaceae) in Korea. Korean J. Plant Taxon. **50**(1): 8–16.

Rahman, M.O., Ahmed, S.S., Ali, M.A. and Lee, J. 2025a. The first chloroplast and nuclear genome assemblies of *Prasoxylon excelsum* (Meliaceae) reveal phylogenetics, evolutionary and functional insights. Bangladesh J. Plant Taxon. **32**(2): 139–160.

Rahman, M.O., Ahmed, S.S., Alqahtani, A.S., Rehman, M.T., Sultana, N., Bouhrim, M., Ali, M.A. and Lee, J. 2025b. Identification of stigmasterol derived AChE inhibitors for Alzheimer's disease using high throughput virtual screening and molecular dynamics simulations. Sci. Rep. **15**(1): 36676.

Samad, A., Ajmal, A., Mahmood, A., Khurshid, B., Li, P., Jan, S.M., Rehman, A.U., He, P., Abdalla, A.N., Umair, M., Hu, J. and Wadood, A. 2023. Identification of novel inhibitors for SARS-CoV-2 as therapeutic options using machine learning-based virtual screening, molecular docking and MD simulation. Front. Mol. Biosci. **10**: 1060076.

Scalfani, V.F., Patel, V.D. and Fernandez, A.M. 2022. Visualizing chemical space networks with RDKit and NetworkX. J. Cheminform. **14**(1): 87.

Simmons, M.P., Savolainen, V., Clevinger, C.C., Archer, R.H. and Davis, J.I. 2001. Phylogeny of the Celastraceae inferred from 26S nuclear ribosomal DNA, phytochrome B, rbcL, atpB and morphology. Mol. Phylogenet. Evol. **19**(3): 353–366.

Tamura, K., Stecher, G. and Kumar, S. 2021. MEGA11: Molecular evolutionary genetics analysis version 11. Mol. Biol. Evol. **38**(7): 3022–3027.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E.S., Fischer, A., Bock, R. and Greiner, S. 2017. GeSeq: Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. **45**(W1): W6–W11.

Xu, S., Teng, K., Zhang, H., Gao, K., Wu, J., Duan, L., Yue, Y. and Fan, X. 2023. Chloroplast genomes of four *Carex* species: Long repetitive sequences trigger dramatic changes in chloroplast genome structure. Front. Plant Sci. **14**: 1100876.

Zhang, L.B. and Simmons, M.P. 2006. Phylogeny and delimitation of the Celastrales inferred from nuclear and plastid genes. Syst. Bot. **31**(1): 122–137.

Zhang, Z., Zhang, Y., Song, M., Guan, Y. and Ma, X. 2019. Species identification of *Dracaena* using the complete chloroplast genome as a super-barcode. Front. Pharmacol. **10**: 1441.

*(Manuscript received on 10 January 2026; revised on 7 June 2026)*