

Reducing Structural Ambiguity in Food Compound Identification via Computational Refinement: A Case Study of Catechins and Their Oxidation Products in Chinese Black Tea

Shanbo Zhang

National Univerisity of Singapore

Zhi En Low

National Univerisity of Singapore

Kim Huey Ee

Mane SEA Pte Ltd

Yunle Huang

Mane SEA Pte Ltd

Rui Min Vivian Goh

Mane SEA Pte Ltd

Lingyi Li

National Univerisity of Singapore

Lionel Jublot

Mane SEA Pte Ltd

Chee Sian Gan

Agilent Technologies Singapore (Sales) Pte Ltd

Shao Quan Liu

National Univerisity of Singapore

Dachuan Zhang

National Univerisity of Singapore

Bin Yu

`gsgpbiy@gmail.com`

Mane SEA Pte Ltd

Article

Keywords: Chinese black tea, Catechins, Oxidation products, Structural identification, Metabolomics Standards Initiative, Food omics

Posted Date: June 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9816810/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: Competing interest reported. Author Dachuan Zhang serves as Associate Editor of this journal and had no role in the peer-review or decision to publish this manuscript. All other authors declare no financial or non-financial competing interests.

Version of Record: A version of this preprint was published at npj Science of Food on October 19th, 2026. See the published version at <https://doi.org/10.1038/s41538-026-01190-y>.

Abstract

Characterization of natural products and their metabolites in foods is hindered by isomeric ambiguity, limited availability of authentic standards, and unknown compounds not present from libraries. Here, this study presents a confidence-stratified identification workflow aligned with Metabolomics Standards Initiative (MSI) Levels 1–4, spanning standard-confirmed identification, putative annotation, tentative structural or class assignment, and pathway-supported characterization of unknowns. Using UHPLC-QTOF, the workflow was applied to catechins and their oxidation products in Chinese black tea. A total of 27 compounds were confirmed at MSI Level 1 using an in-house library, while 16 oxidation products were annotated at Level 2 by integrating public libraries with SIRIUS-based formula prediction, structure ranking, and class annotation. For compounds absent from libraries, a pathway-informed *in-silico* pipeline constrained structural space and improved annotation at MSI Levels 3–4. The approach was further validated using oxidation products of theasinensin A and theaflavic acid, demonstrating reduced structural ambiguity in unknown analysis.

Introduction

A substantial fraction of the chemical constituents present in foods remains structurally uncharacterized, posing a fundamental challenge for understanding their contributions to flavor, nutrition, and health ^{1–3}. This challenge is particularly pronounced in chemically complex, minimally defined natural products, where diverse transformation processes generate a vast and dynamic molecular landscape ^{4,5}. Tea (*Camellia sinensis*) is one of the most widely consumed beverages worldwide, with black tea accounting for a significant proportion of global production and trade ⁶. The taste, and health-relevant properties of black tea are influenced by its polyphenolic constituents, among which catechins represent the predominant class in fresh tea leaves ⁷. During the manufacture of black tea, catechins undergo extensive enzymatic and non-enzymatic oxidation, condensation, and polymerization, yielding a diverse array of oxidation products that are critical to the characteristic sensory and biological properties of the final product ⁸. These catechin-derived oxidation products range from well-characterized dimeric compounds such as theaflavins to structurally complex but poorly defined polymeric thearubigins, alongside epimerization-derived isomers and novel oxidation intermediates. These oxidation products render black tea one of the most analytically challenging food matrices for compound identification. Currently, although these oxidation products have potential to enrich the flavor profile and health value of black tea, the characterization of these complex oxidation products remains limited, and their comprehensive identification represents an analytical challenge that has yet to be fully resolved ⁹.

For the characterization of non-volatile compounds especially polyphenols in tea, the widely used analytical platform is liquid chromatography coupled to mass spectrometry (LC-MS), owing to its sensitivity, selectivity, and ability to detect a wide range of compounds in a single run ¹⁰. When paired with high-resolution (HR) instruments such as quadrupole time-of-flight (QTOF) or Q-Orbitrap mass

spectrometry, it enables highly accurate measurement of molecular masses and the generation of MS/MS fragmentation spectra, which together form the basis for compound identification^{11–13}. The conventional identification workflow relies on matching the accurate mass and MS/MS spectrum of a detected compound against spectral libraries — either in-house libraries built from authentic standards, or publicly available libraries such as MassBank, mzCloud and PubChem^{14–16}. When a match is found against an authentic standard measured under the same conditions, the highest level of identification confidence is achieved, corresponding to Level 1 of the Metabolomics Standards Initiative (MSI) framework¹⁷. This approach has been effectively used to characterize many of the well-known tea polyphenols, including the four major theaflavins and eight catechin monomers¹⁸. However, this approach has clear limitations in a chemically complex matrix such as black tea. Authentic standards simply do not exist for the majority of black tea oxidation products, including most thearubigins and related polymeric compounds, which share the same monomeric precursors and thus exhibit high structural similarity; consequently, a large fraction of the chemical space cannot be confirmed by standard-based matching alone¹⁹; furthermore, some genuinely novel compounds absent from all existing libraries remain undetectable by library-search approaches^{11,20}.

To address the limitations of spectral library matching, several computational tools have been developed over the past decade to predict candidate identities without requiring a matching reference spectrum. One widely used approach is molecular networking on the GNPS molecular networking platform, which groups MS/MS spectra by structural similarity and propagates annotations from known compounds to structurally related but uncharacterized neighbors²¹. Although productively applied to food and beverage matrices, molecular networking does not itself assign molecular formulae or structures, and its annotation power remains constrained by the same library coverage limitations as conventional matching. A more powerful complement is the SIRIUS software suite²², which integrates isotope pattern analysis and fragmentation tree computation to predict molecular formulae, uses CSI:FingerID to predict molecular fingerprints and retrieve candidate structures from libraries such as PubChem without requiring a reference spectrum, and extends to compound class prediction via the CANOPUS module even for compounds absent from all structural libraries²³. SIRIUS and MS-FINDER have been benchmarked directly for phenolic compound identification in food matrices, with SIRIUS showing particular advantages at the formula and compound class levels for entries absent from spectral libraries²⁴. Despite this growing uptake, a fundamental challenge remains: the annotation rate achievable by spectral library matching alone in untargeted LC-MS/MS studies of plant-derived foods is typically only 2–15% of detected features at MSI Level 2 or above²⁵, and computational tools such as SIRIUS, while extending coverage beyond library limits, depend heavily on the structural diversity represented in training data and reference libraries. For matrices dominated by processing-derived oxidation products, such as black tea, this limitation is compounded by the near-absence of tea-specific thearubigin and oxidation-product structures in public libraries, as most transformation products have not yet been characterized via LC-MS²⁶. In addition, although computational tools such as SIRIUS can predict a large number of structural candidates, they are often limited to provide more precise

annotations, particularly for structurally similar isomers²⁴. Under these circumstances, refining candidate annotations by incorporating prior knowledge of oxidation pathways and expected structural transformations is crucial. Investigating such knowledge-guided, and confidence-stratified identification strategies therefore represents an unmet need for characterizing black tea catechins and their oxidation products.

Therefore, this study aims to address structural ambiguity in catechins and their oxidation products in Chinese black tea through a confidence-stratified identification workflow. The approach integrates retention time and MS/MS fragmentation for isomer discrimination, combines SIRIUS with public spectral libraries for putative annotation, and employs a pathway-informed in silico strategy to reduce ambiguity for compounds absent from libraries, improving structural assignment of previously uncharacterized catechins and their oxidation products.

Results and discussion

Overview of compound identification strategy

Three principal analytical challenges were identified in the characterization of catechins and their oxidation products in Chinese black tea, namely isomeric complexity, lack of authentic standards, and the presence of unknown compounds (Fig. 1). To address these challenges, a tiered identification strategy was developed based on the level of available structural evidence. Compounds with authentic standards were confirmed by matching retention time and MS/MS features against an in-house library (MSI level 1). For compounds lacking standards, public spectral libraries were combined with SIRIUS-based formula prediction and structure ranking to support putative annotation (MSI level 2). For compounds absent from all existing libraries, a pathway-informed in silico candidate generation approach was integrated with SIRIUS to constrain the structural space and enable tentative structural assignment (MSI level 3–4).

Confirmed identification of catechins and their oxidation products by reference standard comparison and chromatographic resolution

Flavonoids, particularly catechins and their oxidative products, represent the predominant and analytically most studied polyphenol class in black tea²⁷. Therefore, a total of 27 compounds-including catechins, theaflavins and related flavonoids were confirmed at MSI Level 1 identification confidence through concurrent matching of retention time, accurate mass (< 5 ppm), isotope pattern, and MS/MS fragmentation pattern against our in-house spectral library (**Table S2**). Among these, catechins and their oxidation products is challenging to analyze, as this compound class encompasses multiple isomeric pairs that are indistinguishable at the precursor mass level²⁸.

The epigallocatechin gallate/gallocatechin gallate (EGCG/GCG) epimeric pair illustrates this challenge most clearly (Fig. 2). Both epimers share the molecular formula $C_{22}H_{18}O_{11}$ and produce an identical deprotonated precursor ion at $[M - H]^-$ m/z 457.0779; the overlaid isotope patterns (Fig. 2b) are completely superimposable across the M , $M + 1$, and $M + 2$ isotopologue signals, confirming that no precursor-level discrimination is possible²⁹. At the MS/MS level, both epimers produce near-identical spectra at CE = 0 and 10 eV, dominated by the intact precursor and the galloyl fragment at m/z 169.0139 ($[\text{gallic acid} - H]^-$), with no diagnostic fragment ions that permit unambiguous assignment at lower collision energies (Fig. 2c). Chromatographic separation, therefore provides the only reliable basis for discrimination: under the established UHPLC gradient, baseline separation of EGCG (RT $\approx$ 25 min) and GCG (RT $\approx$ 28 min) was achieved (Fig. 2a), consistent with the well-documented earlier elution of the 2,3-*cis* epimer relative to the 2,3-*trans* configuration on reversed-phase C18 stationary phases³⁰.

When authentic reference standards are available, isomer discrimination relies on the interplay between chromatographic and mass spectral resolution. Where both dimensions provide independent confirmation, MSI Level 1 identification is straightforward. However, for pairs such as EGCG/GCG where MS/MS evidence is insufficient, authenticated retention time becomes the sole decisive criterion—underscoring that chromatographic separation is indispensable when spectral discrimination fails³¹. This combined RT and MS/MS strategy was applied systematically across all isomeric pairs in the Level 1 compound inventory. For compounds lacking authentic standards, a different identification strategy and level was required, as described in the following section.

Putative identification of compounds lacking authentic standards via public library matching and SIRIUS-assisted annotation

For the compounds lacking authentic reference standards, putative annotation at MSI Level 2 was achieved by querying accurate precursor masses against public spectral libraries (MassBank, COCONUT, etc.) in conjunction with SIRIUS 6.3.2, integrating molecular formula prediction (SIRIUS/ZODIAC), candidate structure ranking (CSI:FingerID), and compound class annotation (CANOPUS).

Figure 3 illustrates this annotation process using epitheaflavic acid-3-gallate (putatively $C_{28}H_{20}O_{14}$, $[M - H]^-$ m/z 579.0797), a catechin oxidation product, as a representative example. The MS/MS spectrum (Fig. 3a) revealed key fragment ions including a gallic acid neutral loss (-170 Da), producing the fragment at m/z 169.0144 diagnostic of a 3-O-galloyl ester linkage, *retro*-Diels–Alder cleavage products at m/z 365.0663 and 535.0854, and a benzophenone-type fragment at m/z 229.0504 characteristic of the theaflavin benzotropolone scaffold, consistent with the findings of Liu, et al.³². The corresponding fragmentation tree (Fig. 3b) systematically connected these ions, supporting the putative assignment of this feature to epitheaflavic acid-3-gallate. This structural interpretation was corroborated at the formula level by SIRIUS (Fig. 3c), which assigned the molecular formula $C_{28}H_{20}O_{14}$ ($[M - H]^-$ m/z 579.0797) at rank 1 with a score of 99.000% and 38/38 explained MS/MS peaks, substantially reducing formula-level ambiguity relative to mass-only queries. The top-ranked CSI:FingerID structure corresponded to

epitheaflavic acid-3-gallate (Fig. 3d), with CANOPUS compound class prediction concordantly assigning a flavan-3-ol oxidation product classification, providing orthogonal support from three independent lines of computational evidence²². Applying this workflow across all features without authentic standards, a total of 16 catechin oxidation products were putatively annotated at Level 2 identification (**Table S3**), encompassing theasinensins, theacitrins, oolongtheanins, and related dimeric and galloylated derivatives³³.

Among the Level 2 annotated compounds, several isomeric pairs were putatively present. Notable examples included theasinensin B and epigallocatechin-(4 β →8)-epigallocatechin-3-O-gallate, which share the molecular formula C₃₇H₃₀O₁₈ at [M – H][–] *m/z* 761.1359, as well as theasinensin F and epigallocatechin-(4 β →8)-epicatechin-3,3'-digallate, both corresponding to C₄₄H₃₄O₂₁ at *m/z* 897.1520. These isomeric compounds were putatively distinguishable by differences in MS/MS fragmentation patterns and diagnostic fragment ion ratios, enabling SIRIUS to return the correct structure at rank 1 for each. **Figure S2** illustrates this principle with the vitexin/isovitexin pair as a representative case: although both C-glycoside positional isomers share identical molecular formulae and precursor masses, they produced distinctly different fragment ion series at CE = 10 and 20 eV, with characteristic ions at *m/z* 311.0561 and 341.0667 present at different relative abundances, and CSI:FingerID correctly returned each compound at rank 1 when processed independently. This demonstrates that when a structural difference between isomers generates detectable changes in the fragmentation tree, MS/MS-based discrimination can provide some evidence for structural assignment even when authentic standards are unavailable³⁴.

Resolving structural ambiguity in known-unknown oxidation products via importing a small-scale generated candidate list

Despite advances in annotation achievable through in-house library matching and public library-assisted SIRIUS prediction, a substantial fraction of detected chemical features in Chinese black tea poses a more fundamental ambiguity problem: compounds that are structurally unknown, absent from all existing libraries, and for which no spectral reference exists at any confidence level. This situation is not an analytical failure, but a chemical reality rooted in the manufacturing process of Chinese black tea.

During tea processing, catechins undergo enzymatic oxidation by polyphenol oxidase (PPO) and peroxidase (POD) to generate theaflavins, which subsequently undergo further iterative oxygen insertion reactions, each adding + 16 Da per step, to generate a homologous cascade series of polyhydroxylated derivatives (Fig. 4a)³⁵. More fundamentally, each mass increment of + 16 Da can arise from oxygen insertion at multiple distinct regiochemical positions within the theaflavin scaffold. For theaflavin monogallates alone, six sites are available for oxygen insertion — four in each A-ring, and two in the benzotropolone moiety — generating six structural regioisomers that share identical molecular formula and identical precursor masses, and for which no reference MS/MS spectra exist^{36,37}.

When SIRIUS processes a detected feature, the CSI:FingerID module predicts a molecular fingerprint from the fragmentation tree and searches it against the activated structural libraries, returning a ranked list of candidate structures. In practice, for oxidative cascade derivatives absent from all reference libraries, this ranking process produces a set of candidates: multiple structurally distinct compounds receive near-identical fingerprint scores because none of them is represented in the training data, and the fingerprint predictor therefore cannot meaningfully differentiate between chemically plausible regioisomers²³. The annotation output is formally a ranked list with hundreds of candidates, but sometimes the score differences between adjacent candidates are negligibly small, rendering the rank-1 assignment effectively arbitrary. In this circumstance, the candidate pool may contain structures that are entirely unrelated to black tea processing chemistry yet remain indistinguishable from genuine targets on the basis of fingerprint similarity alone. It constitutes the primary source of unresolved structural ambiguity for the theaflavin oxidative cascade series in Chinese black tea.

To resolve this ambiguity, a small-scale *in-silico* candidate library was constructed by manually drawing the two principal regioisomeric oxygenation products of theaflavin — oxygen insertion at the benzotropolone core and at the A-ring — in ChemDraw based on known PPO-mediated oxidation regioselectivity³⁸, and importing these structures into SIRIUS as a custom candidate database. The MS/MS fragmentation pattern (Fig. 4b) provided the primary structural evidence: at higher collision energies, diagnostic fragment ions at m/z 441.0831 and 365.0668 correspond to *retro*-Diels–Alder cleavage products, while the fragment at m/z 229.0504 is characteristic of an oxygenated benzotropolone scaffold, consistent with key fragment assignments for theaflavin oxidative cascade products reported by Liu, et al.³² and illustrated in **Figure S3**. When the CSI:FingerID search space was restricted to the chemically informed candidate list (Fig. 4c, lower panel), the benzotropolone-oxygenated isomer was promoted to rank 1, consistent with the known preferential reactivity of PPO at the benzotropolone moiety³⁸. It should be noted that this candidate restriction does not fully resolve regioisomeric ambiguity; rather, it serves to refine the annotation to structures that are chemically consistent with documented oxidation pathways, enabling a more precise assignment for processing-derived features that would otherwise remain entirely uncharacterized.

These results demonstrate that importing a chemically informed, pathway-constrained candidate list into SIRIUS — even a small-scale, manually constructed one — can convert a candidate ambiguity into a better resolved, confident structural assignment for processing-derived oxidation products that are entirely absent from public databases. The approach is directly scalable: as the number of candidate structures in the custom library increases to cover broader oxidation pathways and precursor scaffolds, the same principle applies, providing a systematic route to Level 3 annotation for the large fraction of Chinese black tea chemical space that currently remains inaccessible to general-purpose computational tools. However, manual construction of candidate structures in ChemDraw is inherently limited in throughput — it is feasible for a small number of well-characterized transformation products but cannot practically cover the full combinatorial space of oxidation products that characterizes the Chinese black

tea thearubigin fraction. This limitation motivates the development of a large-scale computational pipeline for systematic candidate enumeration, as described in the following section.

Large-scale candidate list generation: pipeline design

To narrow down the chemical search space, a computational pipeline is developed to process parent tea polyphenols and generate predicted candidates formed during the different steps of tea processing (Fig. 5a). A detailed write-up of the pipeline architecture is presented in Supplementary Note S1. A set of 41 parent tea polyphenols including catechins³⁸, represented by their SMILES, was processed through two parallel prediction branches: biotransformation using enzymatic reaction rules and chemical transformation using curated chemical reaction rules. Oxidation is one of the major transformations in tea processing³⁹. Hence, an Enzyme Commission (EC) class filter was applied in the biotransformation branch to restrict the reactions with oxidoreductases (EC class 1). The chemical transformation branch uses a reaction database curated from reactions known to occur during thermal processing, like drying and roasting. Both pipelines use substructure matching to ensure only chemically relevant transformations are applied. Additional filtering steps were implemented to retain chemically valid products. The specific thresholds and selection rationale are provided in Supplementary **Table S4**. The pipeline generated 2,919 predicted metabolites from biotransformation and 363 from chemical transformation, which was shown in the supplementary files. The metabolites were merged and deduplicated using canonical SMILES, yielding a final set of 3,272 predicted candidates, which were subsequently used for SIRIUS matching.

An example from each branch of the pipeline is presented to demonstrate how the predicted candidate list can improve SIRIUS matching compared with a public database search. Figure 5b shows theasinensin A predicted by the biotransformation pipeline, together with the matched product and their CSI:FingerID scores. The CSI:FingerID score reflects how well a candidate structure matches the observed fragmentation pattern. A higher score (less negative) indicates that it is a better match for the specific fragmentation pattern. The top hit from the unconstrained PubChem search, dehydrotheasinensin A, yielded a score of -246. After introducing the custom candidate list, theasinensin A product_123 and product_9 yielded scores of -134 and -172, respectively. This suggests that the curated list improved fingerprint-based matching and enabled discrimination between structurally similar products that might otherwise be collapsed into a single annotation. In addition, the unconstrained public database search space may include irrelevant candidates, reducing annotation confidence. A reaction of theaflavic acid predicted by the chemical transformation pipeline is presented in Fig. 5c. Both theaflavic acid product_64 and product_67 were identified as viable candidates, with scores comparable to the top hit from the unconstrained PubChem search. The curated list helped resolve structural ambiguity that might otherwise remain hidden without a constrained candidate space. These two case studies demonstrate that a curated candidate list can improve annotation and expose ambiguity that cannot be readily resolved by an unconstrained search.

Beyond improving candidate ranking and reducing structural ambiguity, the generated list also enables structural annotation when public database searches return no results. The distinction between Level 3 and Level 4 annotations in this study hinges on the availability of a probable structural candidate consistent with the observed MS data. Without the custom candidate list, SIRIUS can assign a molecular formula to catechins and their oxidation products (satisfying the criteria for Level 4), but cannot return a ranked structural candidate, because no relevant structures exist in the public databases against which CSI:FingerID can perform fingerprint matching. Upon importing the pathway-informed candidate list, CSI:FingerID gains access to chemically plausible structures and can rank them against the predicted molecular fingerprint. When the top-ranked candidate is consistent with the accurate mass, predicted formula, and fragmentation pattern, the annotation is elevated to Level 3, as a tentative structural candidate – albeit without spectral library or authentic standard support – is proposed on the basis of convergent mass spectral and chemical evidence. This is illustrated in **Figure S4**, where a compound feature at $[M - H]^-$ m/z 775.1140 returned no structural candidates under PubChem search alone (Level 4), but was assigned a tentative structure as (-)-epigallocatechin gallate dimer_product_20 (Rank 1, CSI:FingerID score: -122.887) upon activation of the generated candidate list (Level 3). The breadth of chemical space covered by the generated list, therefore, directly determines how many such Level 4 features can be upgraded to Level 3. Beyond improving annotation coverage, this approach also enables the identification of previously unreported compounds, potentially assessing regions of ‘chemical dark matter’ that remain underrepresented in standard reference libraries.

To characterize the structural diversity and chemical space coverage of the generated candidate list—which directly determines the breadth of Level 3 annotation achievable in this study—the predicted metabolites were evaluated using scaffold and fingerprint analyses. Murcko scaffold analysis revealed that the predicted metabolites comprised many unique scaffold families, with a small overlap between the biotransformation and chemical transformation pipelines (Fig. 6a). Calculating the diversity ratio, which is the number of unique scaffold families divided by the total number of products, further illustrates this distinction. The biotransformation pipeline yielded a ratio of 0.199, indicating that these products mostly share the same core skeletons, whereas the chemical transformation pipeline achieved a ratio of 0.744, showing that nearly all products possess distinct, unique scaffolds. This suggests that the two pipelines are complementary, thereby increasing the breadth of chemical space. Furthermore, Fig. 6b illustrates that the biotransformation pipeline tends to generate products that are structurally similar to their respective parent compounds, as evidenced by analysis of their Morgan fingerprints.

Physicochemical descriptor distributions (Fig. 6c) provide further evidence of this diversity. The definitions of the descriptors chosen are presented in Supplementary **Table S5**. Regarding the number of aromatic hydroxyl groups (fr_Ar_OH), chemical transformation products tend to retain or gain aromatic hydroxyl groups through condensation⁴⁰, whereas biotransformation products lose them. Similarly, thermal coupling reactions add aromatic units to chemical transformation products, while biotransformation reactions typically preserve the original aromatic count by modifying existing rings. The increase in chiral centers observed in chemical transformation products aligns with epimerization,

rearrangement, and condensation associated with thermal processing of tea ⁴⁰. Despite the differences in product size, both pipelines produced candidates with comparable lipophilicity. Molecular weight (MolWt) analysis further indicates that the biotransformation pipeline typically generates products clustering near the parent compound's molecular weight, consistent with single-step modifications. In contrast, the chemical transformation pipeline exhibits a bimodal distribution, indicating that it simultaneously generates both small fragments and large condensation products. Hence, due to increased hydroxyl groups and molecular size, chemical transformation products tend to be more polar than those from the biotransformation pipeline. Overall, these results demonstrated that the predicted candidate list achieved broad chemical coverage while remaining sufficiently focused on effective structural annotation of tea-derived metabolites.

In summary, a confidence-stratified identification workflow was developed and applied to characterize catechins and their oxidation products in Chinese black tea using UHPLC-QTOF. The workflow successfully confirmed 27 compounds at MSI level 1, and the putative identification of 16 catechin oxidation products at MSI level 2. For theaflavin oxidative cascade derivatives absent from all reference libraries, the incorporation of a pathway-informed candidate list into the CSI:FingerID search space reduced candidate ambiguity and enabled regiochemical assignments. Nevertheless, several limitations remain. The workflow relies on manually curated oxidation pathway rules, which may not fully capture the combinatorial complexity of thearubigin formation, and Level 3–4 assignments remain tentative in the absence of authentic standards confirmation. The annotation confidence of CSI:FingerID is also inherently constrained by the structural diversity of its training data, which remains underrepresented for tea-specific oxidation products. The confidence-stratified framework established here nonetheless provides a transferable template for the systematic characterization of oxidative polyphenol chemistry in other processed food matrices, including oolong tea, coffee, and cocoa.

Methods

Chemicals and reagents

Analytical standards used for Level 1 identification (MSI confidence) were sourced from two suppliers and a full listing of all analytical standards, their purity and sources, and CAS numbers is provided in **Table S1**.

Stock solutions of all standards were prepared in methanol or pure water and stored at – 20°C until use. LC-MS-grade water, methanol (MeOH), and acetonitrile (ACN) were purchased from Thermo Fisher Scientific (Waltham, Massachusetts, USA), while LC-MS-grade formic acid (FA) was obtained from Merck (Darmstadt, Germany).

Sample collection and sample preparation

Chinese black tea (*Camellia sinensis* var. *sinensis*) samples were obtained from Wuyishan County, Nanping City, Fujian Province, China in April 2025. Approximately 1.00 ± 0.05 g of each dried and ground tea sample was extracted with 20.0 mL of methanol/water (50:50, v/v) at 50°C under sonication for 30 min. After centrifugation at $12,000 \times g$ at 4°C for 10 min, the supernatants were filtered through 0.22- μ m PTFE membranes and stored at – 80°C until analysis. Prior to LC-MS injection, samples were appropriately diluted with mobile phase B to keep analyte signals within the detector's linear dynamic range.

UHPLC-QTOF analysis

All analyses were performed on an Agilent 1290 Infinity II UHPLC system coupled with a 6550 iFunnel Quadrupole Time-of-Flight (QTOF) mass spectrometer (Agilent Technologies, Santa Clara, CA, USA). Chromatographic separation was achieved using a Poroshell HPH-C18 column (2.1 $\times$ 150 mm, 2.7 μ m; Agilent Technologies) for profiling. The mobile phases consisted of (A) water with 0.1% formic acid and (B) methanol with 0.1% formic acid. A multi-step gradient was employed: 5% B held for 3 min, increased to 16% B over 12 min, to 18.5% B over 5 min, to 21% B over 3 min, to 23.5% B over 5 min, and then to 50% B over 10 min, followed by a wash and re-equilibration step at 95% B for 5 min. The flow rate was 0.12 mL/min, the column temperature was maintained at 40°C, and the injection volume was 1 μ L.

QTOF data were acquired in negative ionization mode using full-scan MS (m/z 50–1500, 4 spectra/s) combined with data-dependent Auto MS/MS (isolation width 1.3 m/z , up to 5 precursors per cycle; collision energies 10, 20, and 40 eV), following our previous methods²⁸. Source parameters were: gas temperature of 200°C, drying gas flow of 13 L/min, nebulizer pressure of 35 psi, sheath gas temperature of 350°C, sheath gas flow of 12 L/min, capillary voltage of 2500 V were used. Mass accuracy was maintained below 5 ppm using continuous internal reference mass correction with reference ions at m/z 112.9856 and 1033.9881. Raw LC-QTOF data were converted to the .cef format using MassHunter Qualitative Analysis (v 12.0, Agilent Technologies) and MassHunter Explorer (v 2.0) and imported into SIRIUS.

Compound identification library construction

In-house spectral library build-up

Authentic reference standards were injected into the UHPLC-QTOF under the same chromatographic and ionization conditions as those used for the study samples. For each compound, an entry was first established in Agilent MassHunter PCDL Manager (v 8.0). Subsequently, experimentally acquired data, including retention time, accurate precursor mass, molecular formula, and MS/MS spectra obtained at three collision energies (10, 20, and 40 eV) in targeted MS/MS mode, were imported and curated using Agilent MassHunter Qualitative Analysis (v 12.0). The resulting PCDL library was subsequently loaded into MassHunter Qualitative Analysis for compound identification, whereby the study sample features were matched against library entries by concurrent comparison of retention time, accurate mass (≤ 5 ppm), molecular formula, and MS/MS fragmentation pattern.

Small-scale in-silico candidate library build-up

For black tea compounds without public library support, candidate structures were manually constructed in ChemDraw (v 25.5.0; PerkinElmer) based on documented chemical oxidative transformations applied to known catechin and theaflavin scaffolds³⁵. Structures were exported as the individual .mol files and imported into Agilent ChemVista library manager (v 1.0; Agilent Technologies), where each entry was populated with compound name, molecular formula, exact mass, IUPAC International Chemical Identifier Key (InChIKey), and Simplified Molecular Input Line Entry System (SMILES). The compiled structure library was exported in the .sdf format and imported into SIRIUS 6 as a custom candidate library via the GUI import panel. Compound identification using this library was performed within the SIRIUS annotation workflow described in SIRIUS-assisted compound identification, with candidate structures searched against the in-house .sdf library alongside public libraries (MassBank, COCONUT, etc.)⁴¹.

Large-scale computational candidate library build-up

A list of parent compounds and their SMILES was processed using a two-tool *in-silico* pipeline implemented in Python (v 3.9) with RDKit (v 2025.9.2) (Fig. 5a). The compounds were transformed using enzyme-rule-based (biotransformation) and chemical-reaction-rule-based (chemical transformation) approaches. For the biotransformation pipeline, SMILES Arbitrary Target Specification (SMARTS)-based reaction rules derived from the RxnFinder database^{5,42} were pre-filtered to retain only oxidoreductase rules and applied to generate candidate metabolites. The products were subsequently filtered using Tanimoto similarity between the parent compound and the reference substrate of each matched rule, computed from Morgan fingerprints centered on the reaction-center atoms and expanded to radii 0–4. Products that are below the similarity threshold were discarded. As for the chemical transformation pipeline, a curated library of chemical reaction rules was applied via substructure matching. The predicted metabolites were filtered based on structural criteria and ranked by synthetic accessibility, retaining only structurally valid products. The outputs of both tools were then merged and deduplicated by canonical SMILES.

SIRIUS-assisted compound identification

SIRIUS (v 6.3.2; Böcker lab, Friedrich-Schiller-Universität Jena, Jena, Germany) was employed for compound identification. The software integrates four complementary computational modules: (i) SIRIUS for molecular formula prediction using isotope pattern analysis and fragmentation tree scoring²²; (ii) ZODIAC for network-based molecular formula re-ranking⁴³; (iii) CSI:FingerID for structure elucidation by fingerprint prediction and PubChem library matching²³; and (iv) CANOPUS for chemical class assignment using the ClassyFire taxonomy⁴⁴.

The *in-silico* candidate library .sdf file was imported into SIRIUS 6 via two complementary routes: the SIRIUS custom-db CLI command for reproducible batch processing (used for all final reported analyses), and the SIRIUS 6 GUI custom library import panel for iterative validation and visual inspection during

method development. Both routes read the same source .sdf file. The custom-db library was activated alongside MassBank, COCONUT and Pubchem within the SIRIUS 6 project configuration.

The annotation pipeline was executed in batch mode: (i) SIRIUS formula (mass accuracy 5 ppm, negative ionization, adduct types $[M - H]^-$, CHNO element space); (ii) ZODIAC for network-level formula re-ranking; (iii) CSI:FingerID for structure prediction against the combined in-house and public library; and (iv) SIRIUS CANOPUS (ClassyFire ontology). Results were exported via SIRIUS write-summaries and post-processed in Python (pandas v 2.1) applying: SIRIUS score $\geq 70\%$, CSI:FingerID confidence ≥ 0.3 , and explained MS^2 peaks $\geq 70\%$. The confidence score (CSI:FingerID score, scale A) and SIRIUS score were recorded for each annotation. The full SIRIUS workflow is schematically depicted in **Figure S1**.

Compound identification levels

Compound identification confidence level was assigned and modified according to the MSI framework¹⁷ as extended by Schymanski, et al.³³. Level 1 identification (Confirmed Structure) was performed as follows: RT + accurate mass + isotope pattern + MS/MS all matched against an authentic standard. Level 2 identification (Putative Candidate) was carried out as follows: MS/MS library match (MassBank, COCONUT, etc.) corroborated by SIRIUS formula/structure prediction and consistent CANOPUS class. Level 3 identification (Tentative Candidate) was conducted as follows: Accurate mass and SIRIUS molecular formula prediction consistent with a plausible structure; no spectral library matches available. Level 4 identification (Unequivocal Molecular Formula and mass of interest) was undertaken as follows: Molecular formula assigned by SIRIUS using in-house biotransformation library, structure proposed from biotransformation logic but not matched in any library.

Abbreviations

ACN, acetonitrile; CE, collision energy; CLI, command-line interface; EC, Enzyme Commission; EGCG, epigallocatechin gallate; FA, formic acid; GCG, gallic acid; GUI, graphical user interface; HR, high-resolution; InChIKey, IUPAC International Chemical Identifier Key; LC-MS, liquid chromatography coupled to mass spectrometry; MeOH, methanol; MSI, Metabolomics Standards Initiative; NMR, nuclear magnetic resonance; PCDL, Personal Compound Database and Library; POD, peroxidase; PPO, polyphenol oxidase; PTFE, polytetrafluoroethylene; QTOF, quadrupole time-of-flight; RT, retention time; SMARTS, SMILES Arbitrary Target Specification; SMILES, Simplified Molecular Input Line Entry System; UHPLC, ultra-high-performance liquid chromatography.

Declarations

Acknowledgements

This research was funded by Mane SEA PTE LTD and the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (A-8003718-00-00). The authors were grateful to Mane SEA PTE LTD for providing technical assistance during the study. The authors would also like to thank Mr. Teck Maan Lim,

Dr. Eric Tan, and Dr. Eugene Goh from Agilent Technologies Singapore (Sales) PTE LTD. for their valuable technical assistance.

Competing interests

Author Zhang Dachuan serves as Associate Editor of this journal and had no role in the peer-review or decision to publish this manuscript. All other authors declare no financial or non-financial competing interests.

Data availability statement

The data that support the findings of this study are available upon request.

Credit authorship contribution statement

Shanbo Zhang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing –original draft, Writing –review & editing.

Zhi En Low: Conceptualization, Data curation, Methodology, Investigation, Software, Writing –original draft, Writing –review & editing.

Kim Huey Ee: Conceptualization, Methodology, Formal analysis, Investigation, Project administration, Supervision, Writing –review & editing.

Yunle Huang: Methodology, Investigation, Writing –review & editing.

Rui Min Vivian Goh: Methodology, Investigation, Writing –review & editing.

Lingyi Li: Resources.

Lionel Jublot: Resources, Supervision, Project administration, Funding acquisition, Writing –review & editing.

Chee Sian Gan: Resources, Supervision, Project administration, Writing –review & editing.

Shao Quan Liu: Resources, Supervision, Project administration, Funding acquisition, Writing –review & editing.

Dachuan Zhang: Conceptualization, Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition, Writing –review & editing.

Bin Yu: Conceptualization, Formal analysis, Methodology, Investigation, Resources, Supervision, Project administration, Funding acquisition, Writing –review & editing.

References

1. Ke, Q., Zhang, J., Huang, X., Kou, X. & Zhang, D. Machine learning unveils three layers of food complexity. *npj Science of Food* 10, 87 (2026). <https://doi.org/10.1038/s41538-026-00730-w>
2. Leow, M. Z., Fu, C. & Zhang, D. Food Informatics: Leveraging data and knowledge to advance food systems. *The Innovation Informatics*, 100038 (2026). <https://doi.org/10.59717/j.xinn-inform.2026.100038>
3. Shi, P., Huang, X., Ke, Q., Kou, X. & Zhang, D. Mapping sleep-promoting volatiles in aromatic plants with machine learning: a comprehensive survey of 2300 molecules. *Digital Discovery* 5, 1068–1078 (2026). <https://doi.org/10.1039/d5dd00173k>
4. Ding, S. *et al.* From reactants to products: computational methods for biosynthetic pathway design. *Synthetic and Systems Biotechnology* 10, 1038–1049 (2025). <https://doi.org/10.1016/j.synbio.2025.05.005>
5. Zhang, D. *et al.* Data-Driven prediction of molecular biotransformations in food fermentation. *Journal of Agricultural and Food Chemistry* 71, 8488–8496 (2023). <https://doi.org/10.1021/acs.jafc.3c01172>
6. Huda, H. S. *et al.* Exploring the ancient roots and modern global brews of tea and herbal beverages: A comprehensive review of origins, types, health benefits, market dynamics, and future trends. *Food Science & Nutrition* 12, 6938–6955 (2024). <https://doi.org/10.1002/fsn3.4346>
7. Farag, M. A. *et al.* Metabolomics in tea products; a compile of applications for enhancing agricultural traits and quality control analysis of *Camellia sinensis*. *Food Chemistry* 404, 134628 (2023). <https://doi.org/10.1016/j.foodchem.2022.134628>
8. Deka, H., Sarmah, P. P., Devi, A., Tamuly, P. & Karak, T. Changes in major catechins, caffeine, and antioxidant activity during CTC processing of black tea from North East India. *RSC Advance* 11, 11457–11467 (2021). <https://doi.org/10.1039/D0RA09529J>
9. Piyasena, K. G. N. P. & Hettiarachchi, L. S. K. Identification of key flavor compounds and color substances in tea: a review. *Discover Food* 5, 250 (2025). <https://doi.org/10.1007/s44187-025-00545-w>
10. Lucci, P., Saurina, J. & Núñez, O. Trends in LC-MS and LC-HRMS analysis and characterization of polyphenols in food. *TrAC Trends in Analytical Chemistry* 88, 1–24 (2017). <https://doi.org/10.1016/j.trac.2016.12.006>
11. Chen, T. *et al.* Screening and identification of unknown chemical contaminants in food based on liquid chromatography–high-resolution mass spectrometry and machine learning. *Analytica Chimica Acta* 1287, 342116 (2024). <https://doi.org/10.1016/j.aca.2023.342116>
12. Piovesana, S. *et al.* Developments and pitfalls in the characterization of phenolic compounds in food: From targeted analysis to metabolomics-based approaches. *TrAC Trends in Analytical*

- Chemistry* 133, 116083 (2020). <https://doi.org/10.1016/j.trac.2020.116083>
13. Vinaixa, M. et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* 78, 23–35 (2016). <https://doi.org/10.1016/j.trac.2015.09.005>
 14. Zweigle, J. et al. Prioritization strategies for non-target screening in environmental samples by chromatography – High-resolution mass spectrometry: A tutorial. *Journal of Chromatography A* 1751, 465944 (2025). <https://doi.org/10.1016/j.chroma.2025.465944>
 15. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* 45, 703–714 (2010). <https://doi.org/10.1002/jms.1777>
 16. Jansen, L. J. M. et al. Systematic assessment of acquisition and data-processing parameters in the suspect screening of veterinary drugs in archive matrices using LC-HRMS. *Food Additives & Contaminants: Part A* 39, 272–284 (2022). <https://doi.org/10.1080/19440049.2021.1999507>
 17. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3, 211–221 (2007). <https://doi.org/10.1007/s11306-007-0082-2>
 18. Yu, Z. et al. Transformation of catechins into theaflavins by upregulation of CsPPO3 in preharvest tea (*Camellia sinensis*) leaves exposed to shading treatment. *Food Research International* 129, 108842 (2020). <https://doi.org/10.1016/j.foodres.2019.108842>
 19. Wawroszová, S., Čumová, M., Štěpán, R., Pluháčková, H. & Komendová, R. Development of LC-MS method enabling full chromatographic separation of 36 pyrrolizidine alkaloids in plant-based matrices. *Analytica Chimica Acta* 1375, 344577 (2025). <https://doi.org/10.1016/j.aca.2025.344577>
 20. Zhang, S. et al. A systematic approach to analyzing catechins and catechin derivatives in Ceylon black tea using liquid chromatography coupled with triple quadrupole mass spectrometry. *Food Chemistry: X* 28, 102621 (2025). <https://doi.org/10.1016/j.fochx.2025.102621>
 21. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* 34, 828–837 (2016). <https://doi.org/10.1038/nbt.3597>
 22. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* 16, 299–302 (2019). <https://doi.org/10.1038/s41592-019-0344-8>
 23. Hoffmann, M. A. et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology* 40, 411–421 (2022). <https://doi.org/10.1038/s41587-021-01045-9>
 24. Mallmann, L. P., O. Rios, A. & Rodrigues, E. MS-FINDER and SIRIUS for phenolic compound identification from high-resolution mass spectrometry data. *Food Research International* 163, 112315 (2023). <https://doi.org/10.1016/j.foodres.2022.112315>
 25. Wang, X. et al. TidyMass2: advancing LC-MS untargeted metabolomics through metabolite origin inference and metabolic feature-based functional module analysis. *Nature Communications* 17, 1755 (2026). <https://doi.org/10.1038/s41467-026-68464-7>
 26. Tang, M.-g. et al. A comprehensive review of polyphenol oxidase in tea (*Camellia sinensis*): Physiological characteristics, oxidation manufacturing, and biosynthesis of functional constituents.

27. Engelhardt, U. H. Tea chemistry – What do and what don't we know? – A micro review. *Food Research International* 132, 109120 (2020). <https://doi.org/10.1016/j.foodres.2020.109120>
28. Huang, Y. *et al.* Characterisation of catechins and their oxidised derivatives in Ceylon tea using multi-dimensional liquid chromatography and high-resolution mass spectrometry. *Journal of Chromatography A* 1682, 463477 (2022). <https://doi.org/10.1016/j.chroma.2022.463477>
29. Wisanpitayakorn, P. *et al.* Spatial mapping of stereoisomeric and isobaric alkaloids in mitragyna speciosa tissues by high-resolution DESI-cIM-MS. *Analytical Chemistry* 97, 26088–26096 (2025). <https://doi.org/10.1021/acs.analchem.5c04730>
30. Dhaubhadel, U., Patel, A. & Armstrong, D. W. Back to basic: Using ammonium hydroxide to improve peptide epimer/isomer liquid chromatography separations. *Analytica Chimica Acta* 1378, 344707 (2025). <https://doi.org/10.1016/j.aca.2025.344707>
31. Zhang, Y. *et al.* Generic and accurate prediction of retention times in liquid chromatography by post-projection calibration. *Communications Chemistry* 7, 54 (2024). <https://doi.org/10.1038/s42004-024-01135-0>
32. Liu, H. *et al.* Analysis of non-volatile and volatile metabolites during Ziziphus jujube leaf black tea processing via widely targeted metabolomics. *LWT* 205, 116507 (2024). <https://doi.org/10.1016/j.lwt.2024.116507>
33. Schymanski, E. L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environmental Science & Technology* 48, 2097–2098 (2014). <https://doi.org/10.1021/es5002105>
34. Tanaka, T. *et al.* New degradation mechanism of black tea pigment theaflavin involving condensation with epigallocatechin-3-O-gallate. *Food Chemistry* 370, 131326 (2022). <https://doi.org/10.1016/j.foodchem.2021.131326>
35. Kuhnert, N., Clifford, M. N. & Müller, A. Oxidative cascade reactions yielding polyhydroxy-theaflavins and theacitrins in the formation of black tea thearubigins: Evidence by tandem LC-MS. *Food & Function* 1, 180–199 (2010). <https://doi.org/10.1039/C0FO00066C>
36. Li, Y., Shibahara, A., Matsuo, Y., Tanaka, T. & Kouno, I. Reaction of the black tea pigment theaflavin during enzymatic oxidation of tea catechins. *Journal of Natural Products* 73, 33–39 (2010). <https://doi.org/10.1021/np900618v>
37. Zhao, Y. *et al.* Thermal activation in oxidative model drives catechin assembly into theaflavins with enhanced antioxidant capacity. *Food Chemistry* 508, 148334 (2026). <https://doi.org/10.1016/j.foodchem.2026.148334>
38. Drynan, J. W., Clifford, M. N., Obuchowicz, J. & Kuhnert, N. The chemistry of low molecular weight black tea polyphenols. *Natural Product Reports* 27, 417–462 (2010). <https://doi.org/10.1039/B912523J>

39. Teng, J. *et al.* Purification, characterization and enzymatic synthesis of theaflavins of polyphenol oxidase isozymes from tea leaf (*Camellia sinensis*). *LWT* 84, 263–270 (2017).
<https://doi.org/10.1016/j.lwt.2017.05.065>
40. Fan, F. Y. *et al.* Differential behaviors of tea catechins under thermal processing: Formation of non-enzymatic oligomers. *Food Chemistry* 196, 347–354 (2016).
<https://doi.org/10.1016/j.foodchem.2015.09.056>
41. Chandrasekhar, V. *et al.* COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Research* 53, D634-D643 (2025).
<https://doi.org/10.1093/nar/gkae1063>
42. Hu, Q. N., Deng, Z., Hu, H., Cao, D. S. & Liang, Y. Z. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. *Bioinformatics* 27, 2465–2467 (2011). <https://doi.org/10.1093/bioinformatics/btr413>
43. Ludwig, M. *et al.* Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence* 2, 629–641 (2020). <https://doi.org/10.1038/s42256-020-00234-6>
44. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* 39, 462–471 (2021).
<https://doi.org/10.1038/s41587-020-0740-8>

Figures

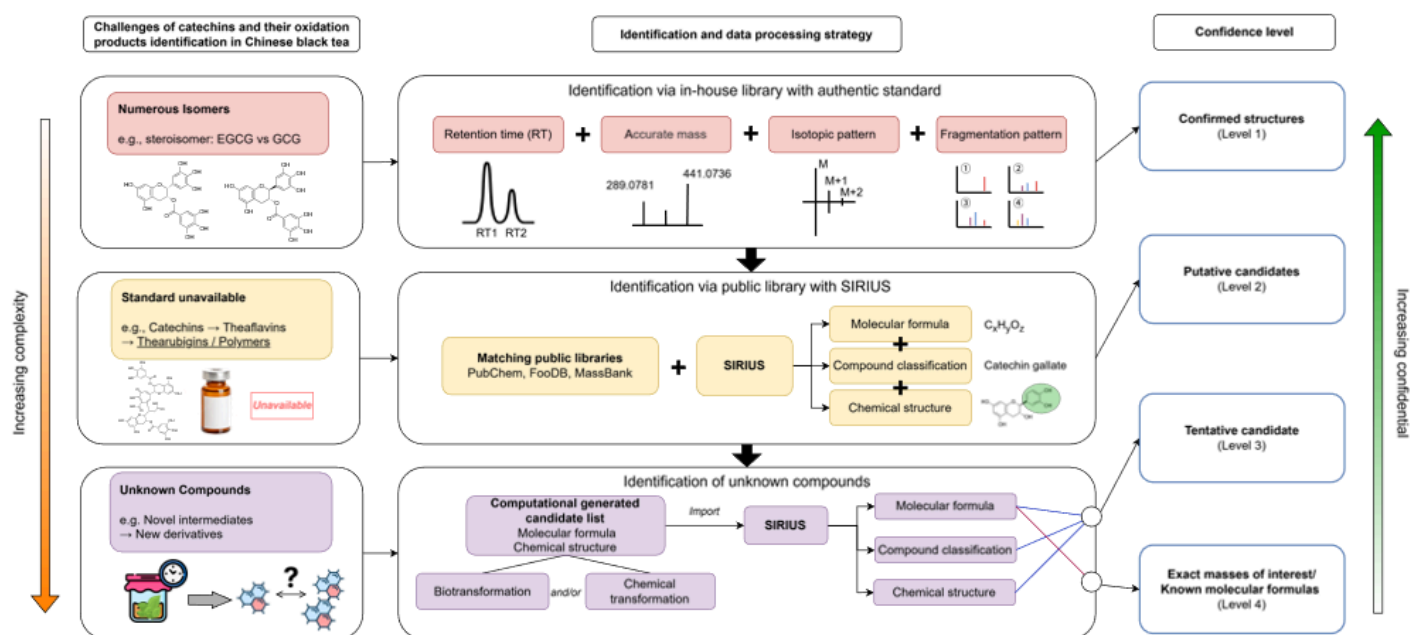


Figure 1

Challenges, identification and data processing strategy, and identification confidence level for catechins and their oxidation products in Chinese black tea. Three principal challenges are recognized (left, increasing complexity): the presence of numerous isomers (e.g., epimeric pairs such as EGCG and GCG); the unavailability of authentic reference standards for oxidative products (e.g., catechins → theaflavins → thearubigins/polymers); and the occurrence of unknown compounds including novel intermediates and new derivative structures. The corresponding identification strategies (center) are applied according to compound class and available evidence. For compounds with authentic standards, identification is based on concurrent matching of retention time, accurate mass, isotope pattern, and fragmentation pattern against an in-house library, yielding confirmed structural assignments (Level 1). For compounds lacking authentic standards, accurate masses are queried against public spectral libraries (PubChem, FooDB, MassBank) in conjunction with SIRIUS for molecular formula prediction, compound classification, and chemical structure ranking, supporting putative candidate annotation (Level 2). For genuinely unknown compounds absent from all existing databases, a computationally generated candidate list — derived from biotransformation and/or chemical transformation rules applied to known precursor structures — is imported into SIRIUS; depending on the completeness of the resulting annotation, these compounds are assigned as tentative candidates where molecular formula, compound classification, and a plausible chemical structure can be proposed (Level 3), or as exact masses of interest where only the molecular formula is retrievable (Level 4).

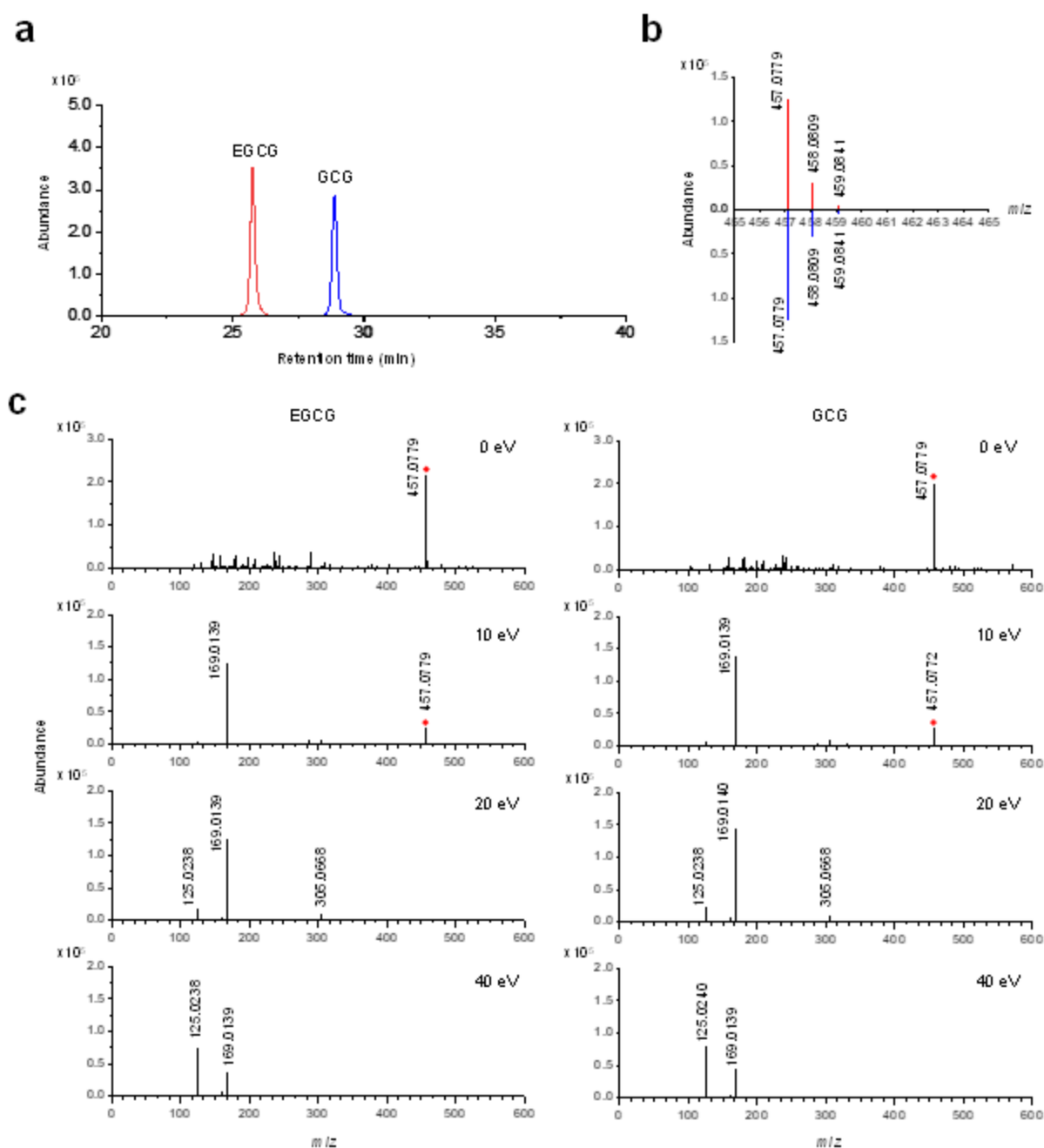


Figure 2

Figure 2

Chromatographic and spectral differentiation of epigallocatechin gallate (EGCG) and gallicocatechin gallate (GCG) as representative epimeric isomers. (a) Extracted ion chromatograms of EGCG (red, RT $\approx$ 25 min) and GCG (blue, RT $\approx$ 28 min), demonstrating baseline chromatographic separation of the two epimers sharing identical molecular formulae. (b) Overlaid isotope patterns of EGCG and GCG ($[M-H]^-$, m/z 457.0779, 458.0809, 459.0841), confirming identical elemental compositions and indistinguishable

mass spectral profiles at the precursor level. (c) Collision energy-dependent MS/MS spectra of EGCG (left) and GCG (right) at CE = 0, 10, 20, and 40 eV.

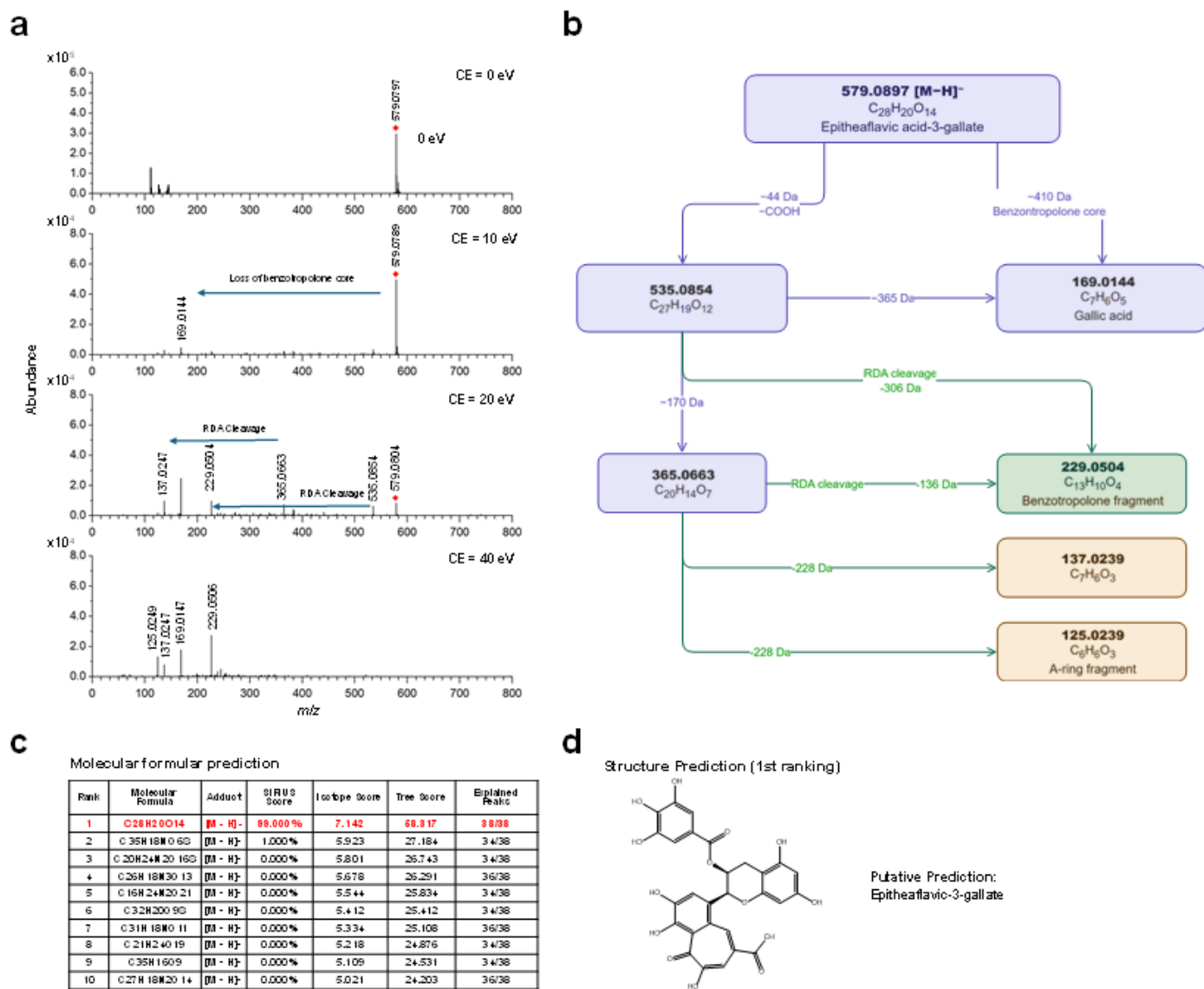


Figure 3

Figure 3

Putative identification of a theaflavin-related compound (putatively epitheaflavic acid-5-gallate, C₂₈H₂₀O₁₄, [M-H]⁻ *m/z* 579.0797) by SIRIUS-assisted molecular formula prediction and fragmentation interpretation. (a) Collision energy-dependent MS/MS spectra at CE = 0, 10, 20, and 40 eV, showing diagnostic gallic acid formation (169.0144) and characteristic low-mass fragments at *m/z* 137.0247 and 125.0249; (b) MS/MS fragmentation tree annotating key neutral losses and fragment assignments, including benzophenone-type and acylium fragment ions at *m/z* 229.0504 and 365.0663; (c) SIRIUS molecular formula prediction table showing rank 1 assignment of C₂₈H₂₀O₁₄ with a SIRIUS score of 99.000%, isotope score of 7.142, tree score of 58.317, and 38/38 explained peaks, alongside the top-

ranked predicted structure; (d) Putative predicted compounds showing rank 1 assignment of epitheaflavic-3-gallate.

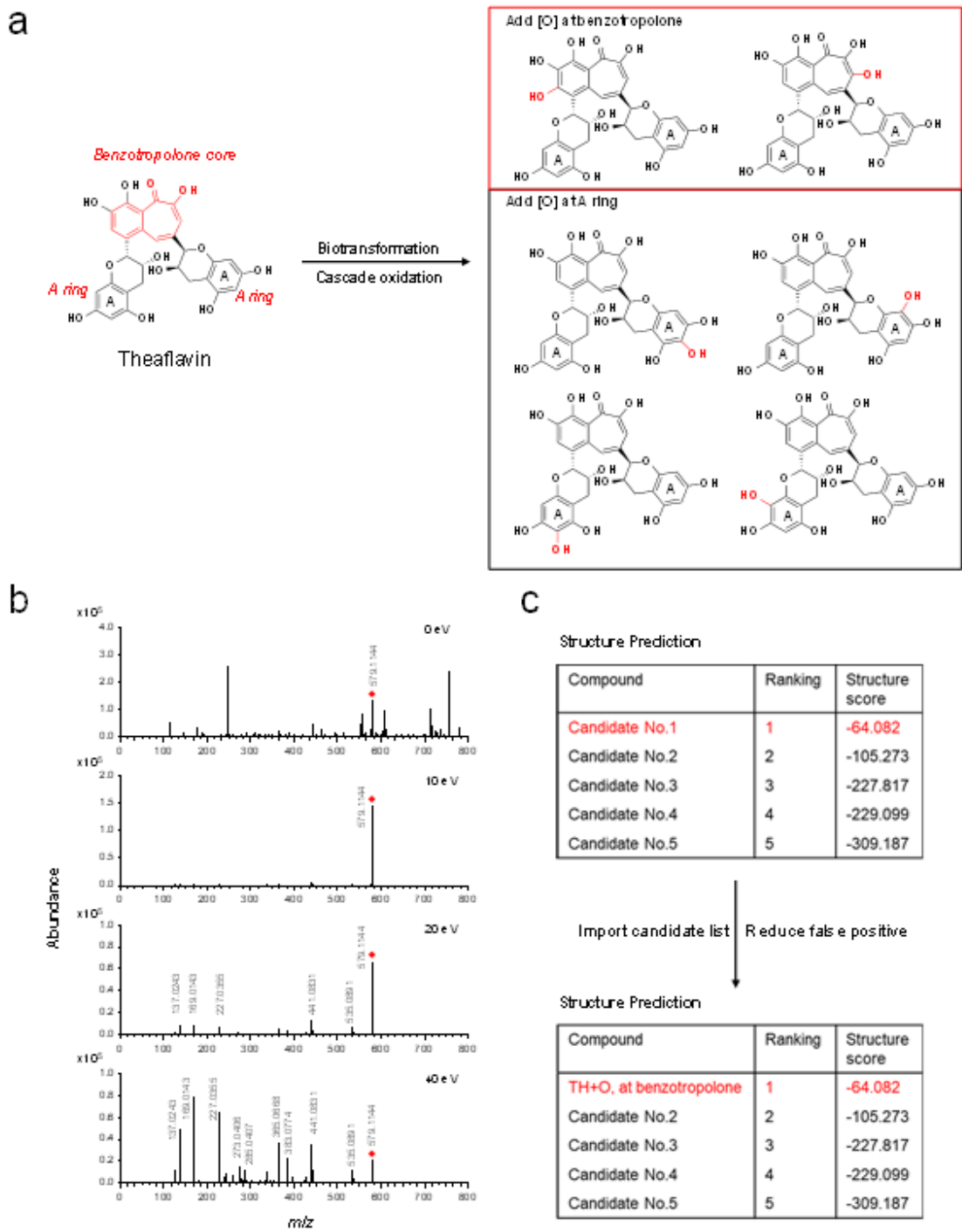


Figure 4

Figure 4 Structural prediction and MS/MS-based identification of theaflavin cascade oxidation products. (a) Proposed biotransformation pathways of theaflavin via cascade oxidation, generating structural isomers

with additional oxygen functionalities (+16 Da per step) at either the benzotropolone core or the A ring. (b) Collision energy-dependent MS/MS spectra (0, 10, 20, and 40 eV) of the representative precursor ion at m/z 579.1144, showing the progressive generation of diagnostic fragment ions including m/z 441.0831, 535.0891, 365.0668, 383.0774, 273.0406, 285.0407, 227.0355, 169.0143, and 137.0243. (c) CSI:FingerID structure prediction workflow illustrating the reduction of false positives through candidate list importation: prior to filtering, "Candidate No.1" ranked first (structure score: -64.082); following importation of the in silico-generated candidate list, the top-ranked candidate was reassigned as TH+O at benzotropolone (structure score: -64.082), providing possible oxidation site and improving annotation.

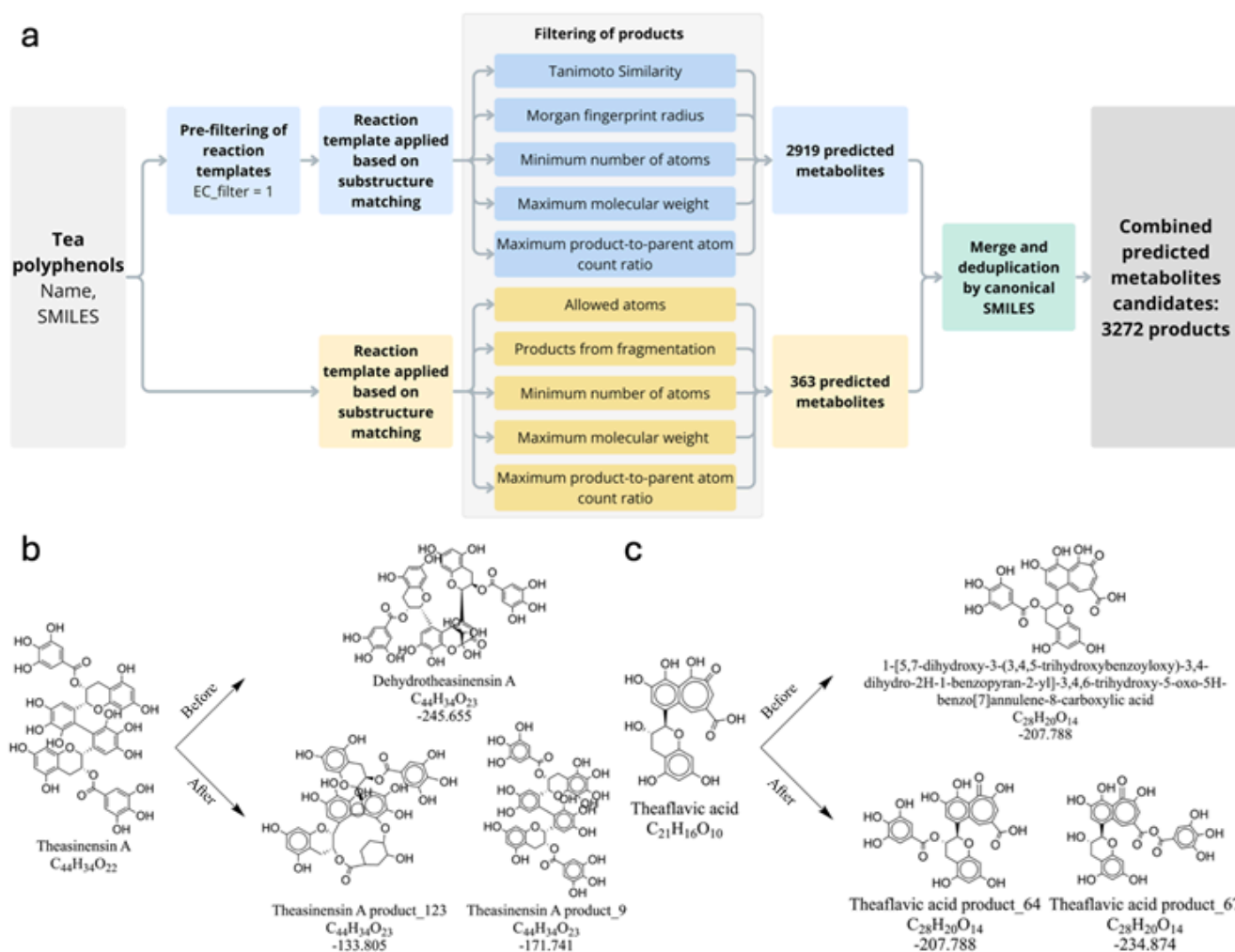


Figure 5

Figure 5

Pipeline architecture and case study applications. (a) Schematic overview of the computational predictive pipeline that comprises two branches, biotransformation (blue) and chemical transformation

(yellow), including template-based reaction application, filtering of products, and consolidation of final candidates. (b) Case study of theasinensin A using the biotransformation pipeline, together with the product names, their respective molecular formula, and their CSI:FingerID score. A higher score indicates better matching. CSI:FingerID score is a measure of the similarity between the candidate's molecular fingerprint and the fragmentation pattern. (c) Case study of theaflavic acid using the chemical transformation pipeline, together with the product names, their respective molecular formula, and their CSI:FingerID score. Note: the products on the top row are matched using SIRIUS without the candidate list and those on the bottom row are matched with the candidate list.

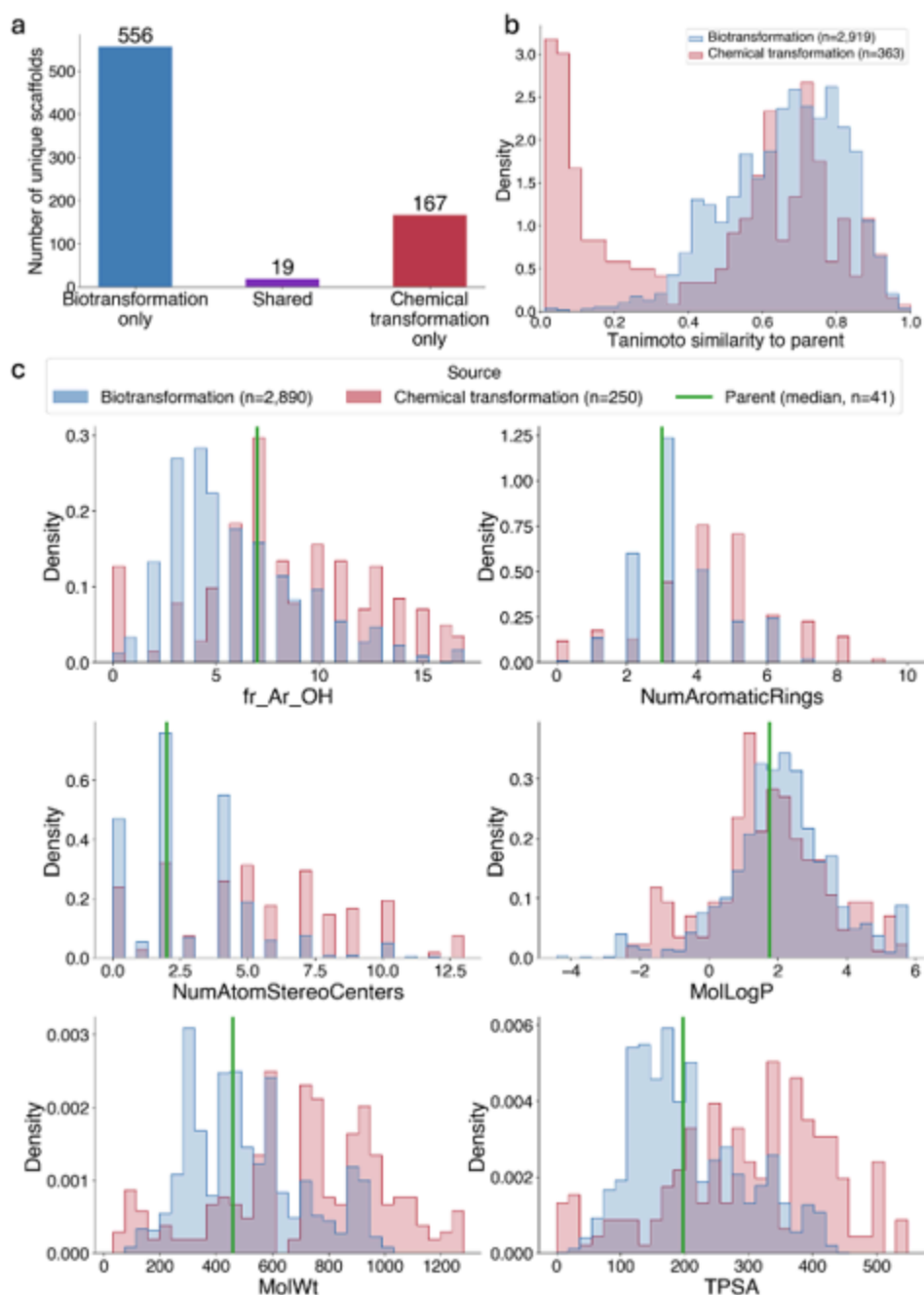


Figure 6

Figure 6

Diversity and physicochemical characterization of the predicted candidate list. (a) Murcko scaffold analysis of products generated by biotransformation and chemical transformation pipelines. (b) Distribution of Tanimoto similarity scores between predicted products (biotransformation, n=2919, and chemical transformation, n=363) and their respective parent compounds. (c) Comparison of physicochemical descriptor distributions for biotransformation products (n=2890), chemical transformation products (n=250), and the median value of parent compounds (n=41). Note: fr_Ar_OH:

number of aromatic hydroxyl groups; NumAromaticRings: Number of aromatic rings;
NumAtomStereoCenters: Number of chiral centers; MolLogP: Calculation of LogP; MolWt: Molecular weight; TPSA: Topological polar surface area

Supplementary Files

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

- [Supplementary.docx](#)