Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2026, **151**, 1071

AminoacidDB: a liquid chromatography-tandem mass spectrometry-based toolkit for the untargeted analysis of non-protein amino acids

Pawanjit K. Sandhu, [a] Ryland T. Giebelhaus, [b] Ryan Hayward, [c] Tingting Zhao,[d] Alix Tucker,[a] Daniel Gaudet, [a] Tao Huan [d] and Susan J. Murch *[a]

Non-protein amino acids (npAAs) are produced by microbes, plants and humans, with previous estimates suggesting that there are ≈1000 of such metabolites. Most of the npAAs were discovered as human toxins, intermediates in metabolism and byproducts of organic and pharmaceutical synthesis. We used a text-mining approach to identify chemicals with the NH_x-R-COOH moiety in PubChem and cross-checked those for classification against amino acid databases including Web of Science, LOTUS and HMDB to generate a dataset of compounds, which was cleaned and curated, resulting in a library of 332,154 amino acids. We established a standard set of 41 npAAs, selected to cover a wide array of structural and isomeric space for training the machine learning model and predicting chromatography elution using the Retip tool. Derivatization added a 6-aminoquinoline (6-AMQ) tag to the N[H] group, thus selecting amine-carrying compounds from the sample extract, which can be identified by cleaving the 6-AMQ carbonyl and producing the common product ion of 171.0555 *m/z* in positive ionization mode to selectively target amino acids in unknown datasets. AminoacidDB (https://www.aminoacidDB.ca) annotates amino acids by matching the features of accurate mass and retention time from untargeted mass spectrometry datasets against the aminoacidDB library. In a proof-of-concept experiment, we putatively annotated 103 amino acids and their derivatives in *Arabidopsis thaliana* and *Cannabis sativa* leaf tissues. Our original data hypothesize a wider distribution of npAAs and peptides in plants than was previously known and indicate the need for more research to understand the prevalence and metabolism of npAAs.

Received 27th November 2025,
Accepted 15th December 2025

DOI: 10.1039/d5an01248a

rsc.li/analyst

Introduction

Metabolomics is broadly defined as the use of high-resolution analytical instrumentation for the untargeted analysis of biological samples to determine their full complement of metabolites.[1,2] Such experiments generate large, complex datasets with thousands or tens of thousands of features corresponding to unidentified metabolites.[3,4] The key challenge in any metabolomics experiment is the careful and accurate identification of individual metabolites, biochemical pathways, clusters of metabolites and patterns of biological significance. The development of tailored software and data-driven approaches for processing and analyzing metabolomics data has increased the accuracy power of metabolomics analysis.[2,4–7] This is especially evident in untargeted metabolomics, where thousands of small molecules can be routinely detected in a single analysis.[1,3,8] Further development of new analytical strategies and instruments has the potential to greatly increase our understanding of metabolism.

There is an increasing interest and need for technologies that enable to understand the biochemical mechanisms and biological functions of amino acids. Amino acids are commonly defined as organic molecules with a central carbon backbone and at least one basic amino (NH_2) moiety and one acid carboxyl (COOH) moiety.[9] Amino acids are differentiated by the molecular configuration of their central carbon backbone and functional definition. The term canonical amino acids describes the 20 metabolites that commonly make up proteins.[10] Non-protein amino acids (npAAs) are characterized by the fact that they are not normally found in proteins.[11,12] The number, diversity, prevalence and metabolic importance of npAAs in nature remain unknown. Bell (2003) estimated the presence of about 900 to 1000 npAAs in plants.[12,13] Most of

[a]*Department of Chemistry, University of British Columbia, Syilx Okanagan Nation Territory, Kelowna, BC, Canada – V1V 1V7. E-mail: susan.murch@ubc.ca*
[b]*Department of Chemistry, University of Victoria, Victoria, BC, Canada – V8P 5C2*
[c]*Supra Research and Development, Kelowna, BC, Canada – V1X 6Y5*
[d]*Department of Chemistry, University of British Columbia, Vancouver, BC, Canada – V6T 1Z1*

the literature only describes a small group of npAAs found incidentally in foods,[14] meteorites[15] or in the context of human diseases.[13] For example, azetidine-2-carboxylic acid from sugar beets was discovered because it causes tissue malformations in animals.[16] Djenkolic acid, produced by the plant *Archidendron jiringa*, was discovered because it causes severe illness and acute kidney failure.[14,17] β-*N*-Oxalyl-L-α,β-diaminopropionic acid (ODAP), found in *Lathyrus* seeds and plants, was identified to be associated with the neurological disorder lathyrism.[18] β-Methylamino-alanine (BMAA) was isolated from cycad seeds in 1967, which is a neurotoxin associated with amyotrophic lateral sclerosis/parkinsonism-dementia complex.[19] One intriguing possibility under investigation is the potential of natural npAAs to be mis-incorporated into proteins due to errors in protein synthesis or *via* currently unknown RNA mechanisms.[20–24] For example, levodopa has been shown to be mis-incorporated into proteins in place of tyrosine in patients with Parkinson's disease.[24] BMAA has been shown to be mis-incorporated into proteins in *in vitro* synthesis systems[22] and cell cultures.[23]

New analytical tools are required to fully understand the metabolism, biochemistry and health impacts of npAAs. Accordingly, the objectives of our work were to (1) develop a comprehensive database of npAAs and (2) develop and validate a database toolkit for the analysis of npAAs in all types of samples including plants, microbes, animals, humans and ecosystems. Our toolkit combines methods for the analysis of large datasets and databases[4] with predictive algorithms[1,7,25,26] and high-resolution mass spectrometry[27,28] in an accessible online, open-source format[7] for easy use. Our technology combined with emerging omics technologies and advances in mass spectrometry will enable studies to fully understand the amino acid complexity in proteins, cells and organisms.

Methods

Chemicals

Acetonitrile (CAS No. 75-05-8, Optima® LC/MS, Fisher Chemical, Ottawa ON), water (CAS No. 7732-18-5, Optima® LC/MS, Fisher Chemical, Ottawa ON; 18.2 MΩ cm, Direct Q3, Millipore, Mississauga, ON), methanol (CAS No. 67-56-1, Optima® LC/MS, Fisher Chemical, Ottawa ON), formic acid (CAS No. 64-18-6, Optima® LC/MS grade, Fisher Chemical, Mississauga, ON), 6 N hydrochloric acid (HCl, CAS No. 7732-18-5, 7647-01-0, Fisher Chemical, Mississauga, ON), 0.1 N trichloroacetic acid (TCA), which was made by dissolving TCA reagent (>99.0%, CAS No. 76-03-9; Sigma-Aldrich) in ultrapure water), 0.2 M borate buffer and 6-aminoquinolyl-*N*-hydroxysuccinimidyl carbamate (AQC) (AccQ·Tag™ Ultra Derivatization Kit, Part No. 186003836, Waters, Mississauga, ON; kit reconstituted according to manufacturer's instructions).

Standards

41 authentic amino acid standards were used for method development (Table S1). They include a standard mix of 17 protein amino acids called amino acid standard H (Thermo Fisher Scientific™), structural isomers of protein amino acids such as beta-alanine, DL-norvaline, and DL-norleucine, and other npAAs including L-β-*N*-methylamino-L-alanine (BMAA) and its 3 structural isomers, 1-aminocyclopropane-1-carboxylic acid (ACC), and DL-β-3,4-dihydroxyphenylalanine (DOPA), among others (Table S1). The amino acids were selected to cover a wide array of structural and isomeric space. The stock solutions of most amino acids were created by dissolving their powder in 0.1 M aqueous HCl (details in Table S1), while further dilutions were performed with 20 mM aqueous HCl (Table S2). The stock solutions were stored at −20 °C.

UHPLC-MS/MS method development

A method for the untargeted analysis of amino acids by LC-MS/MS was developed for amino acids derivatized with AQC. Derivatization makes the polar zwitterionic amino acids amenable to reverse phase separation and increases the reproducibility of the method.[29,30] The method was developed by the modification of targeted analysis methods.[29–32]

AQC derivatization

Employing stock solutions of the 41 amino acid standards, a mixture of amino acid standards was created based on the preliminary analysis of MS detector response (see Table S1 for detailed list of authentic standards). 10 μL of the amino acid mixture was diluted with 70 μL of borate buffer in an autosampler vial (2 mL amber glass with pre-slit Teflon-coated caps; Waters Corp.) fitted with a conical bottom spring insert (250 μL glass; Canadian Life Science, Peterborough, ON, Canada) and derivatized with 20 μL of AccQ·Tag™, followed by vortex mixing (Vortex Genie 2; Scientific Industries, Bohemia, USA) and incubation at 55 °C for 10 min to complete the reaction.

UHPLC parameters

The analysis was performed on a Vanquish UHPLC (Thermo Scientific) system fitted with a Vanquish autosampler, coupled to a Q Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo Scientific) with a heated electrospray ionization (HESI) probe (Waltham, MA, USA) for amino acid detection. 10 μL of the derivatized amino acids was injected into a reverse-phase CORTECS® UPLC® C_{18} column (2.1 × 150 mm, 1.6 μm; Part No. 186007096, Waters, Mississauga, ON) fitted with a CORTECS® C_{18} VanGuard Pre-column (90 Å, 1.6 μm, 2.1 mm × 5 mm; Part No. 186007123, Waters, Mississauga, ON) heated to 55 °C. The amino acids were eluted with a gradient-elution of water/formic acid (99.9 : 0.1; v/v) (solvent A) and acetonitrile/formic acid (99.9 : 0.1; v/v) (solvent B) at a flow rate of 0.4 mL min^{-1}. The needle and seal wash solvent consisted of water : methanol (80 : 20; v/v). The wash time was 10 s at a speed of 25 μL s^{-1}. The sample manager temperature was set to 4 °C. The method was optimized for the separation of structural isomers along the retention time (RT) axis, while the

amino acids with similar masses were resolved by high-resolution mass spectrometry (HRMS) in the Orbitrap. The optimized method had a run time of 30 min with the gradient curve, as given in Table S3.

Mass spectrometer parameters

MS data was collected between 0.5–29 min of chromatographic run time using a full MS/data-dependent MS^2 (dd-MS^2) experiment in a Q Exactive hybrid quadrupole-Orbitrap mass spectrometer (Thermo Scientific) coupled to HESI. The ions were detected in positive mode with electrospray ionization (ESI) and default charge state of 1. The full MS data was collected in profile mode for a scan range of 75 to 800 m/z at a resolution of 70,000 at m/z 200. The automated gain control (AGC) target was set to 1×10^6 or 50 ms as the maximum injection time. The top 5 precursor ions from each scan were fragmented in dd-MS^2 with a stepped collision energy of 17.5, 30 and 47.5 eV. The isolation window was set to 2.0 m/z to select the precursor ions. The parameters for MS^2 were as follows, mass resolution: 17,500 at m/z 200, scan range: 120–2000 m/z, AGC: 1e5, Maximum IT: 50 ms, and dynamic exclusion: 2 s.

Method performance

The sensitivity, selectivity, limit of detection, limit of quantification, accuracy and precision of the method were assessed by repeated intra- and inter-day analysis of the authentic amino acid standards (Table S4). A representative standard mix was run four times per day along with the entire dilution series over three days (Table S2) to calculate the residual standard deviation in MS response and shift in RT (Table S4).

Data processing

Raw files from Q Exactive Orbitrap were processed using Compound Discoverer v3.3.2.31 (Thermo Scientific) for RT alignment, compound detection, predicting elemental compositions and compound identification. For compound detection, the mass tolerance was set to 5 ppm, min peak intensity: 10000, and min #scans per peak: 5, S/N threshold: 1.5. The detected compounds were scored for the presence of 171.0555 m/z in MS^2 as the product ion is produced upon cleavage of the 6-aminoquinoline (6-AMQ) tag from the AQC-derivatized amino acids in MS/MS. The m/z and RT information from the amino acid standards was used as input to train the RT prediction models for the development of aminoacidDB.

Developing web-tool for amino acid identification, aminoacidDB

The code for the aminoacidDB web-tool was written using R, RStudio and RShiny. This web-tool is publicly available *via* aminoaciddb.ca. The code performs amino acid matching between user-uploaded data and aminoacidDB datasets based on accurate mass and RT. Putative compound identity is assigned through m/z matching based on the user-defined mass tolerance. Once putative amino acid(s) are identified for a given m/z, then the predicted RTs are compared to the experimental RT. This allows analysts to eliminate isobar amino acids with predicted RTs that are considerably different from the experimental RT. The predicted RTs are considered after m/z as RT is more variable across instruments. The datasets for aminoacidDB were curated using the following methods. "Amino acids" were defined as chemical compounds containing at least one basic amine (primary or secondary) functional group and one carboxyl functional group and include the known and unknown biological functions of both protein and non-protein classes. The "*amino acid*" categories in databases such as LOTUS and HMDB (Human Metabolome Database) contain compounds where either the amine or carboxylic acid functional group is modified, leaving no free amine or carboxylic acid, respectively, and these compounds were designated as amino acid derivatives in aminoacidDB.

LOTUS

LOTUS is a natural product database with >250 000 metabolites from biological organisms including plants.[33] LOTUS includes a biological and chemical classification of metabolites based on biological origin and chemical structure, respectively. Based on the chemical classification, we downloaded the small peptides class in LOTUS from PubChem containing amino acids, dipeptides, tripeptides and non-classified categories [June 2025]. The dataset was filtered to remove compounds not containing C, N or less than 2 O atoms as well as duplicated chemical structures (represented by duplicate SMILES/InChIKey). The dataset was manually investigated to avoid removing any unique amino acid structures. The dataset contains name, monoisotopic mass, molecular formula, chemical structure (SMILES and InChIKey) and LOTUS classification for each of the compounds.

HMDB

HMDB is a database of >200 000 small molecule metabolites found in the human body.[34] The compounds in HMDB are classified into various classes using ClassyFire[35] including a class called amino acids, peptides and analogues. We downloaded the classified HMDB dataset from the ClassyFire website (released 2016-08-31) and selected metabolites in classes containing the term "*amino acid*". The HMDB IDs from the selected metabolites were searched against the latest HMDB data (released 2021-11-17) to retrieve metabolite information including name, monoisotopic mass, molecular formula, chemical structure (SMILES and InChIKey) and ClassyFire classification. Similar to LOTUS, the dataset was filtered to remove duplicate amino acid entries and combined with the LOTUS dataset.

Web of science

The above-mentioned LOTUS and HMDB datasets were supplemented with known npAAs in plants from Web of Science™. Text-mining was performed using Web of Science™ with terms ("non protein* amino acid*" OR "nonprotein* amino acid*" OR "noncanonical amino acid*" OR "non cano-

nical amino acid*" OR "noncoded amino acid*" OR "non-protein* amino acid" OR "non-canonical amino acid" (all fields) + plant* (all fields)) [May 2025]. The publications from the search were manually curated to extract known npAAs. The SMILES, InChIKey, monoisotopic mass and molecular formula of npAAs were retrieved from PubChem. The npAAs were combined with the above-mentioned LOTUS and HMDB datasets.

PubChem

The PubChem database was downloaded in SDF format on January 16, 2021 from https://pubchem.ncbi.nlm.nih.gov/source. 109,050,179 chemicals were extracted and saved in CSV files. All data processing was achieved in R (v4.2.1). Prior to any further processing, deuterated chemicals and chemicals with nitrogen in the form of an ammonium cation were excluded from the PubChem database. Afterwards, several criteria were applied to find the chemicals containing both carboxylic and amine (primary or secondary) groups. The presence of the carboxylic group was determined by checking the presence of '(C(=O)O)' and 'C(=O)O' in SMILES. The presence of non-charged primary and secondary amine groups was determined using fingerprint. The SMILES was first converted to fingerprint using the get.fingerprint() function with the 'maccs' method in the 'fingerprint' R package. The chemicals with bit of 84 in their fingerprint are primary amine compounds and directly retained. The chemicals with bit of 151 in their fingerprint are possible secondary amines(–NH) and subject to further confirmation. If their –NH substructures are all in the form of a peptide (amide) bond, these possible secondary amines will be excluded. For the selected compounds, their metadata, including name, molecular formula, PubChemID, SMILES, InChIKey and type of amine groups, were output in CSV files. The above-mentioned extracted compounds were filtered to remove heavy isotopes and chemical mixtures, from which unique SMILES representing unique compounds were selected. The dataset was limited to uncharged chemicals containing CHNOS with the number of carbons ranging from 2 to 11 to limit the database similarity to protein amino acids.

For each of the compounds in the above-mentioned datasets, their functional information including number of primary and secondary amine groups, and number of carboxylic acid groups was calculated using RDKit v2025.03.5[36] in Python v3.13.5. The figures were plotted using the ggplot2[37] and ggbreak[38] packages in R and RStudio, respectively. Further, seven common MS adducts including $[M + H]^+$, $[M - H]^-$, $[M + NH_4]^+$, $[M + Na]^+$, $[M + K]^+$, $[2M + H]^+$ and $[2M + NH_4]^+$ were calculated for each analyte in all the datasets.

Retention time prediction using Retip 2.0

The RT for selected amino acids in aminoacidDB (LOTUS, HMDB, WoS, and PubChem) was predicted based on a model built using Retip 2.0[39] in R (v4.5.0). 198 known analytes covering a wide array of metabolite space including amino acid standards (non-derivatized and AQC derivatized), phenolics, hormones, cannabinoids, and terpenes (Table S5) were used to build a robust RT prediction model. The standards were analyzed using the above-mentioned optimized UHPLC-MS/MS method and their RT and *m*/*z* were extracted from Compound Discoverer. The structural information (SMILES and InChIKey) along with RT was used as input in Retip to build an RT prediction model. Based on the structural information, Retip 2.0 calculates chemical descriptors using cdk for each compound, all of which are used as predictors in model building. The data from 198 analytes was split randomly, where 80% was used for training the model and 20% was left out for external validation of the model performance. Within the training dataset, 10-times *k*-fold cross validation was also performed.

Retip 2.0 has five built-in algorithms for RT prediction including random forest (RF), bidirectional recurrent neural networks (BRNNs), XGBoost, light gradient-boosting machine (lightGBM), and H_2O auto machine learning (autoML). Multiple models were built within each of the algorithms using multiple iterations of parameters (*e.g.*, 10 values of tuning parameters for random forest, 5 unique neural network architectures for BRNN, and 30 for autoML), the best of which was selected using 10-times *k*-fold cross validation.[39] The best model among them was selected based on the R^2, root mean square estimate (RMSE) and mean absolute error (MAE) of the external validation dataset and was used to predict RT for the amino acids in aminoacidDB.

Method testing and validation

Test plant materials

Arabidopsis thaliana L. Heynh (Col. 0). *A. thaliana* Col 0 plants were propagated from seeds and grown *via in vitro* tissue culture under sterile conditions. The sterile controlled growth environment room was set at 25 °C, under a 16 : 8 photoperiod at a light intensity of 25–45 μmol m^{-2} s^{-1} at the bench level, and at a positive pressure (+2.9–3.7 Pa) under HEPA-filtered air. Briefly, the seeds were surface sterilized with a sterilization solution (70% Clorox and 0.1% Tween in sterile water) for 5 min, followed by multiple washes with sterile water. The sterilized seeds were planted in magenta boxes containing Murashige & Skoog Modified Basal Medium with Gamborg B5 Vitamins (PhytoTechnology Laboratories, Lenexa, Kansas) solidified with 0.3% Phytagel™ (Sigma-Aldrich) and 1% sucrose at pH 5.7. The medium was sterilized by autoclaving at 121 °C and 15 psi for 20 min (Steris, Mississauga, ON). One magenta box served as one replication unit. Shoots from 4 replicates were harvested 3 weeks after sowing.

***Cannabis sativa* L. CV 'Black Cherry Punch #2'.** Samples of *C. sativa* 'Black Cherry Punch #2' leaves were provided by a commercial research facility (Hawthorne/Flowr R&D Facility, Kelowna, BC) to our licensed research cannabis facility (Health Canada License LIC-0IW2D2L5JY-2024-1) under a Material Transfer Agreement for research. The cannabis plants were grown in a controlled environment growth room under full-spectrum LED light (Gavita 1700e, Hawthorne) with a 14-day propagation stage, 22-day vegetative growth phase and 49-day

flowering cycle. The plants were grown in rock wool cubes (6″ cubes, Grodan) with optimized fertigation (FloraPro Series 'Expert'; General Hydroponics). Samples were provided from the dry trim cut 7 days after transplanting propagation plugs into the rockwool cubes, transported to UBC and stored at −80 °C until analysis. Four biological replicates were used for further analysis.

Amino acid extraction and sample preparation

Free soluble amino acids were extracted from the leaf samples using 0.1 N TCA to avoid highly abundant protein amino acids in the protein pellet. Briefly, leaf tissue was ground with liquid N_2 and weighed (~200 mg) into a 1.5 mL microcentrifuge tube (Fisher Scientific, USA). The ground leaf tissue was homogenized with 1000 μL of 0.1 N TCA by vortexing at full speed for 30 s. The homogenized samples were centrifuged (5 min at 13 000 rpm) to pellet the proteins. An aliquot of 800 μL of the supernatant was filtered through an Ultrafree®–MC GV centrifugal filter tube (Durapore® 0.22 μm PVDF membrane; Merck Millipore) by centrifugation (5 min at 13 000 rpm) to yield the filtered free amino acid extract. The filtered amino acid extract was stored at −20 °C until analysis. 10 μL of the extract was derivatized with 20 μL of AccQ·Tag™ similarly to the standards and used for untargeted analyses in UHPLC-MS/MS. A derivatization blank was prepared, where 10 μL of extraction buffer (0.1 N TCA) was added instead of the sample extract. An extraction blank was prepared following the extraction protocol without the plant material. 10 μL of the derivatized mixture was injected onto the instrument for analysis using the above-mentioned optimized LC and Orbitrap parameters. A sample of *A. thaliana* was prepared separately and used as a quality control sample for the analysis. Additionally, solvent blanks were run at the start, middle and end of the runs to assess carryover.

Data processing and annotation of non-protein amino acids

The raw files from Orbitrap were processed using Compound Discoverer (v3.3.2.31), as described above. For the predicted compositions, the minimum element count was described as C10H7N2O similar to the elemental formula of 6-AMQ tag added by AQC derivatization to limit the predicted compositions. Features with a peak area greater than 5e4 were selected and manually curated to remove artifacts, adducts and tag peaks. The *m*/*z* and RT information from the curated peaks was used as input for aminoacidDB for putative compound annotation. The features where accurate mass (<5 ppm mass error), MS/MS, RT and predicted composition can be matched with authentic analytical standards were annotated at Metabolomics Standard Initiative (MSI) level 1 confidence.[40] The features that matched with <5 ppm mass error, percent RT match >75% and predicted composition were putatively annotated at MSI level 2.[40] The features with <5 ppm mass error and percent RT match <75% or that matching multiple isomers with <5 ppm mass error and percent RT match >75% were putatively annotated at MSI level 3.[40] Finally, the features where only elemental composition could be predicted with confidence were putatively annotated at MSI level 4.[40] A t-test followed by Benjamini-Hochberg correction was applied for the comparison of npAAs detected in both species.

Meta-analysis of previous metabolomics studies

To validate the use of the aminoacidDB web interface for versatile metabolomics datasets, we performed npAA annotation on untargeted metabolomics datasets from the Metabolomics Workbench (https://www.metabolomicsworkbench.org/). We decided to focus on studies conducted on our selected study species, *A. thaliana* and *C. sativa*. As of June 2025, there were 24 studies on *A. thaliana* and 0 studies on *C. sativa* in the Metabolomics Workbench. The list was filtered to remove studies using nuclear magnetic resonance (NMR), gas chromatography-mass spectrometry (GC-MS), hydrophilic interaction chromatography (HILIC), or reversed phase LC techniques with low-resolution detection methods. We also excluded studies with only unaligned raw data. In the case of studies analyzing multiple species, features not detected in *A. thaliana* were removed. This resulted in 3 studies with 6 datasets including 3 in ESI+ mode and 3 in ESI− mode (Table S7). The features that were common between ionization modes were counted as one. The features within 0.01 mass tolerance were considered common across the studies. The *m*/*z* and RT were searched against aminoacidDB data using a mass tolerance of 10 ppm and RT threshold of 50% given that none of the selected studies had the exact chromatography conditions (column, mobile phases and gradient) as the LC method used to predict the elution profiles in aminoacidDB.

Results and discussion

Standardized operating protocol for npAA analysis and identification

The standardized operating protocol (SOP) including sample preparation, UHPLC-MS/MS method and data processing was optimized for the analysis of npAAs in plants (Fig. 1). The amino acids were targeted during sample preparation by the derivatization of their amine group with AQC. AQC derivatization adds a 6-AMQ tag to the N[H] group, thus selecting amine-carrying compounds from the sample extract.[29,30] Further, in MS/MS, the 6-AMQ carbonyl cleaves from the derivatized structure, producing a common product ion of 171.0555 *m*/*z* in positive ionization, which is used to selectively target amino acids post-acquisition.[29,30] Among the 41 amino acid standards, DL-4-chlorophenylalanine methyl ester and kynurenic acid failed to derivatize with AQC. This might be due to the presence of delocalized electrons in the quinoline ring of kynurenic acid, resulting in a weakly nucleophilic N. Only one dissociation event was observed for kynurenic acid at pK_a of 2.43,[41] with no dissociation event at basic pH, indicating that the quinoline N does not dissociate, contrary to its structural inference.[41] These findings suggest that weakly nucleophilic amines (with an aromatic ring system or bonding to an electron-withdrawing group) might not derivatize with

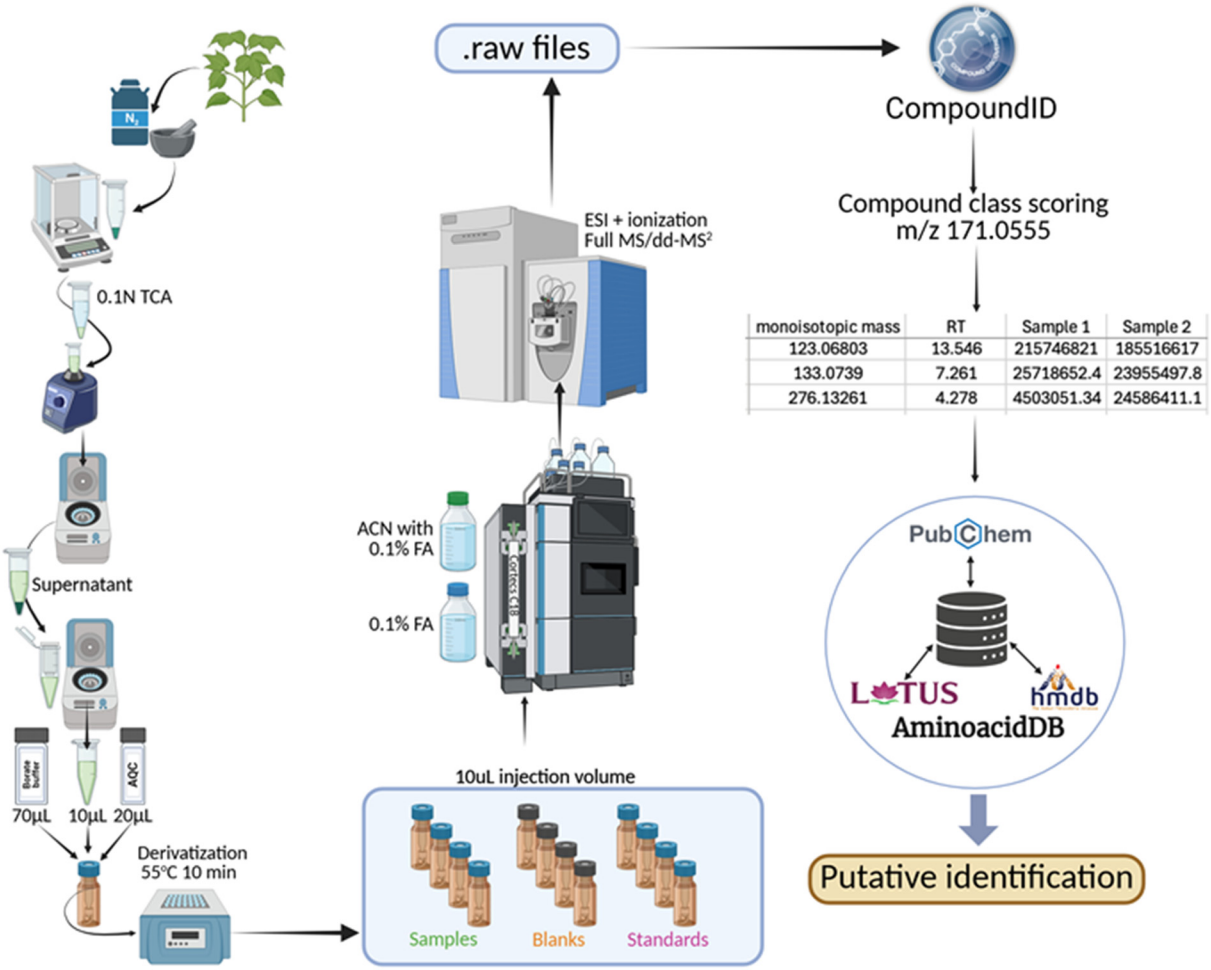

Fig. 1 Standard operating protocol (SOP) for the analysis, discovery and identification of amino acids using aminoacidDB.

AQC, and hence are not targeted by the current method. The LC-MS/MS method parameters were optimized for the remaining 39 amino acid standards. The finalized method has a runtime of 30 min (Table S3) and uses solvents and instruments employed for routine amino acid analyses[29,30] for wide-range applicability across biology and analytical chemistry laboratories. HRMS helped separate close masses (<5 ppm accuracy), which is instrumental in the annotation of npAAs in mass spectrometry datasets with accurate mass matching and prediction of elemental compositions. Structurally similar isomers cannot be differentiated in MS due to their same mass, and thus were resolved along the RT axis and differential MS/MS patterns (Fig. S1). Recent advancements in ion mobility (IM) techniques such as trapped ion mobility spectrometry (TIMS) offer the orthogonal separation of isomers based on their gas-phase mobility (shape, size, and charge) *via* their collision with a buffer gas (nitrogen or helium) under the influence of an electric field.[42] A unique advantage of IM experiments is that the collision cross section (CCS, Ω) value of an ion could be used to characterize isomers, especially in cases where chromatographic separation and MS/MS are insufficient.[42,43] Comprehensive two-dimensional separation (LC × LC and GC × GC) is another technique for resolving structural isomers such as amino acids by subjecting analytes to two stationary phase chemistries in one injection before MS or MS/MS.[44] Unlike IM, separation occurs before ionization, mitigating ion suppression by physically separating the coeluting matrix components before ionization.[45] Structured chromatograms also arise in comprehensive separations, revealing characteristic patterns of analytes sharing physiochemical properties.[46] The developed SOP expands amino acid analyses by including npAAs missing in traditional amino acid analyses, along with a derivatization-based selection strategy, allowing the discovery of novel amino acids using tools readily available in laboratories.

AminoacidDB web-tool

AminoacidDB is an open-source web-tool (available at aminoaciddb.ca) for the identification of npAAs post-acquisition. It allows users to annotate amino acids including npAAs in untargeted metabolomics datasets from any source (plants, animals, microbes, and humans) by matching accurate mass

and RT against the aminoacidDB datasets. Currently, aminoacidDB contains 332,154 amino acids and derivatives from LOTUS, HMDB, and PubChem. Additionally, there are 4 different adduct classes users can search against: monoisotopic (the monoisotopic mass of amino acids from LOTUS/HMDB or PubChem datasets), $[M + H]^+$ ($[M + H]^+$ adduct of amino acids in ESI+ mode), $[M–H]^-$ ($[M–H]^-$ adduct of amino acids in ESI− mode), and adducts ($[M + Na]^+$, $[M + K]^+$, $[M + NH_4]^+$, $[2M + H]^+$, and $[2M + NH_4]^+$ adducts of amino acids in ESI+ mode).

Data input and output. User data are uploaded in CSV format with *m*/*z* in the first column, RT in the second column and sample concentration (area or intensity) in the remaining columns. The CSV format provides versatility to aminoacidDB given that metabolomics datasets could be collected using any available HRMS instrument (Orbitrap, time-of-flight (TOF), FT-ICR *etc*.) as well as processed with a variety of software (such as XCMS, MS-DIAL, MZmine and vendor software). Prior to uploading data, the user selects the datasets (LOTUS/HMDB and/or PubChem) and adducts to search against: monoisotopic, $[M + H]^+$, $[M - H]^-$ or adducts, along with the mass tolerance (±Da or ppm). The mass error to be selected depends on the type of instrument and methods used for data collection. For example, 0–5 ppm mass error is generally recommended for Orbitrap instruments and 10–15 ppm for TOF instruments. The output of aminoacidDB can be viewed in the "Screener Output" tab on the app or downloaded as a CSV file by clicking the "Download Results" on the "*m*/*z* Screener" tab. Searching is performed first by *m*/*z*, followed by RT, as predicted RTs are more variable across systems. However, no holistic scoring metric is used considering both *m*/*z* and RT. An example dataset to familiarize users with the formatting and use of aminoacidDB is provided on the website.

AminoacidDB dataset curation. There are 332,154 amino acids in the current version of aminoacidDB curated from LOTUS, HMDB or PubChem resources. The "All Amino Acids" tab presents all the relevant information for amino acids including name, structure, formula, monoisotopic mass, source of the amino acid, class (amino acid or amino acid derivative), LOTUS classification, and HMDB ClassyFire classification. The datasets are also available to download in CSV format using the "Download All Amino Acids" button in the "Instructions" tab.

LOTUS/HMDB dataset. Amino acids from the LOTUS and HMDB datasets could be considered as classified amino acids. Using Lotus, there were 3194 analytes in the small peptide category including amino acids (1509), dipeptides (1103), tripeptides (503) and not classified (79). The classification in LOTUS is non-exclusive, resulting in the same compound classified into multiple categories. Only one unique entry per compound was kept, along with the removal of duplicate structures (represented by SMILES/InChIKey), resulting in a dataset of 2460 compounds from LOTUS.

The ClassyFire dataset of HMDB contained 4686 unique HMDB IDs from the "*amino acid*" classes. Searching HMDB did not retrieve any results for 159 of these HMDB IDs. Further, multiple IDs matched one compound, only one of which was kept, along with the removal of duplicate and non-amino acid structures, producing a dataset of 4467 compounds from HMDB. 429 analytes were common between the HMDB and LOTUS datasets. The dataset was supplemented with 27 known npAAs from Web of Science™ (WoS) that were not found in either the HMDB or LOTUS datasets. This resulted in a final dataset of 6525 putative npAAs and derivatives, highlighting the extent of known amino acids that are missing in traditional metabolomics and amino acid analysis and can be targeted using the aminoacidDB protocol.

More than 42% of the analytes from LOTUS, HMDB and WoS had either their amine or carboxylic acid functional moiety modified, leaving no free amine/carboxylic acid group, and were classified as amino acid derivatives; the rest were classified as amino acids (Fig. 2A). Approximately 2/3rd of the compounds contained a primary amine group, while the rest contained a secondary or tertiary amine or other N functional groups (Fig. 2A). In the amino acid class, >75% of primary amino acids had one NH_2 moiety, while <5% had 3 or more NH_2 groups. 46% of the secondary amino acids had one N[H], 23% had two, and 5% had more than 4 N[H] groups (Fig. 2A). Further, >90% of the amino acids had 2 or less COOH functional groups (Fig. 2A). Given that the amino acid derivatives had their carboxylic or amino group modified, >90% of the compounds in this class did not have a free COOH group (Fig. 2A), while ~90% of the compounds had at least one NH_2 or N[H] moiety (Fig. 2A). The wide variation in molecules based on their functional groups and position and length and structure of their carbon chain highlights the structural diversity of amino acids known from biological sources (plants, microbes or humans) beyond the 20 protein amino acids. Profiling this diversity of npAAs in biological samples would provide an opportunity to better understand their role in protein chemistry and metabolism.

PubChem dataset. The PubChem search for amino acids (>1 C(=O)O + >1 (N[H])) resulted in 2,631,019 compounds, which were filtered to remove artifacts, heavy isotopes and stereoisomers. The remaining 2,184,228 unique structures had a monoisotopic mass following a normal distribution, with 90% of compounds falling between 209.1052 Da and 665.2957 Da (Fig. S2). This is interesting considering that the heaviest protein amino acid, tryptophan, has a mass of 204.2 Da and majority of the PubChem matches are larger. These data may reflect the composition of the PubChem literature, which includes both naturally occurring and synthetic npAAs as well as short peptides and amino acid conjugates. To better reflect the biological relevance in the dataset, we restricted the included compounds to those with a molecular formula containing C, H, O, N and S and between 2–11 carbons. This resulted in a dataset of 325,843 compounds (Fig. 2B). Within this subset of data, we identified 8358 unique molecular masses, 2800 of which were non-isomeric, while 5558 accounted for most of the observed structural diversity. The full dataset can be accessed through the online tool and has been made available through Borealis (https://borealisdata.ca/

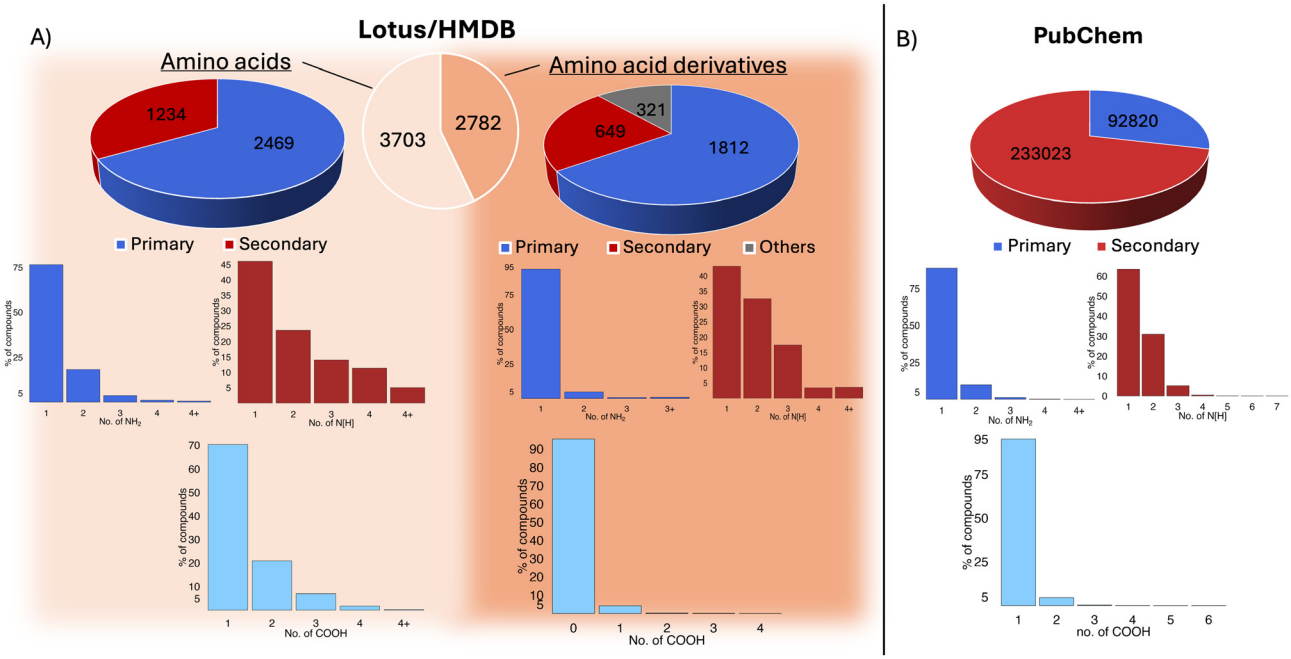

Fig. 2 Functional analyses of amino acids in the aminoacidDB dataset: (A) LOTUS/HMDB and (B) PubChem. Blue, red and grey colors in the pie charts indicate the proportion of compounds containing primary amine, secondary amine or other N-functional groups, respectively. The bar graphs show the number of NH_2 (dark blue), N[H] (red) or COOH (light blue) moieties per compound on the *X*-axis plotted against percentage of compounds on the *Y*-axis.

dataset.xhtml?persistentId=doi:10.5683/SP3/DTAZ1G). The average number of isomers per compound was 58.2. There were >2200 structural isomers for 20 protein amino acids, with as many as 525 isomers for tryptophan, highlighting the structural diversity of the dataset (Fig. S3). Most of the structural diversity is in the length of the carbon chain and position of the functional groups given that >50% of the amino acids contained one amino and one carboxylic group (Fig. 2B). Only 28.5% of the compounds were primary amino acids (*i.e.* containing NH_2), while the rest were secondary amino acids (Fig. 2B). Within the primary amino acids, 88% had one amino group, 10% had two, and <2% had 3 or more amino groups. Among the secondary amino acids, 63% of the compounds had one N[H] group, 30% had two, and <1% of the compounds had 3 or more amino groups. >95% of the amino acids contained only 1 COOH group (Fig. 2B).

RT prediction of aminoacidDB datasets. Retip 2.0, an R package for predicting the retention times of small molecules in HPLC,[39] was used to build models for predicting the RT of amino acids in the aminoacidDB datasets. Among the five algorithms tested, autoML performed the best with R^2 of 0.99, mean absolute error (MAE) of 0.89 min and 95% confidence interval for the predicted retention time of ±1.8 min for the external validation dataset (Table 1 and Fig. 3). The model performed well for training data with R^2 of 0.96 and standard error of 1.9 ± 3.1 min (Table 1 and Fig. 3). Retip 2.0 could not predict the RT for compounds containing unusual elements (Si, Gd, and As). The final npAA datasets with predicted RT are LOTUS/HMDB (6485 compounds) and PubChem (325,669 compounds). We advise users to use our method conditions to appropriately use the RT matching criterion in aminoacidDB, which annotate npAAs at MSI level 2,[40] along with running a set of authentic standards covering the space of gradient elution parameters to account for unintentional differences across systems. If the authentic standards are not run with a similar LC method (column type and solvent system), an RT matching algorithm could be employed to reject candidates, thereby reducing the target space. With the use of different LC conditions, the RT matching criterion is not advised as the RT tends to vary. The use of aminoacidDB for hypothesis generation should be guided by an assessment of biological plausibility, statistical significance analysis, metabolite cluster analysis and pathway mapping to ensure the valid interpretation of the results.

Table 1 Model performance characteristics for testing (external validation) and training datasets for auto machine learning (autoML) Retip model for retention time prediction. Retention time units are in minutes. RMSE: root mean square estimate and MAE: mean absolute error (in minutes)

autoML	RMSE	R^2	MAE (minutes)	95% ± min
Training	1.92	0.96	0.98	3.06
Testing	1.22	0.99	0.89	1.79

AminoacidDB custom search tool. AminoacidDB has a special feature where the code was modified to create the "Custom Database Search", which allows users to use the exist-

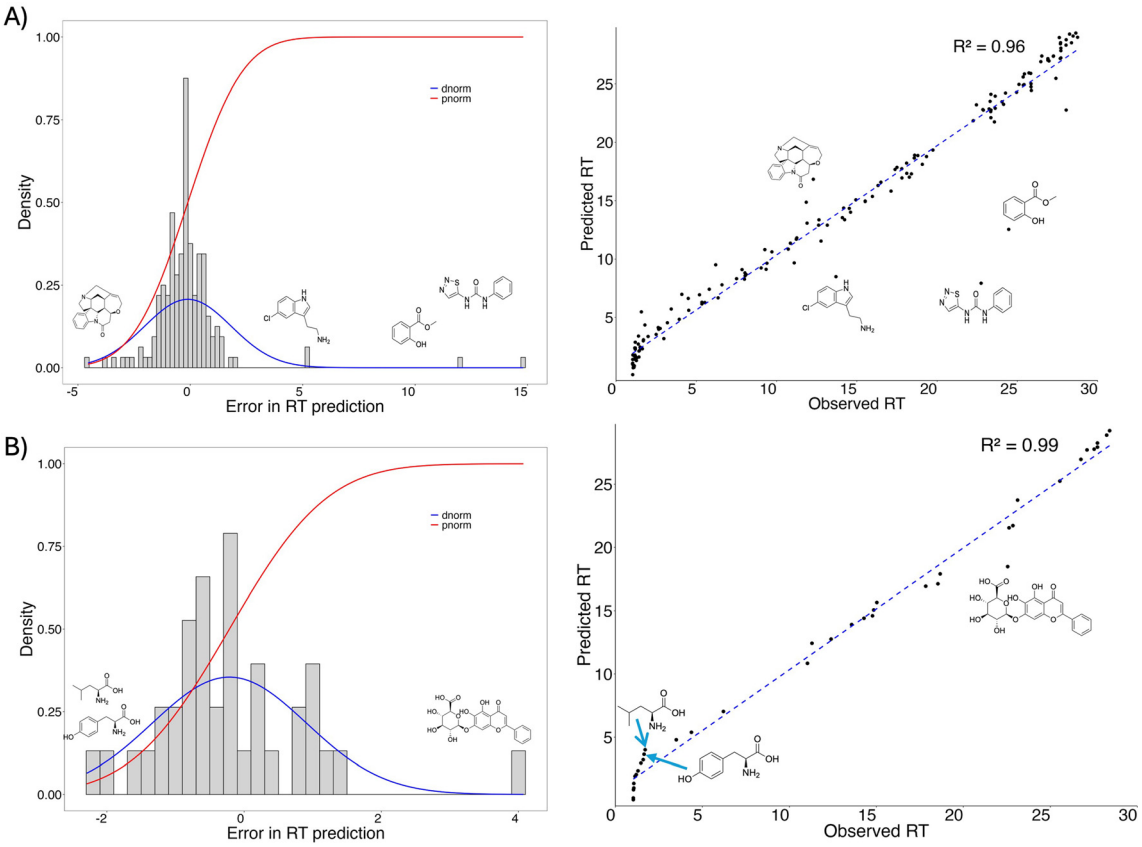

Fig. 3 Model performance for Retip retention time prediction model, autoML: (A) training dataset and (B) external validation dataset.

ing search algorithm and user interface to search against a user uploaded database rather than the internal amino acid datasets. To use this feature, the user first makes their own database of compounds in a CSV file using the same standard format as the "*m/z* screener" and uploads it to "Custom Database Search" along with their input files. The results can be viewed in "Custom Database Output" or downloaded using the "Download Results" button.

Method testing and validation

Putative annotation of npAAs in plants. After manual curation, 229 compounds were found across *A. thaliana* and *C. sativa* leaves, 103 of which were putatively annotated using aminoacidDB and 106 were predicted at MSI level 4 with a molecular formula (Fig. 4). By matching with authentic standards, 15 npAAs were identified at MSI-1 with fully validated analytical methods,[29,40,47] 19 were putatively annotated at MSI-2[7,25,40] by matching the accurate mass (<5 ppm) and RT (>75%) against aminoacidDB datasets, and 50 were putatively annotated at MSI-3[1,40] (Tables 2, S6 and Fig. 4). Given that aminoacidDB contains amino acid derivatives and peptides, 29 dipeptides and 2 tripeptides were also putatively annotated.

The npAA profile of *C. sativa* was more diverse and unique than *A. thaliana*, with 44 npAAs putatively annotated (Fig. 4). This may be attributed to the difference in their species, stage

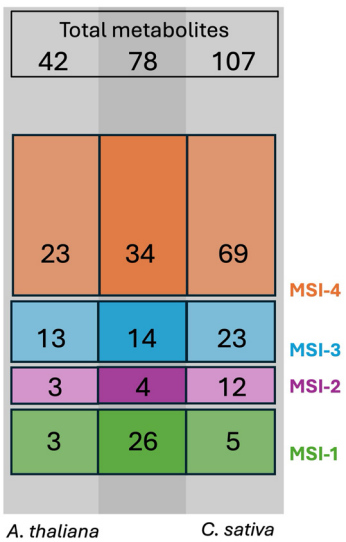

Fig. 4 Level of certainty and distribution of metabolites putatively annotated in *Arabidopsis thaliana* and *Cannabis sativa* leaf tissues using aminoacidDB. Metabolites in the darker shaded region are commonly present in both species.

of growth and/or environmental conditions.[12,27,28,48,49] npAAs are used as a tool in chemotaxonomy as many are produced by specific species, families or taxa of plants,[50,51] while a change

Table 2 Detection of non-protein amino acids in *Arabidopsis thaliana* and *Cannabis sativa*

Amino acid detected	MSI confidence level[40]	Detected in Arabidopsis	Literature report on Arabidopsis	Detected in Cannabis	Literature report in Cannabis
Glutamate, GABA, and alanine metabolism					
L-Alanine	1	✓		✓	
L-Glutamine	1	✓		✓	
L-Glutamic acid	1	✓		✓	
Beta-alanine	1	✓	✓	✓	
Gamma-aminobutyric acid	1	✓	✓	✓	
L-2,4-Diaminobutyric acid	1	✓		✓	
N-(2-Aminoethylglycine)	1	✓			
Beta-*N*-methylamino-L-alanine	1			✓	
Pyroglutamic acid	2	✓	✓	✓	
Alpha-aminobutyric acid	2		✓	✓	
N-Methylalanine	2			✓	
Gamma-acetyldiaminobutyric acid; 4-acetamido-2-aminobutanoic acid	3	✓			
3-(*N*-methylamino)glutaric acid, 4-Amino-5-methoxy-5-oxopentanoic acid; (±)-2,2′-iminobispropanoic acid	3	✓		✓	
N-Carboxyethyl-g-aminobutyric acid	3	✓			
Gamma-glutamyl-beta-cyanoalanine	3	✓			
Iminodiacetic acid	3			✓	
4-Methylglutamate, (2*S*,4*S*,5*S*)-4,5-dihydroxypiperidine-2-carboxylic acid, *O*-acetyl-L-homoserine, 4-Amino-5-methoxy-5-oxopentanoic acid, glutamic acid gamma-methyl ester, 2-aminoadipic acid, (±)-2,2′-iminobispropanoic acid	3			✓	
(3*S*,4*R*)-3-Hydroxy-4-methyl-L-glutamic acid, (2s,4r)-4-hydroxy-4-methyl-glutamic acid, (2*S*,3*R*,4*R*,5*S*)-3,4,5-trihydroxypiperidine-2-carboxylic acid	3			✓	
Norophthalmic acid, glutaminylglutamic acid, glutamyl-Gamma-glutamate, glutamylglutamine, *N*-gamma-Glutamylglutamine	3			✓	
(2*S*)-2-Amino-3-ethylsulfinylpropanoic acid; $C_5H_{11}NO_3S$	4	✓		✓	
3-Aminopropyl(methyl)carbamic acid; $C_5H_{12}N_2O_2$	4	✓			
2-Amino[(propan-2-yl)amino]butanoic acid; $C_7H_{16}N_2O_2$	4			✓	
4-Amino-5-mercapto-2-methylpentanoic acid; C6H13NO2S	4			✓	
2-(3-Furanylmethylamino)propanoic acid; $C_8H_{11}NO_3$	4			✓	
2-(5-Ethyl-1*H*-1,2,4-triazol-3-yl)propanoic acid; C7H11N3O2	4			✓	
Arginine, proline, citrulline, and ornithine metabolism					
L-Asparagine	1	✓		✓	
L-Aspartic acid	1	✓		✓	
L-Arginine	1	✓		✓	
DL-Citrulline	1	✓	✓	✓	✓
L-Proline	1	✓		✓	
DL-Ornithine	1	✓	✓	✓	✓
4-Oxo-L-proline	2	✓		✓	
4-Hydroxyornithine	2			✓	
N-Acetyl-L-ornithine	2	✓	✓	✓	✓
Glycylprolylhydroxyproline	2			✓	
Prolyl-arginine, arginylproline	3	✓		✓	
2-Aminoacrylic acid	3			✓	
Spermic acid 2	3			✓	
Asymmetric dimethylarginine, symmetric Dimethylarginine	3			✓	
Valylproline, dethiobiotin, monascustin	3	✓			
Isoleucylproline, leucylproline, 5*S*-hydroxynorvaline-S-Ile, Pro-leu	3	✓			
Histidine metabolism					
Leucyl-histidine	2	✓			
2-Pyrrolidinecarboxylic acid; C5H9NO2	4	✓			
3-Methyl-2,3-dihydro-1*H*-pyrrole-5-carboxylic acid; C6H9NO2	4	✓			
2,3-Dihydro-1*H*-pyrazine-4-carboxylic acid; $C_5H_8N_2O_2$	4	✓		✓	
3-(2-Aminoethyl)-4-imidazolecarboxylic acid; $C_6H_9N_3O_2$	4	✓		✓	
Thiazolidine-2,5-dicarboxylic acid; $C_5H_7NO_4S$	4	✓			
2,6-Dimethyl-1,4-dihydropyridine-3-carboxylic acid; $C_8H_{11}NO_2$	4			✓	
2-(4,5-Dihydro-1*H*-imidazol-2-yl)ethyl carbamic acid; $C_6H_{11}N_3O_2$	4			✓	
Lysine degradation pathways					
L-Lysine	1	✓		✓	
N(6)-Methyllysine	2	✓		✓	
Saccharopine; $C_{11}H_{20}N_2O_6$	4	✓		✓	
Shikimate Metabolism					
L-Phenylalanine	1	✓		✓	
L-Tyrosine	1	✓		✓	
DL-Beta-3,4-dihydroxyphenylalanine	1			✓	
4-Chloro-L-phenylalanine	1			✓	
N-(1-Deoxy-1-fructosyl)tyrosine	2			✓	

Table 2 *(Contd.)*

Amino acid detected	MSI confidence level[40]	Detected in Arabidopsis	Literature report on Arabidopsis	Detected in Cannabis	Literature report in Cannabis
Histidylphenylalanine, phenylalanylhistidine	3	✓		✓	
(1*R*,6*S*)-6-Amino-5-oxocyclohex-2-ene-1-carboxylic acid, (5*R*,6*R*)-6-amino-5-hydroxycyclohexa-1,3-diene-1-carboxylic acid	3			✓	
2,5-Dihydrophenylalanine, L-dihydrophenylalanine, (2*S*)-2-amino-3-cyclohexa-2,4-dien-1-ylpropanoic acid, Norpandamarilactonine A	3			✓	
Tryptophan metabolism					
L-Tryptophan	1	✓		✓	
5-Hydroxy-L-tryptophan	1	✓		✓	
L-Kynurenine	1			✓	✓
Asparaginyl-tryptophan, tryptophyl-asparagine	3			✓	
Lysyltryptophan, tryptophyl-lysine	3			✓	
Serine family metabolism					
Glycine	1	✓		✓	
L-Serine	1	✓		✓	
L-Cysteine	1	✓			
L-Methionine	1	✓		✓	
L-Cystine	1	✓		✓	
1-Aminocyclopropane-1-carboxylic acid	1	✓	✓	✓	
Oxiglutathione	2		✓	✓	✓
Homo-L-arginine	2		✓	✓	
O-Acetylserine	2			✓	
Rhizobitoxine	2			✓	
DL-Homoserine	3	✓		✓	
Threonylglycine, alanylserine	3	✓		✓	
4-Hydroxyphenylglycine	3			✓	
Branched chain amino acid metabolism					
L-Valine	1	✓		✓	
L-Threonine	1	✓		✓	
L-Isoleucine	1	✓		✓	
L-Leucine	1	✓		✓	
L-2-Amino-5-hydroxypentanoic acid; (2s,4s)-4-hydroxynorvaline; L-pentahomoserine; 3-hydroxynorvaline	3	✓		✓	
5-Methylnorleucine, *N*-methyl-L-isoleucine, (2*S*)-2-amino-4-methylhexanoic acid	3	✓			

in environment, including exposure to stress, also affects the production of npAAs in plants.[12,27,28,48,49] 12 npAAs including ACC, beta-alanine, 5-hydroxy-L-tryptophan, gamma-aminobutyric acid (GABA), Cit, Orn, *N*-acetyl-L-ornithine (AcOrn), L-cystine, L-2,4-diaminobutyric acid (DAB), pyroglutamic acid, 4-oxo-L-proline (4-OP), and *N*(6)-methyllysine were detected in both species (Fig. 5). *N*-(2-Aminoethylglycine) (AEG) was detected in *A. thaliana*, while 13 npAAs including 6-aminohexanoic acid, 4-chloro-L-phenylalanine, β BMAA, L-kynurenine, DOPA, alpha-aminobutyric acid (AABA), *N*-methylalanine, *O*-acetylserine, homo-L-arginine, rhizobitoxine, *N*-(1-deoxy-1-fructosyl)tyrosine, terazosin and 4-hydroxyornithine were specific to *C. sativa* (Table 2 and S6 and Fig. 5). Cit, Orn, AcOrn, GABA, pyroglutamic acid, 4-OP, and L-cystine are found commonly across the plant kingdom as part of primary metabolism.[52–55] L-kynurenine and 5-hydroxy-L-tryptophan are metabolites of tryptophan metabolism involved in phytohormone regulation in plants[25,48,56] and neuromodulating functions in humans.[56,57] Kynurenine was recently quantified in the μg g^{-1} range in *C. sativa.*[56] DAB, BMAA and AEG are structural isomers produced by cyanobacteria and other microbes.[29,58,59] DAB and BMAA are potentially neurotoxic, whereas AEG functions as a component of peptide nucleic acids, which are genetic materials before RNA/DNA.[22,29,60] Within the plant kingdom, DAB and BMAA have been reported in *Cycas revoluta* and *Lathyrus sativus* previously.[21,61,62]

Our data demonstrate the wider distribution of these npAAs in plants. Given that npAAs contain highly reactive amino and carboxylic acid functional groups, it would be interesting to understand their role in protein chemistry through (mis)-incorporation and direct or indirect interactions.[20,63] For example, post-translational methylation of the ε-amino groups of lysine in proteins is a known mechanism for regulating protein–protein interactions, protein function, and transportation.[64,65] The presence of *N*(6)-methyllysine in *A. thaliana* and *C. sativa* indicates the free-form availability of this metabolite in plants with unknown roles.

Among the peptides, leucyl-histidine and cyclo(-his-phe) were found in *A. thaliana*, while valylvaline, glycylprolylhydroxyproline, and oxiglutathione were observed in *C. sativa* (Table 2 and S6 and Fig. 5). Dipeptides (linear or cyclic) have emerged as small molecule regulators in plants but limited knowledge is available on their diversity, distribution and function.[66,67]

Meta-analysis of previous data from published studies. Three previous studies on *A. thaliana* with their data shared

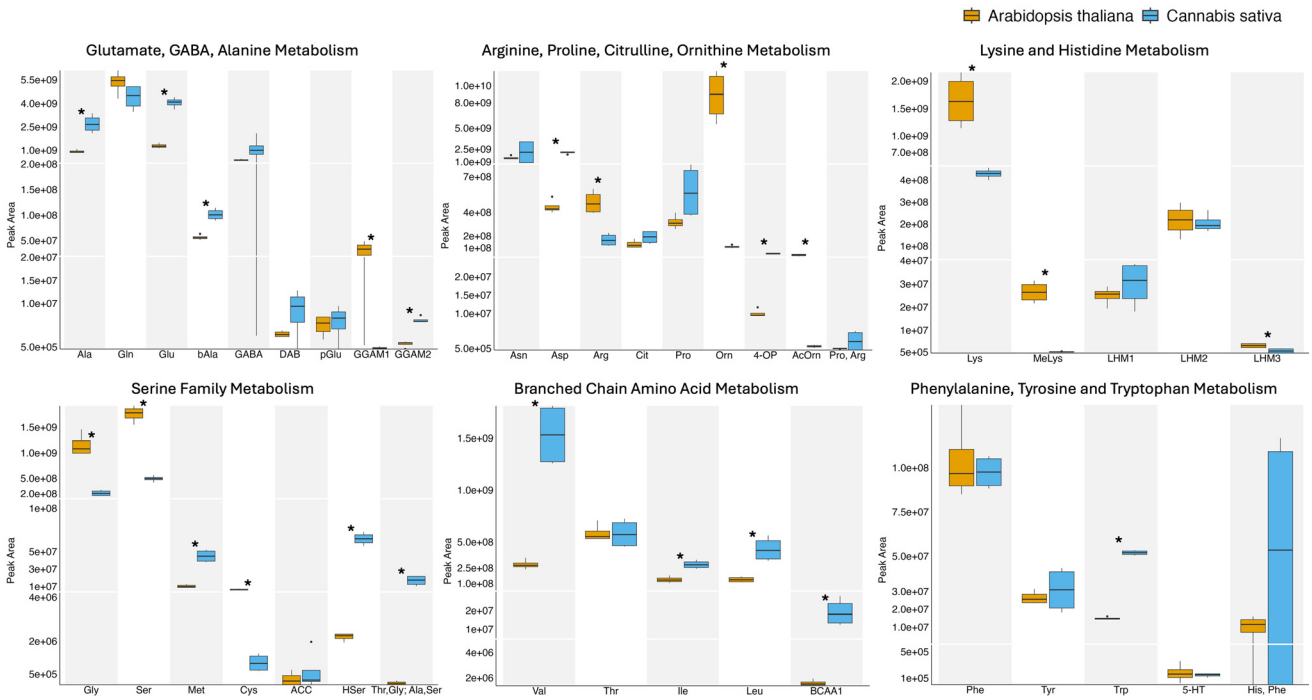

Fig. 5 Box plots depicting the distribution of selected amino acids across various amino acid families in the leaf tissues of *Arabidopsis thaliana* and *Cannabis sativa*. *X*-Axis shows peak areas (n = 4) for each of the amino acids on *Y*-axis. Asterisk denotes significant differences at $p < 0.05$.

publicly were compared to our primary data (Fig. 6). 26% of the metabolites in our *A. thaliana* dataset were also present in at least one of the studies from the Metabolomics Workbench, while <1% (3 features) were common across all the datasets (Fig. 6). The previously published data was generated from different extraction solvents and sample preparation methods that were not optimized for amino acids. Also, the plants were grown under different conditions and the data acquisition and processing parameters varied across the studies (Table S7). More than 84% of the features that were common between at least two studies from the Metabolomics Workbench were shared by study 1 and 2 (Fig. 6), which were conducted by the same researchers using the same analysis (column, solvents, gradient and instruments) conditions (Table S7). Using aminoacidDB datasets, we putatively annotated 86 metabolites across the three studies, 20 of which were also present in our *A. thaliana* dataset (Table S8) including 12 protein amino acids and 6 npAAs such as pyroglutamic acid, 4-oxo-L-proline, DL-norleucine, *N*-allylglycine, and 5-hydroxy-4-oxonorvaline (Table S8). Norvaline was added as an internal standard by the researchers in one of the previously published datasets (Table S7) and was annotated with ~88% RT match and ~2 ppm mass error, highlighting the wider applicability of our method in annotating npAAs given that the study employs a different C_{18} column and LC-system than the method used to develop aminoacidDB (Table S7). 12 additional npAAs and 23 peptides were also putatively annotated across the *A. thaliana* studies, demonstrating the unexplored diversity of npAAs in existing datasets.

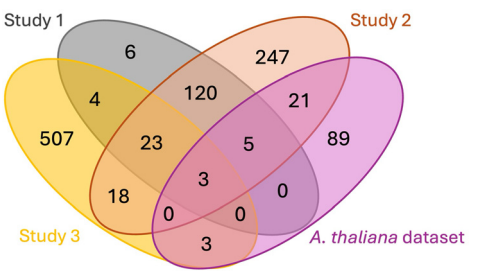

Fig. 6 Comparison of public data from three previous studies on *Arabidopsis thaliana* from the Metabolomics Workbench (Study 1–3) with *A. thaliana* dataset from the current study.

Future directions and applications

npAAs remain an under-investigated class of small molecules despite their potent roles in physiology, stress response and ecosystems. AminoacidDB contains over 300,000 amino acids from diverse sources, along with various MS adducts for use with untargeted MS datasets collected with a wide variety of instruments, techniques and methodologies. Using aminoacidDB, researchers can dissect the identity of wide diversity of npAAs and unravel the complexity of proteins in combination with emergent protein sequencing platforms. Future studies to profile npAAs in free-form or in complex with proteins would help elucidate their roles in protein biochemistry and cellular physiology, with applications in human health, agriculture and food safety.

Conflicts of interest

The authors declare no conflict of interest.

Data availability

Data are available in our website application at https://www.aminoaciddb.ca. The LC-MS files are available through Borealis (https://borealisdata.ca/dataset.xhtml?persistentId=doi:10.5683/SP3/DTAZ1G).

Supplementary information (SI): Fig. S1–S3 and Tables S1–S8. See DOI: https://doi.org/10.1039/d5an01248a, .

Acknowledgements

The work was supported by the Vanier Canada Graduate Scholarships (Vanier CGS) program and Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1 C. E. Turi, J. Finley, P. R. Shipley, S. J. Murch and P. N. Brown, Metabolomics for phytochemical discovery: Development of statistical approaches using a cranberry model system, *J. Nat. Prod.*, 2015, **78**, 953–966.
2 A. Aharoni, R. Goodacre and A. R. Fernie, Plant and microbial sciences as key drivers in the development of metabolomics research, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2217383120.
3 Y. Xu and R. Goodacre, Mind your Ps and Qs – Caveats in metabolomics data analysis, *TrAC, Trends Anal. Chem.*, 2025, **183**, 118064.
4 J. Guo, H. Yu, S. Xing and T. Huan, Addressing big data challenges in mass spectrometry-based metabolomics, *Chem. Commun.*, 2022, **58**, 9979–9990.
5 B. B. Misra, New software tools, databases, and resources in metabolomics: updates from 2020, *Metabolomics*, 2021, **17**, 49.
6 M. Krassowski, V. Das, S. K. Sahu and B. B. Misra, State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing, *Front. Genet.*, 2020, **11**, 610798.
7 R. T. Giebelhaus, L. A. E. Erland and S. J. Murch, HormonomicsDB: a novel workflow for the untargeted analysis of plant growth regulators and hormones, *F1000*, 2024, **11**, 1191.
8 I. Gertsman and B. A. Barshop, Promises and pitfalls of untargeted metabolomics, *J. Inherited Metab. Dis.*, 2018, **41**, 355–366.
9 M. K. Reddy, amino acid, *Encycl. Brit.*, 2025, **29**, https://www.britannica.com/science/amino-acid.
10 K. Bailey and F. Sanger, The Chemistry of Amino Acids and Proteins, *Annu. Rev. Biochem.*, 1951, **20**, 103–130.
11 J. B. Hedges and K. S. Ryan, Biosynthetic Pathways to Nonproteinogenic α-Amino Acids, *Chem. Rev.*, 2020, **120**, 3161–3209.
12 E. A. Bell, Nonprotein Amino Acids of Plants: Significance in Medicine, Nutrition, and Agriculture, *J. Agric. Food Chem.*, 2003, **51**, 2854–2865.
13 P. B. Nunn, E. A. Bell, A. A. Watson and R. J. Nash, Toxicity of Non-protein Amino Acids to Humans and Domestic Animals, *Nat. Prod. Commun.*, 2010, **5**, 1934578X1000500329.
14 B. A. Boughton, P. Reddy, M. P. Boland, U. Roessner and P. Yates, Non-protein amino acids in Australian acacia seed: Implications for food security and recommended processing methods to reduce djenkolic acid, *Food Chem.*, 2015, **179**, 109–115.
15 U. J. Meierhenrich, G. M. M. Caro, J. H. Bredehöft, E. K. Jessberger, W. H.-P. Thiemann and E. F. van Dishoeck, Identification of Diamino Acids in the Murchison Meteorite, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 9182–9186.
16 K. J. Rodgers, J. Kabalan and C. R. Phillips, A comprehensive review of the proline mimic azetidine-2-carboxylic acid (A2C), *Toxicology*, 2025, **510**, 153999.
17 N. Wang, N. C. Bunawan, A. Rastegar and K. White, Djenkolism: case report and literature review, *Int. Med. Case Rep. J.*, 2014, **16**, 79–84.
18 R. J. Bridges and C. Hatalski, *Gliotoxic properties of the Lathyrus excitotoxin fl-N-oxalyl-L-a,fl-diaminopropionic acid (fl-L-ODAP)Brain Research*, 1991, **561**, 262–268.
19 A. Vega and E. A. Bell, α-Amino-β-methylaminopropionic acid, a new amino acid from seeds of Cycas circinalis, *Phytochemistry*, 1967, **6**, 759–762.
20 J. R. Steele, C. J. Italiano, C. R. Phillips, J. P. Violi, L. Pu, K. J. Rodgers and M. P. Padula, Misincorporation proteomics technologies: A review, *Proteomes*, 2021, **9**, 1–20.
21 S. J. Murch, P. A. Cox and S. A. Banack, A mechanism for slow release of biomagnified cyanobacterial neurotoxins and neurodegenerative disease in Guam., *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 12228–12231.
22 W. B. Glover, D. C. Mash and S. J. Murch, The natural non-protein amino acid *N*-β-methylamino-L-alanine (BMAA) is incorporated into protein during synthesis, *Amino Acids*, 2014, **46**, 2553–2559.
23 R. A. Dunlop, P. A. Cox, S. A. Banack and K. J. Rodgers, The Non-Protein Amino Acid BMAA Is Misincorporated into Human Proteins in Place of L-Serine Causing Protein Misfolding and Aggregation, *PLoS One*, 2013, **8**, 1–8.
24 S. W. Chan, R. A. Dunlop, A. Rowe, K. L. Double and K. J. Rodgers, l-DOPA is incorporated into brain proteins of patients treated for Parkinson's disease, inducing toxicity in human neuroblastoma cells in vitro, *Exp. Neurol.*, 2012, **238**, 29–37.
25 L. A. E. Erland, C. E. Turi, P. K. Saxena and S. J. Murch, Metabolomics and hormonomics to crack the code of filbert growth, *Metabolomics*, 2020, **16**, 62.
26 S. J. Murch, H. P. V. Rupasinghe, D. Goodenowe and P. K. Saxena, A metabolomic analysis of medicinal diversity

in Huang-qin (*Scutellaria baicalensis* Georgi) genotypes: Discovery of novel compounds, *Plant Cell Rep.*, 2004, **23**, 419–425.

27 R. T. Giebelhaus, L. Biggs, S. J. Murch and L. A. E. Erland, Untargeted and Targeted Metabolomics to Understand Plant Growth Regulation and Evolution in Wollemi Pine (*Wollemia nobilis*), *Botany*, 2023, **101**, 377–390.

28 A. L. Greene, P. K. Sandhu, B. A. Hall, R. O'Brien, R. Hayward and S. J. Murch, Metabolomics methods to understand the impacts of wildfires on wines, *ACS Food Sci. Technol.*, 2025, **5**, 3759–3773.

29 W. B. Glover, T. C. Baker, S. J. Murch and P. Brown, Determination of β-*N*-methylamino-L-alanine, *N*-(2-aminoethyl)glycine, and 2,4-diaminobutyric acid in Food Products Containing Cyanobacteria by Ultra-Performance Liquid Chromatography and Tandem Mass Spectrometry: Single-Laboratory Validation, *J. AOAC Int.*, 2015, **98**, 1559–1565.

30 C. Salazar, J. M. Armenta and V. Shulaev, An UPLC-ESI-MS/MS Assay Using 6-Aminoquinolyl-*N*-Hydroxysuccinimidyl Carbamate Derivatization for Targeted Amino Acid Analysis: Application to Screening of Arabidopsis thaliana Mutants, *Metabolites*, 2012, **2**, 398–428.

31 P. K. Sandhu, J. T. Solonenka and S. J. Murch, Neurotoxic non-protein amino acids in commercially harvested Lobsters (*Homarus americanus* H. Milne-Edwards), *Sci. Rep.*, 2023, **14**, 8017.

32 W. Xu, C. Zhong, C. Zou, B. Wang and N. Zhang, Analytical methods for amino acid determination in organisms, *Amino Acids*, 2020, **52**, 1071–1088.

33 A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson and P.-M. Allard, The LOTUS initiative for open knowledge management in natural products research, *eLife*, 2022, **11**, e70780.

34 D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V. W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng, R. Mandal, N. Karu, M. Dambrova, H. B. Schiöth, R. Greiner and V. Gautam, HMDB 5.0: the Human Metabolome Database for 2022, *Nucleic Acids Res.*, 2022, **50**, D622–D631.

35 Y. D. Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D. S. Wishart, ClassyFire: automated chemical classification with a comprehensive, computable taxonomy, *J. Cheminf.*, 2016, **8**, 61.

36 G. Landrum, P. Tosco, B. Kelley, R. Rodriguez, D. Cosgrove, R. Vianello, P. Sriniker, G. Jones, E. Kawashima, N. Schneider, D. Nealschneider, A. Dalke, M. Tadhurst-cdd, B. Cole, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, H. Faara, V. F. Scalfani, R. Walker, D. Probst, K. Ujihara, N. Maeder, J. Monat, J. Lehtivarjo and G. Godin, 2025, DOI: DOI: 10.5281/zenodo.15286010.

37 H. Wickman, *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York, 2016.

38 S. Xu, M. Chen, T. Feng, L. Zhan, L. Zhou and G. Yu, Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers, *Front. Genet.*, 2021, **12**, DOI: 10.3389/fgene.2021.774846.

39 P. Bonini, T. Kind, H. Tsugawa, D. K. Barupal and O. Fiehn, Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics, *Anal. Chem.*, 2020, **92**, 7515–7522.

40 L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden and M. R. Viant, Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI), *Metabolomics*, 2007, **3**, 211–221.

41 L. Buzásy, K. Mazák, B. Balogh, B. Simon, A. Vincze, G. T. Balogh, T. Pálla and A. Mirzahosseini, Physicochemical Characterization of Kynurenine Pathway Metabolites, *Antioxidants*, 2025, **14**, 589.

42 Q. Wu, J.-Y. Wang, D.-Q. Han and Z.-P. Yao, Recent advances in differentiation of isomers by ion mobility mass spectrometry, *TrAC, Trends Anal. Chem.*, 2020, **124**, 115801.

43 R. Joshi, S. Sharma and D. Kumar, Advances of Ion Mobility Platform for Plant Metabolomics, *Crit. Rev. Anal. Chem.*, 2024, **54**, 175–191.

44 E. A. H. Keppler, C. L. Jenkins, T. J. Davis and H. D. Bean, Advances in the application of comprehensive two-dimensional gas chromatography in metabolomics, *TrAC, Trends Anal. Chem.*, 2018, **109**, 275–286.

45 R. Pascoe, J. P. Foley and A. I. Gusev, Reduction in Matrix-Related Signal Suppression Effects in Electrospray Ionization Mass Spectrometry Using On-Line Two-Dimensional Liquid Chromatography, *Anal. Chem.*, 2001, **73**, 6014–6023.

46 P. Dugo, F. Cacciola, T. Kumm, G. Dugo and L. Mondello, Comprehensive multidimensional liquid chromatography: Theory and applications, *J. Chromatogr. A*, 2008, **1184**, 353–368.

47 F. J. M. Tymm, S. L. Bishop and S. J. Murch, A Single Laboratory Validation for the Analysis of Underivatized β-*N*-Methylamino-L-Alanine (BMAA), *Neurotoxic. Res.*, 2020, **11**, 1–23.

48 J. A. Forsyth, L. A. Erland, P. R. Shipley and S. J. Murch, Plant Perception of Light: The role of indoleamines in Scutellaria species, *Melatonin Res.*, 2020, **3**, 161–176.

49 L. Fowden, P. J. Lea and E. A. Bell, in *Advances in Enzymology – and Related Areas of Molecular Biology*, ed. A. Meister, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006, pp. 117–175.

50 E. A. Bell, A. A. Watson and R. J. Nash, Non-Protein Amino Acids: A Review of the Biosynthesis and Taxonomic

Significance, *Nat. Prod. Commun.*, 2008, **3**, 1934578X0800300.

51 M. Gibson, W. T. Santos, A. R. Oyler, L. Busta and C. A. Schenck, A new spin on chemotaxonomy: Using non–proteogenic amino acids as a test case, *Appl. Plant Sci.*, 2025, **13**, e70006.

52 V. Joshi and A. R. Fernie, Citrulline metabolism in plants, *Amino Acids*, 2017, **49**, 1543–1559.

53 L. Li, N. Dou, H. Zhang and C. Wu, The versatile GABA in plants, *Plant Signaling Behav.*, 2021, **16**, 1862565.

54 R. Majumdar, R. Minocha and S. C. Minocha, in *Amino acids in higher plants*, ed. J. P. F. D'Mello, CAB International, UK, 1st edn, 2015, pp. 156–176.

55 B. Van de Poel and D. Van Der Straeten, 1-Aminocyclopropane-1-carboxylic acid (ACC) in plants: more than just the precursor of ethylene!, *Front. Plant Sci.*, 2014, **5**, 640.

56 F. Russo, F. Tolomeo, M. A. Vandelli, G. Biagini, R. Paris, F. Fulvio, A. Laganà, A. L. Capriotti, L. Carbone, G. Gigli, G. Cannazza and C. Citti, Kynurenine and kynurenic acid: Two human neuromodulators found in *Cannabis sativa* L., *J. Pharm. Biomed. Anal.*, 2022, **211**, 114636.

57 M. E. Maffei, 5-Hydroxytryptophan (5-HTP): Natural Occurrence, Analysis, Biosynthesis, Biotechnology, Physiology and Toxicology, *Int. J. Mol. Sci.*, 2020, **22**, 181.

58 S. L. Bishop, J. K. Kerkovius, F. Menard and S. J. Murch, *N*-β-Methylamino-L-Alanine and Its Naturally Occurring Isomers in Cyanobacterial Blooms in Lake Winnipeg, *Neurotoxic. Res.*, 2018, **33**, 133–142.

59 T. C. Baker, F. J. M. Tymm and S. J. Murch, Assessing Environmental Exposure to β-*N*-Methylamino-l-Alanine (BMAA) in Complex Sample Matrices: a Comparison of the Three Most Popular LC-MS/MS Methods, *Neurotoxic. Res.*, 2018, **33**, 43–54.

60 S. L. Bishop, F. J. M. Tymm, K. Perry, J. K. Kerkovius, F. Menard, A. Brady, G. Slater, D. S. S. Lim, J. S. Metcalf, S. A. Banack, P. A. Cox and S. J. Murch, Early-earth nonprotein amino acid metabolites in modern cyanobacterial microbialites, *Environ. Chem. Lett.*, 2020, **18**, 467–473.

61 T. Krüger, B. Mönch, S. Oppenhäuser and B. Luckas, LC-MS/MS determination of the isomeric neurotoxins BMAA (β-*N*-methylamino-l-alanine) and DAB (2,4-diaminobutyric acid) in cyanobacteria and seeds of *Cycas revoluta* and *Lathyrus latifolius*, *Toxicon*, 2010, **55**, 547–557.

62 P. A. Cox, S. A. Banack and S. J. Murch, Biomagnification of cyanobacterial neurotoxins and neurodegenerative disease among the Chamorro people of Guam, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 13380–13383.

63 X. Luo, C. C. A. Ng, H. H. N. Lam and Z.-P. Yao, Advances in protein sequencing: Techniques, challenges and prospects, *TrAC, Trends Anal. Chem.*, 2025, **191**, 118341.

64 N. B. C. Serre, C. Alban, J. Bourguignon and S. Ravanel, An outlook on lysine methylation of non-histone proteins in plants, *J. Exp. Bot.*, 2018, **69**, 4569–4581.

65 N. Mezey, W. C. S. Cho and K. K. Biggar, Intriguing Origins of Protein Lysine Methylation: Influencing Cell Function Through Dynamic Methylation, *Genomics, Proteomics Bioinf.*, 2019, **17**, 551–557.

66 P. Agarwal, H. D. Fischer, M. D. Camalle and A. Skirycz, Not to be overlooked: dipeptides and their role in plant stress resilience, *J. Exp. Bot.*, 2025, **76**, 5738–5747.

67 R. I. Minen, M. D. Camalle, T. J. Schwertfeger, F. Abdulhakim, H. Reish, L. Perez de Souza, J. C. Moreno, A. Schilmiller, V. P. Thirumalaikumar, P. Agarwal, C. F. Plecki, A. R. Fernie, H. Hirt, F. C. Schroeder and A. Skirycz, Characterization of the cyclic dipeptide cyclo (His-Pro) in Arabidopsis, *Plant Physiol.*, 2025, **198**, kiaf174.