Analyst

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2025, **150**, 2612

Hyphenated mass spectrometry methods for enlarged capacity data storage systems based on chemical mixtures†

Victor Flors,[a] Raquel Cerveró,[a] Cristopher Tinajero,[b] Victor Sans [b] and Cristian Vicent *[c]

Encoding abstract information in chemical mixtures uses the selective presence or absence of specific analytes, creating a binary-based framework for data storage. Data storage capacity (*C* in bits) can be maximized by encoding with large analyte libraries (*M*) at distinguishable concentration levels (*L*), where $C = M \cdot \log_2 L$. However, roboust decoding of such complex libraries remains challenging for practical applications. This study introduces hyphenated mass spectrometry (MS) methods, liquid chromatography (LC) and flow injection analysis (FIA) that meet the dual requirements of high analyte coverage and precise quantitation to maximize data storage capacity. Encoding and decoding use plant metabolite libraries to create specific mixtures. Using LC-MS, it is feasible to encode and decode up to 200 bits per mixture, with scalability reaching 10^3–10^4 bits at the cost of low decoding rates (*ca.* 0.5 bits per sec). FIA-MS offers a high-throughput alternative, handling 100 bits at faster rates (*ca.* 3 bits per sec). The data storage capacity can be three-fold expanded by incorporating up to eight quantitation levels, supporting binary, quaternary, or octal encoding schemes. To demonstrate the practical application of these methods, we encode and decode various digital file formats such as texts and multicolor images.

Received 28th March 2025,
Accepted 6th May 2025

DOI: 10.1039/d5an00353a

rsc.li/analyst

Introduction

Molecular data storage systems have emerged as a promising alternative for information storage.[1,2] Recent advancements in molecular-scale data systems have primarily leveraged DNA sequences[3–6] and synthetic sequence-defined polymers[7–13] for storing digital information. An alternative approach involves using chemical mixtures, where the presence or absence of specific analytes can be encoded as binary "1" and "0" sequences. Such chemical mixtures resemble naturally-occurring metabolomes, in which metabolic states are encoded by collections of small-molecule metabolites. For example, plants have evolved to encode complex chemical information in metabolomes comprising metabolites, cations, and anions, in response to environmental stimuli such as stress, pests, or nutrient availability.[14,15] In contrast to natural metabolomes, where metabolite identities are determined by biology, chemical mixtures can be selectively designed for stability under long-term storage conditions or tuned for decoding through standard analytical methods. Diverse molecular classes, including the pioneer works using oligopeptides,[16] or synthetic metabolomes,[17] have been explored for storing digital information using Mass Spectrometry methods. Likewise, the use of nicotinic acid derivatives,[18] quinoline isotopologues[19] or mixtures of waste chemicals[20] has been also reported. Combinatorial chemistry has further advanced the encoding of extensive digital datasets in small organic molecules.[21,22]

Most studies utilize laser desorption ionization (LDI) methods coupled with mass spectrometry (MS) for abstract information encoding and decoding. Highly efficient data assembly, achieved by positioning chemical mixtures on predefined μm^2 sized arrays on LDI plates, enables these methods to achieve high information storage densities.[16,21,22] However, the limited analyte coverage (the number of identified metabolites) inherent in standalone mass spectrometry constrains the attainable storage capacity per mixture to a few tens of components. This is critical for encoding large datasets in minimal physical space. The selection of analytes based on

[a]*Plant Immunity and Biochemistry Laboratory, Biochemistry and Molecular Biology Section, Department of Biology, Biochemistry and Natural Sciences, Universitat Jaume I, Castelló, Spain*
[b]*Institute of Advanced Materials (INAM), Universitat Jaume I, Av. Sos Baynat s/n, Castelló 12071, Spain*
[c]*Serveis Centrals d'Instrumentació Científica Universitat Jaume I, Av. Sos Baynat s/n, 12071 Castelló, Spain. E-mail: barrera@uji.es*

†Electronic supplementary information (ESI) available: Detailed experimental procedures and instrumentation was well as codes, output quali- and quantitation results and python scripts for message retrieval are given as supplementary material in *CODES_SI.zip*; .raw data are deposited in **https://doi.org/10.5281/zenodo.15101114**; the complete scripts for identification and quantitation using binary, quaternary and octal schemes are available at **https://github.com/catm542-ai/ChemDataProcessor**. See DOI: **https://doi.org/10.1039/d5an00353a**

distinct molecular characteristics (such as NMR chemical shifts, fluorescence, or Raman shifts) also allows for straightforward encoding and decoding of information using analytical tools like gas chromatography (GC-FID),[23] fluorescence,[24] Raman spectroscopy[25] or nuclear magnetic resonance (NMR).[23,26,27] Table S1† provides an overview of analyte coverage achieved by these encoding and decoding methods. For example, GC-FID can encode around 20 compounds per mixture. Raman spectroscopy and fluorescence-based approaches are restricted to less than 10 analytes per mixture due to signal overlap. NMR, while useful, requires larger sample quantities than MS and presents challenges such as solubility issues and signal overlap (particularly for ^{1}H detection), which could hinder the construction of large analyte sets.

To meet the increasing demands for storage capacity, high-throughput performance, and reliable data readout while maintaining operational simplicity, alternative methods for molecular data storage are being explored. Hyphenated mass spectrometry (MS) techniques represent a promising approach to expand analyte coverage in a single analysis. In particular, liquid chromatography coupled with mass spectrometry (LC-MS) can potentially detect thousands of metabolites in a single analysis.[28–30] For example, chemical mixtures consisting of thousands of synthetic peptide-based molecules can reliably identified by LC-MS.[31] Following the selection of appropriate library compounds that meet criteria for ESI amenability and effective separation in both the *m*/*z* axis and retention time, the encoding of substantial amounts of information (Kb scale) within a single chemical mixture may become feasible. LC-based separation bears the disadvantage of negatively affecting sample throughput, necessitating the compromise between throughput and analyte coverage desired for a specific encoded message. Alternatively, flow injection analysis coupled with mass spectrometry (FIA-MS) offers a streamlined approach that involves transient sample injection into a continuous solvent carrier directly connected to the electrospray ionization (ESI) source of the mass spectrometer.[32] FIA-MS provides lower analyte coverage (typically detecting hundreds of metabolites per run) but offers simplicity, rapid analysis times, and comparable sensitivity and accuracy to LC-MS.[33,34] Despite the high diversity of commercially-available LC-MS platforms and their widespread use in routine analysis, FIA-MS and LC-MS methods remain largely unexplored for data storage applications based on chemical mixtures. Herein, we demonstrate the potential of these methods for high-capacity molecular data storage by successfully encoding and retrieving various computer file formats, including textual data and multicolour images.

Results and discussion

Our initial experiment aimed to optimize and validate the application of FIA-MS and LC-MS methods for the encoding and decoding of abstract information. The schematic representation of our write-encode-decode-process workflow is illustrated in Fig. 1. Information, such as messages encoded in eight-bit ASCII (American standard code for information interchange) and multicolour images, is converted into an equivalent binary molecular code. These encoded messages are subsequently stored in specific chemical mixtures, either within a single mixture (aimed at encoding large datasets in minimal physical space) or distributed across multiple mixtures arranged spatially. The data assembly process involves placing each chemical mixture at predetermined locations on an LC sample plate. Following this, each mixture is decoded using FIA or LC coupled with electrospray ionization mass spectrometry (ESI-MS), where data are automatically processed to recover the original message.

To achieve the highest information storage capacity, each mixture should contain as many kinds of compounds as possible, with reliable identification of these compounds being a key requirement. In the present study, a library of metabolites including flavonoids, plant hormones and dicarboxylic acids was utilized. The widespread use of electrospray ionization mass spectrometry (ESI-MS) as the analytical tool for identifying most flavonoid classes and plant hormones can be attributed to the presence of functional groups such as phenols and carboxylic acids.[35] These groups readily undergo ionization under negative ESI-MS conditions, producing abundant and single $[M - H]^-$ ions, (with minimal formation of other adducts), thus being highly favourable for information encoding and decoding schemes. Once compiled, the metabolite library remained stable for several months (see ESI_1.2† Standard solutions section), obviating the need to regenerate such a large library for each new message encoding.

LC-MS and FIA-MS methods

Our molecular data system considered 200 metabolites (see Table S2†). They were sampled in a single injection taken 8 minutes and an empirical mass spectra database was built from which each metabolite was identified based on its *m*/*z* value and retention time. Such database was later used for target screening of the presence or absence of metabolites in the specific chemical mixtures. This approach provides a storage capacity of 200 bits per mixture and a decoding rate of 0.4 bits per second. Throughput can be enhanced at expense of analyte coverage using FIA-MS.[34,36] However, we observed that the simultaneous introduction of all analytes inherent to FIA-MS analysis leads to significant ion suppression of certain compounds.[37] This negatively affected our automated processing workflow (see below), requiring continuous adjustments to threshold values for metabolite confirmation and thereby limiting the number of metabolites that could be used. Consequently, when employing this simplified FIA-MS encoding/decoding strategy, a reduced subset of metabolites (typically 80–100) was utilized, with each injection taking 35 seconds. This configuration provides a storage capacity of 100 bits per mixture and a decoding rate of *ca.* 3 bits per second. The encoded messages are first prepared at 2.5 μg mL^{-1} (ppm) and finally diluted to the 50–250 ng mL^{-1} (ppb) concen-

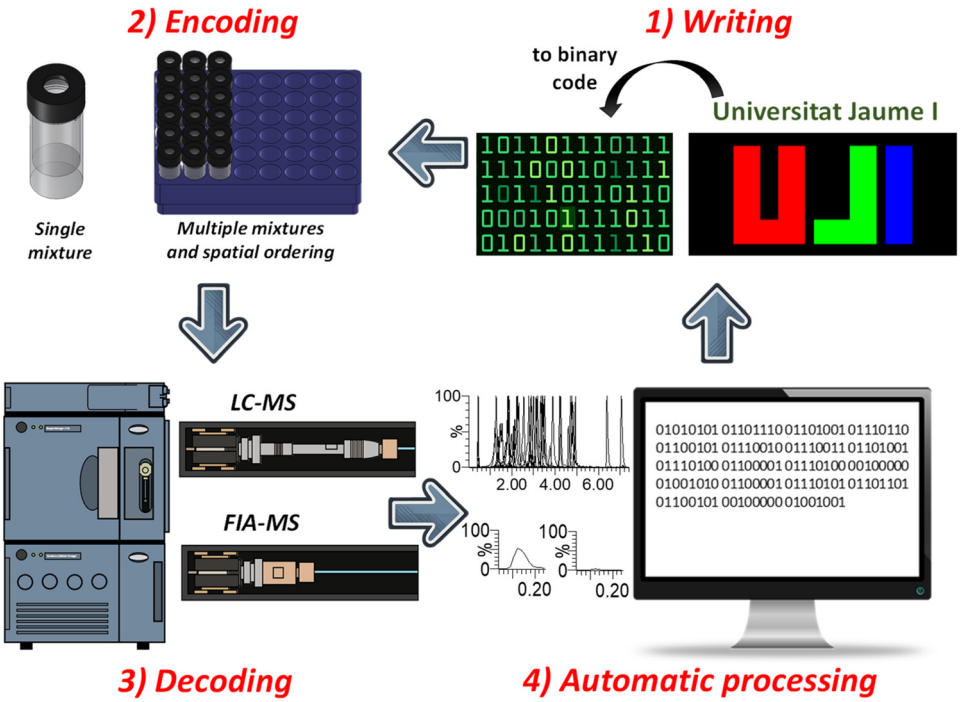

Fig. 1 Schematic representation of the write-encode–decode-process workflow using FIA-MS and LC-MS.

trations compatible with FIA-MS and LS-MS methods (see experimental section). Both methods require minimal sample volumes, typically in the range of 1–5 μL, thus enabling multiple 10^4 readouts of molecular information from any given mixture.

Ranking different procedures for storing information relies on number and density of places where information can be stored and the amount of information that can be stored in each location or single mixture.[16] The present methods offer potential for high-density storage of information per single mixture (200 bits per mixture) requiring minimal physical space. The encoded messages can be stored in the solid state by adding 5 μL (2.5 ppm) of the specific chemical mixture on a small portion of filter paper (10 mm^2), thus attaining a storage density of (250 bytes per cm^2). It can be stored for several months and reconstituted by adding 200 μL of methanol, filtered and directly decoded by FIA-MS and LC-MS. The spatial organization of multiple mixtures enhances storage capacity linearly while reducing the theoretical storage density relative to single metabolite mixtures. As a result, greater physical space becomes essential to support the storage of such information. The density of information achieved using spatial ordering depends on the instrumentation vendor's design. In the present work, sample plates have 48 positions occupying *ca.* 70 cm^2, so the attainable spatial density is modest (19 bytes per cm^2); this can be linearly enhanced by increasing the number of sample plates available (well-plates of 394 positions occupying 70 cm^2 are also available, thus reaching storage densities of 155 bytes per cm^2). This clearly compares unfavourably with LDI-MS methods that can potentially deposit each metabolite mixture in spots falling in the μm^2 range attaining KB per cm^2 densities.

Writing, encoding, decoding and processing chemical mixtures

The name of our institution, "Universitat Jaume I", was encoded into 152-bit data (see CODE1_LCMS in ESI†). Each bit was assigned to a unique metabolite with a specific *m*/*z* value and retention time. Encoding was performed on a single vial and decoding *via* LC-MS was used for the retrieval of the original message. Given the limited analyte coverage of FIA-MS methods, spatial ordering was necessary to encode the message (see CODE1_FIAMS in ESI†). A set of 32 metabolites was utilized, arranged spatially in five distinct locations (or vials) on a sample plate for decoding *via* FIA-MS.

Both the metabolite encoding and the specific positioning of the vials on the sample plate determine the reconstruction of the information in the correct order. Fig. 2 illustrates the binary code associated with "Universitat Jaume I" and the spatial ordering for decoding *via* FIA-MS. Once data were acquired, various commercially available data processing tools were employed. Chrotool, an application integrated within MassLynx 4.2 was particularly useful for rapid data visualization. It enabled the grouping of eight-metabolite subsets (see Fig. 2) according to the sequence defined during the encoding process. The extracted ion chromatograms (XICs) for each eight-metabolite group (bytes) were then automatically displayed in a "one-click" operation, facilitating data visualization. As shown in Fig. 2, strong signals were observed for metabolites encoded in vial 1 at positions 2, 4, 6, 8, 10, 11, 13,

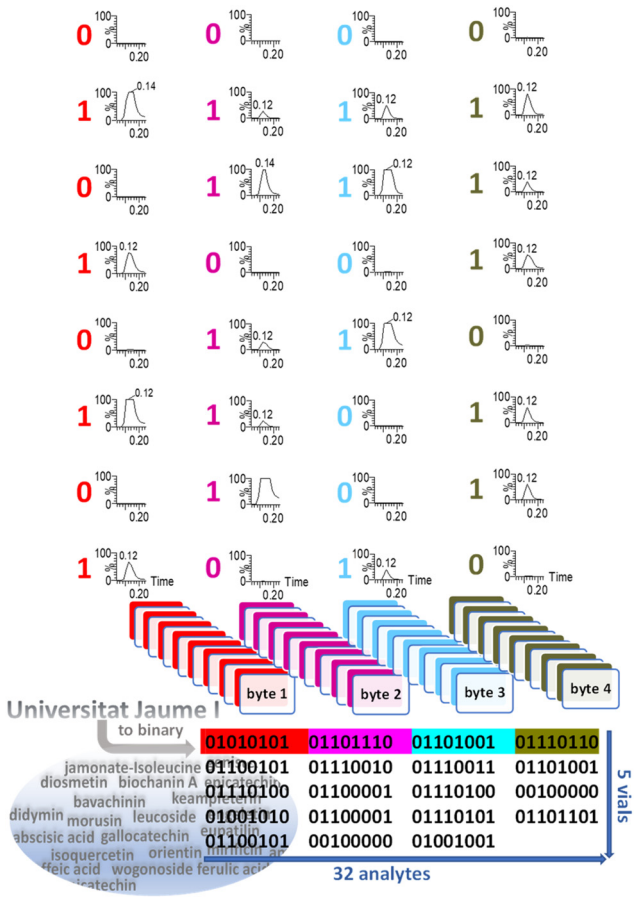

Fig. 2 Binary code corresponding to "Universitat Jaume I" encoded with set of 32 metabolites arranged spatially in 5 distinct locations along with a schematic illustration of the eight bits-grouping performed with Chrotool (Masslynx 4.2). The extracted ion chromatograms (XIC's), obtained with a single "one-click" operation, of each encoding metabolite representing the bytes 1–4 in vial 1, are shown. A consistent threshold of 5×10^5 ion counts was applied to normalize all analytes to the same scale. Note that, under this representation, the apex of some prominent peaks exceeds the range and appears truncated.

14, 15, 17, 18, 20, 24, 26, 27, 28, 30, and 31 within the 32-metabolite sublibrary. This arrangement encoded the following four bytes: 01010101 01101110 01101001 01110110.

Larger texts can be readily encoded/decoded using our approach. A paragraph of the Nobel lecture of J. J. Thomson of 1908 (1112 bits) have been also encoded and decoded successfully (see CODE2_LCMS in ESI†). Due to the limited size of the encoded abstract information (approximately 1000 bits), conventional analyses of error rates and digital error correction are not directly applicable or informative for the present datasets. It is important to note that the peak area associated with the presence of a metabolite was two orders of magnitude higher than the noise level observed in the absence of that metabolite, which (i) eliminated the need for adjusting metabolite concentration ratios during encoding to achieve uniform intensities,[18] (ii) ensured error-free decoding (iii) allowed for reliable automation of the process and (iv) could be further optimized to use more diluted samples (and increase the density expressed in bytes per g) or enlarge the subset of compounds per mixture.

We established an automated and general procedure for recovering the stored information, utilizing TargetLynx, an application integrated within MassLynx 4.2, which automates data processing and reporting. This tool incorporates robust confirmatory checks to identify analytes based on our previous empirical mass spectra library as well as user-defined intensity thresholds. The TargetLynx workflow used to recover the original message is shown in Fig. S1.† TargetLynx allows for the extraction of XICs for each metabolite, along with the calculation of their integrated areas, which can then be saved as a comprehensive summary in a .txt file. A Python-based code was developed in which the identity of each metabolite, as defined during encoding, and its integrated peak area was considered to determine the presence or absence of each metabolite (the detailed scripts are appended in the ESI†). When a metabolite is present, the corresponding bit was assigned a value of "1" (and "0" if absent). The binary values were then concatenated, and the final binary code was converted into the original ASCII message.

Encoding with structurally-related metabolites, isobaric metabolites and regioisomers

Our preliminary results from decoding chemical mixtures using FIA-MS revealed several inherent limitations when using flavonoids and plant hormones (findings that are broadly applicable to other compound classes) for encoding and decoding molecular information: (i) Flavonoid-to-flavonoid interferences: some flavonoids are derived from extensive conjugation, including glycosides (both C- and O-glycosylated flavonoids are commonly found in nature) of smaller flavonoid counterparts. Such conjugation often represents the more labile portion of the metabolite, which is susceptible to cleavage under electrospray ionization (ESI) conditions. Consequently, an analyte of interest may undergo in-source fragmentation, resulting in the formation of the pseudo-molecular $[M - H]^-$ ion of another analyte. This phenomenon is illustrated in Fig. 3(a), where a metabolite mixture containing isoquercetin (m/z 463.0896) is analysed in the absence of quercetin (m/z 301.035).

The presence of isoquercetin is evident in the extracted ion chromatogram (XIC) shown in Fig. 3(b) (bottom), where a minor XIC associated with the ion at m/z 301.0348 (see Fig. 3(b) top), formed by in-source fragmentation of isoquercetin, is also observed. Such in-source dissociation occurs at a low extent (less than 5% relative to the XIC of isoquercetin) and does not result in the detection of false positives (presence of quercetin when indeed it is absent) using our general processing workflow (see ESI†). However, as will be discussed below, such metabolite interference constitutes a major hurdle for quantitative analysis. (ii) Metabolites with the same nominal mass: due to the absence of chromatographic separation, FIA is inherently less selective than LC-MS techniques. In particular, when using low-resolution MS instruments (see

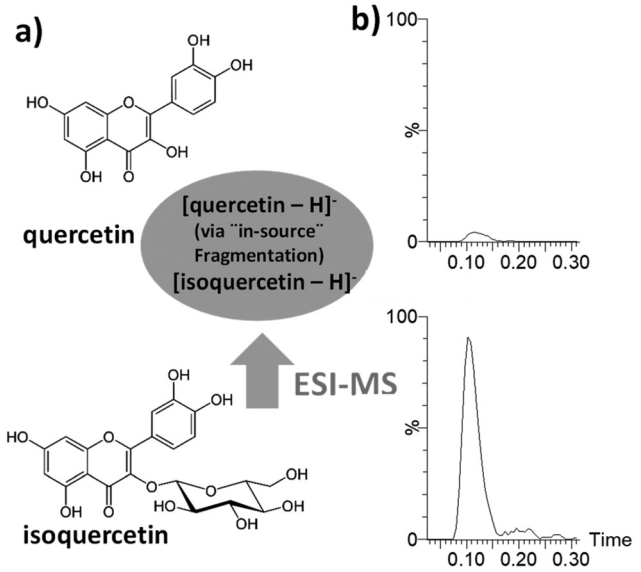

Fig. 3 (a) illustration showing that under ESI conditions, the formation of the [isoquercetin – H]$^-$ ion is accompanied by minor formation of the [quercetin – H]$^-$ ion *via* "in-source" fragmentation; (b) the XIC of the [M – H]$^-$ ion of isoquercetin is displayed at the bottom, while the XIC of the ion at *m*/*z* 301.035 (formally associated to the [quercetin – H]$^-$ species), produced by "in-source" fragmentation, is shown at the top.

Experimental section), decoding messages encoded with metabolites of the same nominal mass becomes problematic and it is critical to select an encoding metabolite library that avoids isobaric analytes. Greater flexibility in choosing metabolites can be achieved by coupling FIA with high-resolution MS (FIA-HRMS), which allows for more precise metabolite identification through accurate *m*/*z* determination. This advantage is exemplified in the successful encoding/decoding of the text "Universitat Jaume I", where alpinetin ($C_{16}H_{14}O_4$; [M – H]$^-$ *m*/*z* 269.0814) and genistein ($C_{15}H_{10}O_5$; [M – H]$^-$ *m*/*z* 269.0450) were used from the 32-metabolite subset; (iii) Interference between regioisomers: identifying regioisomers using FIA-MS poses a significant challenge due to the lack of separation before MS analysis. In the case of flavonoids, regioisomers often share identical fragmentation patterns and are difficult to differentiate based solely on the relative branching ratios of their product ions under collision-induced dissociation (CID) conditions. To overcome this challenge, ion mobility spectrometry mass spectrometry (IMS-MS) can provide a solution, as it allows for the distinction of regioisomers based on the shape and size of their associated gas-phase ions. The regioisomers genistein, baicalein, and galangin ($C_{15}H_{10}O_5$) were investigated using IMS-MS. Fig. 4(a) displays their negative ESI mass spectra, while Fig. 4(b) shows the corresponding mobility traces for their [M – H]$^-$ ion at *m*/*z* 269.0501. The ability to differentiate these three regioisomers based on their distinct drift times demonstrates the feasibility of utilizing regioisomers as encoding analytes, provided their identification relies on both *m*/*z* and drift time measurements.

Extending storage capacity through quantitation of chemical mixtures

Encoding data using more than two concentration levels enhances the complexity of the readout process and increases the maximum attainable storage capacity.[38] Such complexity arises in part due to the increased number of analytes and the potential ion suppression effects among them. Both FIA-MS and LC-MS display reliable and operationally simple quantitation capabilities, partly due to advancements in mass spectrometry (MS) instrumentation, which typically offers up to three orders of concentration linear range.[34,39] We investigated potential ion suppression effects for three representative families of the compound investigated, namely myricetin (flavone skeleton), didymin (conjugated flavonoid), and abscisic acid (plant hormone), in the presence of a reduced subset of 25 flavonoids as the matrix. Detailed results are provided in the ESI.† No significant ion suppression was observed when using LC-MS as the decoding method. When each compound was analysed both in neat solvent and within the 25-flavonoid mixture at two concentration matrix levels (75 and 750 ppb), the resulting calibration curves were virtually identical. Accordingly, datasets encoded using both quaternary and octal encoding schemes were successfully decoded with 100% accuracy using LC-MS (see below). In contrast, FIA-MS exhibited a greater potential for ion suppression due to the number of flavonoids analysed simultaneously and the use of short, non-resolving chromatographic runs. Ion suppression was evidenced as indicated by a decrease in the slope of the calibration curves for the three representative compounds, namely myricetin, didymin and abscisic acid in the presence of the 25-flavonoid matrix compared to those obtained in neat solvent (see ESI†). Nevertheless, this suppression did not substantially affect the accurate quantification of a reduced 25 subset of flavonoids (see below) across four predefined and widely spaced concentration levels (0, 100, 500, and 1250 ppb) employed in the quaternary encoding system. However, when more than eight concentration levels were employed, an increase in classification errors led to a higher incidence of false positives, thereby reducing the overall readout accuracy in FIA-MS analysis. This outcome is expected, as denser concentration gradations increase the likelihood of misclassification. Eventually, the use of internal standards would correct such deviations but can be impractical for large metabolite sets due to the lack of appropriate (most often isotopically enriched homologues) flavonoid internal standards.[40]

The increase in storage capacity achieved through quantitation follows a logarithmic (log 2) function, meaning that analyte determination at four or eight levels can result in a two- or three-fold increase in capacity, respectively. This enables the use of higher quaternary or octal encoding systems. The quaternary (base-4) numeral system, which employs the digits 0, 1, 2, and 3, can represent any real number associated with four concentration levels. Interestingly, the concept of quaternary encoding was inspired by the genetic code of DNA, where digital information can be

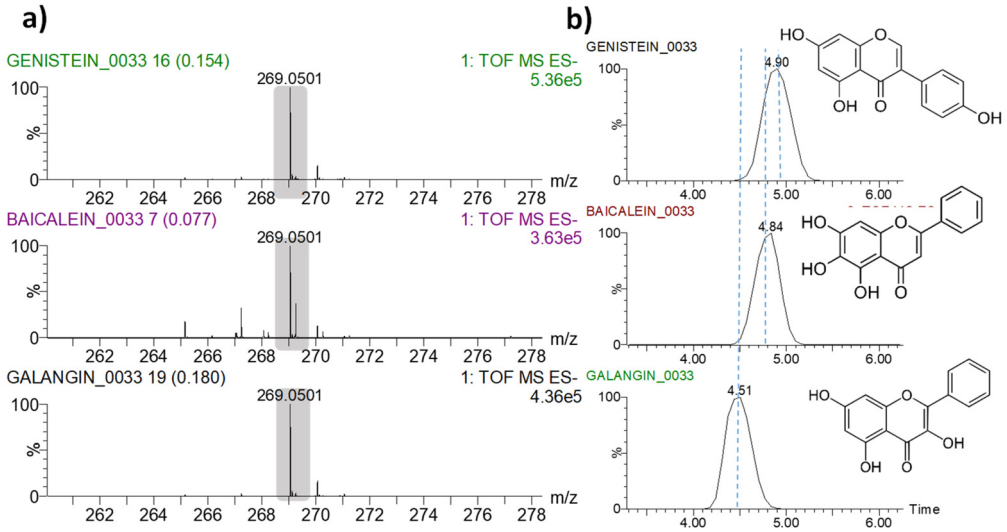

Fig. 4 (a) ESI mass spectra of the regioisomers genistein, baicalein and galangin showing the [M − H]- species as the base peak; (b) ion mobility traces of the [M − H]- species observed at *m/z* 269.0501.

encoded into sequences of A, T, C, and G. A binary number can be converted into quaternary form, where each digit corresponds to a binary pair (00, 01, 10, 11) converted to 0, 1, 2, and 3, respectively. This approach significantly reduces the overall task of implementation (reducing a half both the encoding and decoding steps) and reducing the overall latency of FIA-MS and LC-MS methods. Similarly, the octal (base-8) numeral system, using digits 0 to 7, operates by converting each binary digit triplet (000, 001, 010, 011, 100, 101, 110, 111) into the corresponding octal values. The use of both the quaternary and octal encoding systems also has implications for high-level information encryption. For example, with eight possible concentration levels for each component, half a dozen metabolites can generate approximately 260 000 combinations, each representing a different symbol. This creates the opportunity for developing a degenerate encoding key, where a given symbol is represented by multiple combinations. Such degenerate keys enhance the reliability, flexibility, and security of systems requiring robustness, such as in digital communications and cryptographic applications.[41]

In our experiments, we encoded texts and multicolour images using the identification and quantitation of chemical mixtures at different concentration levels. The logo of our institution, represented as a 25 × 10 grid, was encoded in quaternary code using 10 sets of 25 metabolites (see CODE3_QUAT_FIAMS and CODE4_QUAT_LCMS in ESI†). Four concentration ranges of metabolites were employed to define the 0–3 values in the quaternary encoding system. For each of the 10 vials, metabolites intended for inclusion were manually transferred from their stock solutions. Calibration mixtures were prepared from a single standard mixture of the 25 flavonoids at 1000 ppb by subsequent dilution covering the desired concentration ranges. Information retrieval using both FIA-MS and LC-MS was achieved with 100% efficiency, and the decoded image is shown in Fig. 5(a). To recover the original message, a Python script was utilized, where four threshold values were established to assign the 0–3 values. These values were then concatenated, and the resulting quaternary code was converted back into the original image.

The word "flavonoids" was converted into 80-bit data and stored in a single vial using a subset of 80 metabolites. Decoding *via* FIA-MS was conducted with a 100% readout accuracy (see CODE4_FIAMS in ESI†). To demonstrate the effectiveness of the shortening implementation, consider the binary encoding vector associated with the word "flavonoids".

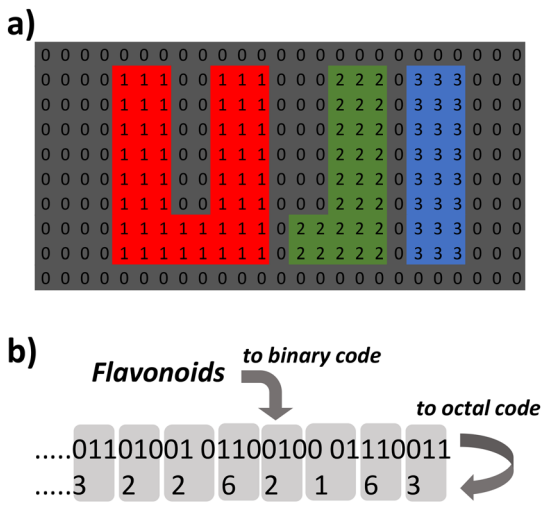

Fig. 5 (a) Encoding of the logo of our institution in quaternary code using 10 sets of 25 metabolites; (b) encoding of "flavonoids" to binary code and then to octal code. Each 000, 100, 010, 010, 011, 100, 101, 110, 111 subsequence in the binary code was transformed to 0, 1, 2, 3, 4, 5, 6 and 7 starting from the right side of the binary encoding vector.

As illustrated in Fig. 5(b) each 24-bit segment (representing 24 different analytes) can be represented by a sequence of just eight metabolites by using eight distinct concentration levels, thus achieving the same data storage capacity. Accordingly, a single vial containing 27 metabolites at eight concentration levels was used for octal encoding and satisfactory LC-MS decoding, replacing the need for 80 metabolites (see CODE5_OCTAL_LCMS in ESI†). To recover the original message, a Python code adapted from the binary data retrieval process (see ESI†) was used. Specifically, eight threshold values were defined to assign values between 0 and 7, which were then concatenated. The octal code was subsequently converted into binary and finally the original text was retrieved.

Conclusions

This study introduces two innovative hyphenated mass spectrometry (MS) methods, namely LC-MS and FIA-MS, for molecular data storage within complex chemical mixtures. These methodologies offer distinct advantages: (i) they provide an alternative to standalone MS analysis, overcoming limitations in analyte coverage per acquisition and thereby enhancing the attainable storage capacity per mixture; (ii) besides ensuring high analyte coverage and moderate injection throughput, they maintain operational simplicity and rely solely on commercially available materials and instrumentation; (iii) they allow seamless switching between both FIA-MS and LC-MS methods using the same equipment within minutes and (iv) the implementation of higher quantitation levels, up to eight, allow for a three-fold storage capacity enhancement that is used here to shorten both the encoding and decoding steps.

The FIA-MS approach achieves maximum storage capacities of 80–100 bits per mixture. Due to the lack of chromatographic separation, metabolite identification in FIA-MS relies solely on m/z values, which narrows the scope of the metabolite list that can be included, especially when utilizing low-resolution FIA-MS. Wider encoding analyte scope can be achieved using high-resolution (HR) MS and ion mobility spectrometry (IMS) MS capabilities. HR-MS enables the discrimination of isobaric compounds, particularly within metabolomic compound classes whereas IMS-MS introduces a rapid, millisecond-scale, post-ionization separation dimension, enhancing the ability to distinguish regioisomers through a combination of m/z and drift time. LC-MS methods, on the other hand, combine the broadest metabolite coverage in comparison to other techniques and a wide flexibility at choosing the encoding library metabolites, as each metabolite is uniquely defined by both m/z value and retention time. The maximum storage capacity for LC-MS in routine metabolomic analysis, demonstrated here at 200 bits per mixture, has the potential to reach 10^4 bits per mixture, representing the highest data storage capacity achieved in molecular data systems based on single chemical mixtures.

The expansion of data storage capacity per mixture as well as the encoding and decoding latency could be significantly optimized by integrating additional libraries, adopting faster robotic-based encoding methods, or leveraging advanced metabolomic data-processing tools. Although the present demonstrations focused on libraries of plant hormones and flavonoids, the approach is adaptable to other metabolite families, provided they are amenable to the electrospray ionization (ESI) process. This compatibility extends the method's applicability to a wide range of metabolites, including small molecules that may pose analytical challenges for laser desorption ionization (LDI) techniques due to matrix interference in low m/z ranges. In this context, lipids, which constitute approximately three-quarters of the human metabolome database (HMDB 4.0, **https://www.hmdb.ca**), encompassing around 5×10^4 lipid species, offer substantial potential for constructing large capacity devices based on chemical mixtures. Furthermore, automated synthesis of expansive synthetic libraries *via* multi-component reactions, coupled with robotic liquid dispensing into vials, can increase encoding efficiency. As shown in this work, metabolomics provides an optimized and user-friendly analytical environment, largely due to its extensive history of software development, allowing for rapid data processing with minimal adjustments. Additionally, other widely used hyphenated techniques in metabolomics, such as gas chromatography-MS and capillary electrophoresis-MS, could also be adapted for high-capacity chemical mixture-based memory devices. Integrating foundational principles from metabolomics into synthetic chemical mixtures can offer novel avenues at the intersection of information technology, plant physiology, analytical chemistry, and chemical sciences.

Author contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Data availability

Detailed experimental procedures and instrumentation was well as codes, output quali- and quantitation results and python scripts for message retrieval are given as ESI† in *CODES_SI.zip*; .raw data are deposited in **https://doi.org/10.5281/zenodo.15101114**; the complete scripts for identification and quantitation using binary, quaternary and octal schemes are available at **https://github.com/catm542-ai/ChemDataProcessor**.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

PID2021_1248130B-C32 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", by the European Union. C. T. thanks the Generalitat Valenciana (CIGRIS/2021/075) for funding. Dr Pere Clemente is acknowledged for his assistance using MATLAB. The authors thank 'Servei Central d'Instrumentació Científica (SCIC) of the Universitat Jaume I'.

References

1 A. Anžel, D. Heider and G. Hattab, *Comput. Struct. Biotechnol. J.*, 2021, **19**, 4904–4918.
2 M. Soete, C. Mertens, N. Badi and F. E. Du Prez, *J. Am. Chem. Soc.*, 2022, **144**, 22378–22390.
3 L. Ceze, J. Nivala and K. Strauss, *Nat. Rev. Genet.*, 2019, **20**, 456–466.
4 M. G. T. A. Rutten, F. W. Vaandrager, J. A. A. W. Elemans and R. J. M. Nolte, *Nat. Rev. Chem.*, 2018, **2**, 365–381.
5 C. K. Lim, S. Nirantar, W. S. Yew and C. L. Poh, *Trends Biotechnol.*, 2021, **39**, 990–1003.
6 Y. Erlich and D. Zielinski, *Science*, 2017, **355**, 950–954.
7 T. Schutz, I. Sergent, G. Obeid, L. Oswald, A. Al Ouahabi, P. N. W. Baxter, J. L. Clément, D. Gigmes, L. Charles and J. F. Lutz, *Angew. Chem., Int. Ed.*, 2023, **62**, e202310801.
8 S. Martens, A. Landuyt, P. Espeel, B. Devreese, P. Dawyndt and F. Du Prez, *Nat. Commun.*, 2018, **9**, 4451.
9 M. Frölich, D. Hofheinz and M. A. R. Meier, *Commun. Chem.*, 2020, **3**, 1–10.
10 N. F. König, A. Al Ouahabi, S. Poyer, L. Charles and J. F. Lutz, *Angew. Chem., Int. Ed.*, 2017, **56**, 7297–7301.
11 M. Soete and F. E. Du Prez, *Angew. Chem., Int. Ed.*, 2022, **61**, e20220281.
12 M. Soete, K. De Bruycker and F. Du Prez, *Angew. Chem., Int. Ed.*, 2022, **61**, e202116718.
13 L. Zhang, T. B. Krause, H. Deol, B. Pandey, Q. Xiao, H. M. Park, B. L. Iverson, D. Law and E. V. Anslyn, *Chem. Sci.*, 2024, **15**, 5284–5293.
14 B. Mauch-Mani, I. Baccelli, E. Luna and V. Flors, *Annu. Rev. Plant Biol.*, 2017, **68**, 485–512.
15 A. R. Fernie, S. de Vries and J. de Vries, *Philos. Trans. R. Soc., B*, 2024, **379**, 20230347.
16 B. J. Cafferty, A. S. Ten, M. J. Fink, S. Morey, D. J. Preston, M. Mrksich and G. M. Whitesides, *ACS Cent. Sci.*, 2019, **5**, 911–916.
17 E. Kennedy, C. E. Arcadia, J. Geiser, P. M. Weber, C. Rose, B. M. Rubenstein and J. K. Rosenstein, *PLoS One*, 2019, **14**, 1–12.
18 V. Pardi-Tóth, Á. Kuki, M.Á Kordován, G. Róth, L. Nagy, M. Zsuga, T. Nagy and S. Kéki, *Sci. Rep.*, 2023, **13**, 1–7.
19 P. Sőregi, M. Zwillinger, L. Vágó, M. Csékei and A. Kotschy, *Chem. Sci.*, 2024, **15**, 14938–14945.
20 S. Gumus, D. Biechele-Speziale, K. E. Manz, K. D. Pennell, B. M. Rubenstein and J. K. Rosenstein, *ACS Omega*, 2024, **9**, 19904–19910.
21 C. E. Arcadia, E. Kennedy, J. Geiser, A. Dombroski, K. Oakley, S. L. Chen, L. Sprague, M. Ozmen, J. Sello, P. M. Weber, S. Reda, C. Rose, E. Kim, B. M. Rubenstein and J. K. Rosenstein, *Nat. Commun.*, 2020, **11**, 691.
22 E. Kennedy, J. Geiser, C. E. Arcadia, P. M. Weber, C. Rose, B. M. Rubenstein and J. K. Rosenstein, *Sci. Rep.*, 2021, **11**, 1–10.
23 P. Bohn, M. P. Weisel, J. Wolfs and M. A. R. Meier, *Sci. Rep.*, 2022, **12**, 1–8.
24 A. A. Nagarkar, S. E. Root, M. J. Fink, A. S. Ten, B. J. Cafferty, D. S. Richardson, M. Mrksich and G. M. Whitesides, *ACS Cent. Sci.*, 2021, **7**, 1728–1735.
25 Y. Tang, C. He, X. Zheng, X. Chen and T. Gao, *Chem. Sci.*, 2020, **11**, 3096–3103.
26 T. Ratner, O. Reany and E. Keinan, *ChemPhysChem*, 2009, **10**, 3303–3309.
27 J. M. Lee, H. Jang, S. W. Lee and K. T. Kim, *JACS Au*, 2022, **2**, 2108–2118.
28 R. Tautenhahn, G. J. Patti, D. Rinehart and G. Siuzdak, *Anal. Chem.*, 2012, **84**, 5035–5039.
29 T. Fuhrer and N. Zamboni, *Curr. Opin. Biotechnol.*, 2015, **31**, 73–78.
30 W. S. Maaty and D. D. Weis, *J. Am. Chem. Soc.*, 2016, **138**, 1335–1343.
31 D. P. Zolg, M. Wilhelm, T. Schmidt, G. Médard, J. Zerweck, T. Knaute, H. Wenschuh, U. Reimer, K. Schnatbaum and B. Kuster, *Mol. Cell. Proteomics*, 2018, **17**, 1850–1863.
32 M. Trojanowicz and K. Kołacińska, *Analyst*, 2016, **141**, 2085–2139.
33 T. Fuhrer, D. Heer, B. Begemann and N. Zamboni, *Anal. Chem.*, 2011, **83**, 7074–7080.
34 S. C. Nanita and L. G. Kaldon, *Anal. Bioanal. Chem.*, 2016, **408**, 23–33.
35 A. De Villiers, P. Venter and H. Pasch, *J. Chromatogr. A*, 2015, **1430**, 16–78.
36 G. Madalinski, E. Godat, S. Alves, D. Lesage, E. Genin, P. Levi, J. Labarre, J.-C. Tabet, E. Ezan and C. Junot, *Anal. Chem.*, 2008, **80**, 3291–3303.
37 C. G. Enke, *Anal. Chem.*, 1997, **69**, 4885–4893.
38 J. K. Rosenstein, C. Rose, S. Reda, P. M. Weber, E. Kim, J. Sello, J. Geiser, E. Kennedy, C. Arcadia, A. Dombroski, K. Oakley, S. L. Chen, H. Tann and B. M. Rubenstein, *IEEE Trans. Nanobiosci.*, 2020, **19**, 378–384.
39 C. Xie, D. Zhong, K. Yu and X. Chen, *Bioanalysis*, 2012, **4**, 937–959.
40 A. Nasiri, R. Jahani, S. Mokhtari, H. Yazdanpanah, B. Daraei, M. Faizi and F. Kobarfard, *Analyst*, 2021, **146**, 6049–6063.
41 T. Sarkar, K. Selvakumar, L. Motiei and D. Margulies, *Nat. Commun.*, 2016, **7**, 11374.