ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Repurposing Waste Chemicals for Sustainable and Durable Molecular Data Storage

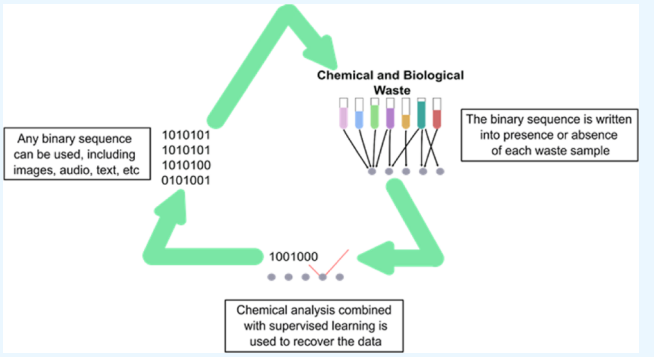
Selahaddin Gumus, Dana Biechele-Speziale, Katherine E. Manz, Kurt D. Pennell, Brenda M. Rubenstein, and Jacob K. Rosenstein*

Cite This: *ACS Omega* 2024, 9, 19904−19910

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular data storage offers the intriguing possibility of higher theoretical density and longer lifetimes than today's electronic memory devices. Some demonstrations have used deoxyribonucleic acid (DNA), but bottlenecks in nucleic acid synthesis continue to make DNA data storage orders of magnitude more expensive than electronic storage media. Additionally, despite its potential for long-term storage, DNA faces durability challenges from environmental degradation. In this work, we demonstrate nongenomic molecular data storage using molecular libraries redirected from chemical waste streams. This approach requires no synthetic effort and can be implemented by using molecules that have a minimal associated cost. While the technique is agnostic about the exact molecular content of its inputs, we confirmed that some sources contained poly fluoroalkyl substances (PFAS), which persist for long periods in the natural environment and could offer extremely durable information storage as well as environmental benefits. These demonstrations provide a perspective on some of the valuable possibilities for nongenomic molecular information systems.



INTRODUCTION

Molecular data systems have been proposed as a path to meet some of the world's ever-growing demands for information storage while reaching greater information density, durability, and sustainability than electromagnetic memory. However, existing molecular storage methods are limited in information capacity by practical scaling challenges associated with the cost and complexity of their chemical synthesis. In this work, we demonstrate that data storage can be achieved with chemicals requiring no explicit synthesis, purification, or prior knowledge of the molecular structure. We encoded digital data using chemical waste sources containing mixtures of unknown byproducts. This work highlights a path for molecular data storage to utilize low-cost chemical waste while avoiding complicated synthesis or purification. In order for any chemical information system to succeed and scale, it will need to overcome the synthesis, purification, and cost bottlenecks that are associated with all previously reported demonstrations, spanning from DNA[1,2] to synthetic polymers[3] to several families of small molecules.[4−6]

Discussions of molecular information on them begin with deoxyribonucleic acid (DNA), whose role in biology makes it a natural candidate for synthetic molecular data systems.[1] However, although the chemical synthesis of DNA has been used and improved upon for decades, DNA oligomer synthesis and purification are still material-intensive,[7] resulting in low yields and producing significant waste. In part, these challenges are inherent to the serial nature of long polymers. For example, even if the yield of a single base incorporation was 99%, the yield of a 150-mer would be only 22%. To reduce such losses, DNA synthesis chemistries commonly use 20× excess of phosphoramidites and tetrazole-based activators, as well as an excess of solvents to wash between each serial synthesis step.[7] Material wastes from DNA synthesis include the protecting group, dimethoxytrityl, which is 35% of the weight of each phosphoramidite; tetrazole, which is explosive; acetonitrile; amidites; pyridine; among others.[8] In addition to costing money and time, many of these inputs and intermediates that do not make it into final products end up in waste streams, with associated environmental and economic impacts. A potential transition to enzymatic DNA synthesis could mitigate some of these costs if new advances can achieve competitive error rates, reliability, and throughput.[2]

Another theoretically valuable feature of DNA is its durability. However, many environmental factors can cause

Received: November 19, 2023
Revised: March 31, 2024
Accepted: April 15, 2024
Published: April 25, 2024

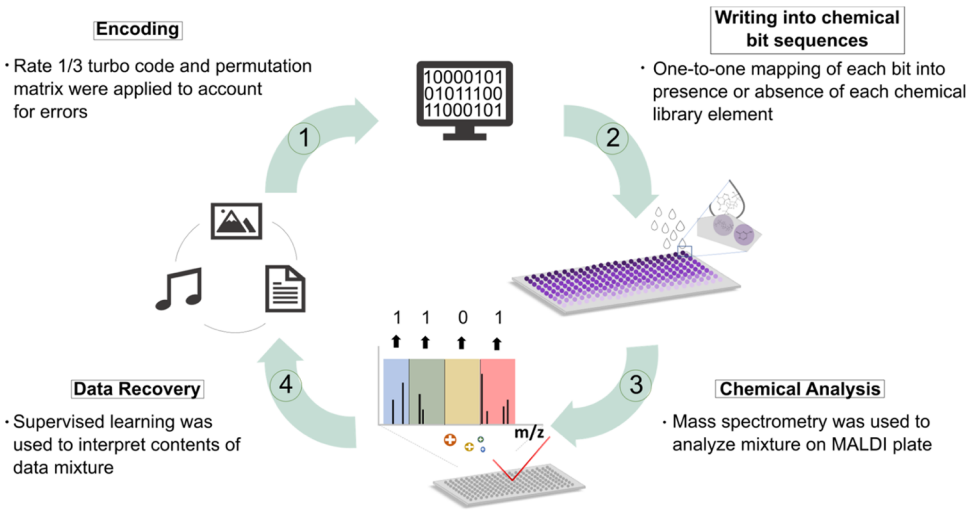


Figure 1. Visualization of the overall process of encoding, writing, analysis, and recovery of data in molecular mixtures. Step 1 involves the encoding of data, including the Turbo error correction bits and the application of a permutation matrix to disperse errors. Step 2 shows the mapping of the encoded data to chemical library elements. Step 3 demonstrates the utilization of chemical analysis techniques to identify the chemical contents of each spot. Step 4 involves the use of supervised learning to interpret the chemical contents and recover the data.

DNA degradation or reactivity, which could result in information loss. These factors include the pH, ionizing radiation and ultraviolet light exposure, the salt concentration of the medium, the presence of nucleases, and hydrolysis, among others.[9,10] Hydrolysis is the dominant mechanism for instability and can lead to information loss through depurination, deamination, and backbone cleavage. To maintain the integrity of data stored in DNA, it needs to be maintained in carefully controlled environments and dried, frozen, or encapsulated. These treatments limit potential commercial markets and can reduce the information density by orders of magnitude.[11]

Some alternative research directions have explored non-genomic molecules for information storage. For example, alternative sequence-defined polymers offer different synthetic paths and a larger library of monomers, which allow for a greater information density using the same polymer length.[12] (However, the upper limit of de novo sequencing of digitally encoded polymers is currently on the order of 77 units.[13]) Demonstrations of synthetic polymers have used oligo(triazole amides),[14] oligourethanes,[15] poly(phosphodiester)s,[16] poly(alkoxyamine amide)s,[17] and poly(l-lactic-*co*-glycolic acid)s (PLGAs),[18] among others.

Small nonpolymeric molecules can also be used to encode digital information. Previously, we demonstrated the use of synthetic metabolite mixtures to store digital information,[19] showing that digital data could be encoded in the presence or absence of vitamins, nucleotides, amino acids, sugars, and other small molecules. A similar encoding scheme was used by other groups to encode digital data using mixtures of oligopeptides[5] and fluorescent dyes.[4] Further extending these methods, we encoded larger digital data sets in organic small molecules using combinatorial chemistry with multicomponent Ugi products,[6] and applied sparse encoding schemes for redundancy and improved error tolerance,[20] and worked toward secret messaging on common objects.[21] While these studies have established the concept of small molecule-based data storage, they usually still require the synthesis of a starting chemical library, which can be time-consuming or cost-prohibitive for larger data sets.

In this work, we demonstrate that molecular data storage can be achieved without any synthesis or purification steps and that data can even be stored using compounds that are available from waste sources at no material cost. This is a general approach that can work with sources whose exact chemical makeup is unknown to the user. An overview of the data storage process is shown in Figure 1, highlighting the ease and scalability of the process. A digital image was encoded by using complex mixtures of unknown waste chemicals. These waste sources were organic synthesis byproducts and toxicology research byproducts, which were later shown to include per- and poly fluoroalkyl substances (PFAS), among other compounds. In addition to providing improved cost and resource efficiency, given appropriate chemical inputs, it can also encode data using highly durable compounds whose stability and lifetime can exceed that of DNA under ambient conditions.

RESULTS

Turning Chemical Waste into Digital Data. To offer a solution to the synthesis-related challenges and costs of molecular data storage, we set out to encode data using chemicals diverted from waste streams. Several waste sources were obtained from two academic laboratories: one developing organic synthesis methodology[22] and one studying the developmental effects of exposure to environmental contaminants.[23] In contrast with most other molecular data storage systems, prior knowledge of the chemical makeup of the source materials was not required for their use. The samples were diluted in dimethyl sulfoxide (DMSO), and a matrix-assisted laser desorption ionization (MALDI) matrix was added to the mixtures.

In previous work, we encoded information in small-molecule samples using the presence or absence of each library element to represent binary values of "1" and "0".[6] In this work, we used seven waste samples as library elements, and thus, each mixture could represent up to 7 bits of information. Since the waste samples themselves contained many compounds, the resulting encoded data was a set of "mixtures of mixtures". Figure 2 shows the process of encoding a 3249-pixel digital

https://doi.org/10.1021/acsomega.3c09234
ACS Omega 2024, 9, 19904−19910

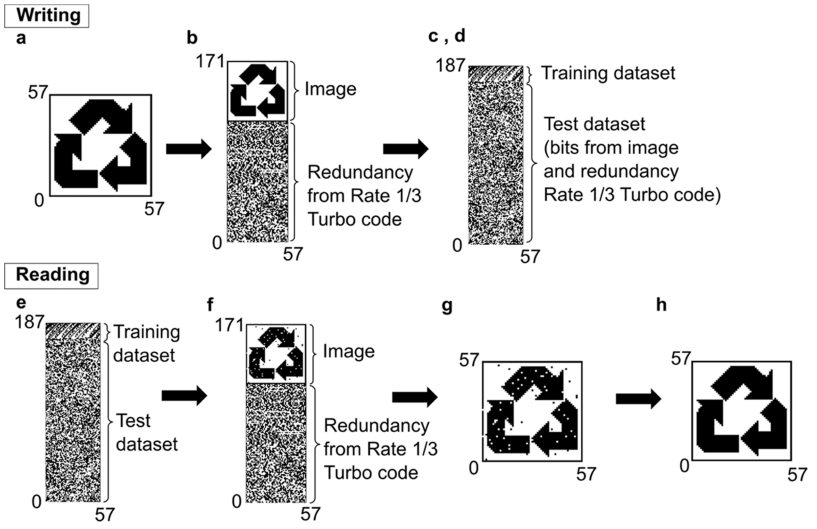


Figure 2. Writing and reading processes of the encoded data. (a) Encoded sustainability logo. (b) 1/3 Turbo error correction bits incorporated into the data set. (c) Permutation matrix was applied. (d) Training set was included at the top. (e) Raw data prediction after readout. (f) Data set was unpermuted. (g) Recovery of the encoded image before error correction. (h) Recovered image after error correction.

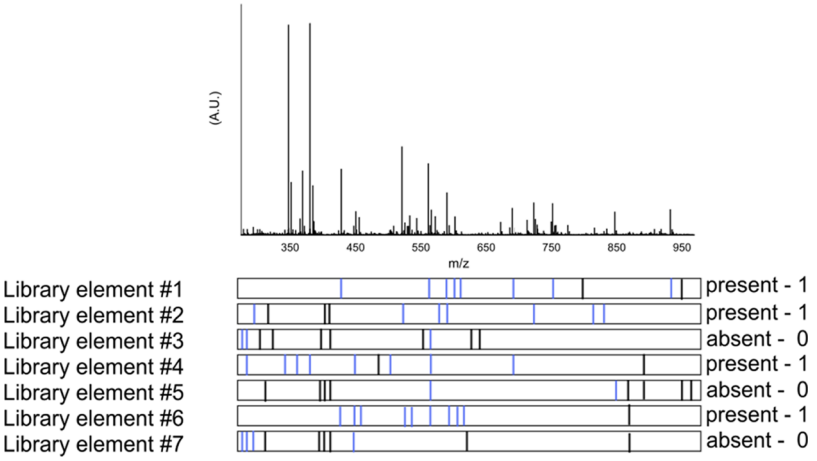


Figure 3. MALDI positive ionization mass spectrum for a sample that encodes the 7 bits of information "1101010". Below the spectrum is a visualization of 10 features that the classifier models have associated with the presence of each of the 7 library elements. A blue bar indicates the presence of the individual feature in this particular spectrum, and a black bar represents the absence of that feature.

image. To provide redundancy and support error correction, the digital image was processed with a Rate 1/3 Turbo code[24] (Figure 2b) and a permutation matrix (Figure 2c). To assist with the recovery of the data, a training set of all $2^7 = 128$ enumerated binary mixtures was concatenated with the encoded data set (Figure 2d). The whole data set, including the training and test data, was rearranged as a 1521×7 matrix, where each column corresponded to a unique library element (one of the seven waste mixtures). Each row of the matrix was mapped to a unique location on a standard metal MALDI target data plate. For each element of the matrix, if the value was 1, an acoustic liquid handler was instructed to dispense 30 nL of the appropriate waste mixture to the designated location. If the value was 0, then no transfer was performed. This process resulted in 1521 mixtures of subsets of the 7 sources; in each mixture, the presence (1) or absence (0) of each waste source encoded the data. The mixtures were dried overnight, leaving crystalline spots behind. This process is scalable to thousands of unique mixtures per data plate.[6]

Reading Back Data from the Encoded Waste Mixtures. To recover the digital data from the chemical samples, each mixture on the data plate was analyzed with a Fourier transform ion cyclotron resonance (FT-ICR) MALDI mass spectrometer in positive ionization mode. Each spectrum contained a varying number of peaks based on the complexity of the mixture. The training set of $2^7 = 128$ enumerated combinations of the library elements was used to train supervised learning algorithms to identify the presence of each library element in each location on the data plate.

The analysis began by extracting candidate features from the training samples by averaging all of the training spectra and identifying the locations of the most significant peaks in the averaged spectrum. For each of the 7 library elements, this list of candidate peaks was ranked by discrimination power, using the area under each peak's receiver operating characteristics curve (AUC/ROC). The highest-ranked features were used to train supervised learning classifiers for each library element. The trained classifiers were then applied to the remaining data mixture spots and used to make soft decisions about the

https://doi.org/10.1021/acsomega.3c09234
ACS Omega 2024, 9, 19904−19910

presence or absence of each library element. The classifiers predicted whether each library element was present (1 atom) or absent ("0") in each data mixture. Figure 3 shows an example of one spectrum obtained during the analysis. This spectrum contains thousands of data points and hundreds of peaks. The classifiers associate each of the library elements with a "fingerprint" of some subset of spectral features that correlate with the presence of that library element in the training mixtures. In Figure 3, we annotate the top 10 features that the model associated with the presence of each library element. Four of the seven library elements are present, and this spectrum encodes the 7 bits of information "1101010".

For the supervised classifier models, we tested two common architectures, logistic regression and random forest regression, with varying levels of complexity. Since the data is encoded with redundancy and error tolerance, it is interesting to consider what level of complexity is required to achieve full recovery of the data payload (Figure 4). Seven classifier models

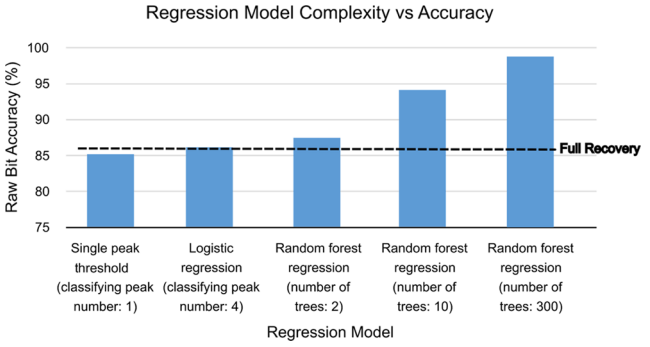


Figure 4. Comparison of logistic and random forest regression models for recovering encoded data. The raw bit accuracy was determined by comparing the binary representation of the encoded sustainability logo with that of the decoded logo prior to Turbo decoding. Since the data was encoded with redundancy and error tolerance, any model that exceeded 86% raw bit accuracy was sufficient to recover the data payload with 100% accuracy.

were trained, with one binary classifier for each library element. The parameters for each model are as follows: the single peak threshold used 1 classifying peak from the largest 50 features, the logistic regression model used 4 classifying peaks from 50 features, and the random forest regression models used 2, 10, and 300 trees with 1000, 1000, and 3000 features, respectively. Although the library mixtures are highly complex, even a simple logistic regression with 4 peaks achieved 86% raw accuracy, which was sufficient to achieve perfect recovery of the data payload after Turbo decoding (Figure 4). Based on simulations, the Rate 1/3 Turbo code was able to correct up to 14% raw bit error rates. Indeed, after Turbo decoding, we achieved 100% recovery of the encoded image (Figure 2h).

Recovery of Data. With the largest random forest regression model tested, the overall raw bit error rate was approximately 2%, but these errors were not uniformly distributed. As shown in Figure 5a, each of the chemical library elements was identified with more than 93% accuracy. However, the individual accuracies varied significantly; the total error from library element 7 was 0.07%, while the error from library element 6 was 6.39%. There were some spatial trends as well, with errors overrepresented near the left edge of the plate, where 2% of the mixtures were responsible for 17% of the total errors (Figure 5b). This could be attributed to mechanical errors or alignment errors in either the liquid handler or the mass spectrometer.

DISCUSSION

In this study, we demonstrated that molecular data storage can be achieved using inputs that do not require synthesis, purification, or even prior knowledge of their molecular contents. Using unpurified waste chemicals, we encoded a 3249-pixel digital image payload, plus redundancy for error correction (9751 raw bits including coding redundancy). Complete recovery of the data was successful despite not knowing the identity or concentration of the compounds within each sample. This construction is possible because training samples are included alongside the molecular data samples. With a training overhead of 8.41% (128 out of 1521 spots), we trained models that achieved a raw bit error rate as low as 2.04%, which was well below the 14% raw error rate tolerated by the Turbo code. There is always a trade-off between error correction capabilities and total payload capacity. Previous molecular data storage demonstrations have used Reed-Solomon[25,26] and fountain codes,[27] but here, we chose to use Turbo codes for their ability to tolerate potentially higher error rates from complex, unknown library mixtures.

To appreciate the contents of the waste input samples, the seven library elements were subsequently analyzed using nontargeted mass spectrometry analysis (Figure S1). The complexity of the waste sources varied significantly, with each sample containing between 13 and 109 identified compounds. Some of the waste inputs were quite similar to each other, containing up to 90% overlap in the lists of identified compounds. Conceptually, it should not matter how many similarities exist between the library elements as long as they

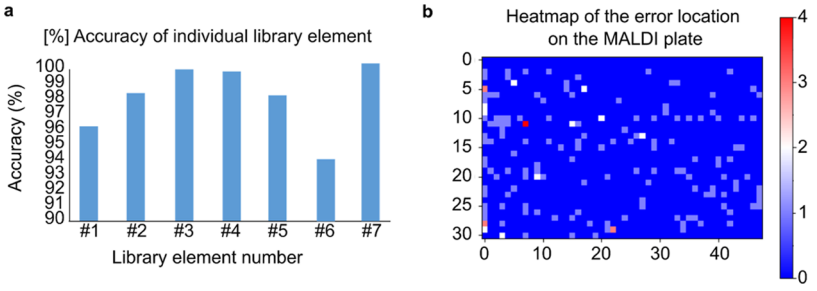


Figure 5. (a) Raw bit accuracy (before error correction) for each individual waste mixture and (b) visualization of the physical location of the raw bit errors on the MALDI plate. The color bar represents the number of errors on a blue-to-red scale, where blue represents fewer errors and red represents the highest number of errors.

https://doi.org/10.1021/acsomega.3c09234
ACS Omega 2024, 9, 19904−19910

also have identifiable differences. The ability to tolerate a wide range of complex unknown mixtures suggests possibilities for encoding data by using many different environmental and industrial inputs as library sources.

Two library samples contained several compounds belonging to the PFAS class, which are highly stable fluorinated hydrocarbon compounds that are persistent in the environment. As a class, PFAS are less chemically reactive and more stable than DNA under ambient conditions. DNA degradation can occur through several mechanisms,[11,28] and the rate of DNA degradation increases substantially with temperature or pH.[29] In contrast, many PFAS are stable up to 200−600 °C and certain PFAS also can tolerate very high pH before significant degradation.[30] PFAS are nicknamed "forever chemicals" owing to their long half-lives of up to 1000 years under ambient conditions.[31,32] The presence of PFAS in these samples suggests that library inputs that are partly or entirely composed of PFAS compounds could be chosen to produce data sets that are extremely stable and may retain their form for longer than DNA. Additionally, there could be value created by removing these contaminants from the environment and using them for long-term molecular data storage.

This study shows that digital information can be encoded in molecular mixtures without prior knowledge of their molecular contents and that the digital data can be fully recovered using chemical analysis. The absence of any explicit library synthesis steps is in contrast to previously demonstrated molecular data storage systems that are based on DNA, synthetic polymers, or small molecules. Despite advanced microarray synthesis platforms, DNA data storage still costs $0.001 per bit,[33] compared to $0.000000000001 per bit for modern electronic storage. In this work, we demonstrated data storage using molecules that had no associated cost. On a larger scale, there may be costs associated with the transportation and storage of chemical waste streams.

There are a wide range of different molecular information storage approaches, with trade-offs between density, capacity, complexity, and speed.[20] This work lies at one extreme, representing what it could mean to write molecular data without any library synthesis at all. This clearly creates some asymmetries, potentially making it more difficult to read back the data and limiting the information density to the limits of the chemical printing and chemical analysis tools used for writing and reading.

It is also important to recognize the limitations of encoding data by using waste streams as input sources. Although this approach lowers some material costs, writing data still incurs costs from liquid handling and has similar throughput and latency as other molecular data storage approaches. If scaled up, it would be important to maintain consistency of the source material streams, and some waste streams may be inappropriate for this approach if they are too dilute, acutely hazardous, complex, or similar to one another. Even if many of the molecules in a sample are very stable, some degradation is inevitable over time and some components in a given library element may cross react with components of another library element. Our approach of including every distinct combination of library elements in training samples and embedded error tolerance should allow for some degradation and potential cross reactivity to be naturally accommodated when it is time for the data to be read.

This demonstration should be taken as an illustration of one way that molecular data storage can be generalized to accommodate nearly any chemical input source, even those that have impurities or variable contents. This work highlights opportunities for molecular data storage to contribute positively to a circular economy while also providing low-cost and long-term durability.

METHODS

Materials. A total of seven waste chemical mixtures were used. Five of the waste mixtures were obtained from Dr. Ming Xian's lab in the Chemistry Department at Brown, and two of the biological waste samples were obtained from Dr. Jessica Plavicki's lab in the Pathology and Molecular Medicine Department at Brown University. All samples obtained were byproducts that were collected after their respective experiments had concluded and would have otherwise been deemed waste. Dimethyl sulfoxide (DMSO, anhydrous, 99.9%, MilliporeSigma) was used as a solvent to dilute all of the solutions in the library. Analytical grade α-cyano-4-hydroxycinnamic acid (HCCA, 99.0%, MilliporeSigma) was used as the matrix material for all MALDI samples.

Library Preparation. Each waste mixture was filtered through a 0.2-μm nylon syringe filter before being dissolved in DMSO. The waste chemicals from Dr. Xian's lab were dissolved in DMSO using a 1:2 ratio of DMSO to waste mixture. The waste chemicals from Prof. Plavicki's lab were dissolved in DMSO using a 3:1 ratio of DMSO to waste mixture. Each waste mixture was finally diluted in MALDI matrix solution (20 mg of HCCA in 1 mL of DMSO) using a 1:2 ratio of matrix solution to waste mixture. They were pipetted into a 384-well plate, with each row representing a distinct waste mixture.

Data Plate Preparation. The digital image of a sustainability logo (57 × 57 pixels) was encoded using a Rate 1/3 Turbo code, multiplied by a permutation matrix, concatenated with the training set, and converted into a one-dimensional binary vector. The training spots are a set of 2^7 = 128 enumerated binary mixtures, which is represented in the binary vector as 128 × 7 = 896 bits. The assembled one-dimensional (1-D) vector was reshaped into an $M \times N$ matrix, where M represents the number of waste mixtures and N represents the number of independent mixtures. For each true ("1") value in the matrix, 30 nL (12 2.5 nL droplets) of the mth waste mixture was transferred from the 384-well library plate to the nth location on a metal MALDI plate using an acoustic liquid handler (Echo 550, Beckman Coulter). This implies that each location was a mixture of up to 7 library elements and contained up to 210 nL. Once all transfers are complete, the data plate was left to dry in a fume hood overnight. The resulting dried mixture spots were typically 1 mm in diameter.

Mass Spectrometry. Mass spectra were acquired with a Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer in positive ion mode. To obtain accurate peak assignment, a mass calibration is performed through electrospray ionization (ESI) before each run using sodium trifluoroacetate as a reference. Samples were ionized by using matrix-assisted laser desorption ionization (MALDI). The sample run was automated, and the typical spectra were acquired for 1.5 s. During the automation process, the instrument serially addresses each crystallized spot and takes about 4 h to record all 1536 spots on a plate. Each measurement is made by ionizing a portion of a sample with

https://doi.org/10.1021/acsomega.3c09234
ACS Omega 2024, 9, 19904−19910

a laser configured to take 500 shots at 1000 Hz, over a scan area of 500−900 μm with a medium focus and ×4 averaging.

Nontargeted Analysis of the Chemical Library Elements. After the chemical data experiment was complete, we performed a separate nontargeted analysis of the waste samples. All instrument parameters, including the chromatography scheme, source settings, full scan parameters, and data-dependent (dd) MS2 settings, for both negative and positive ESI, are provided in detail (Table S1). Data was collected on a high-resolution Thermo QExactive HF-X Orbitrap MS equipped with a Vanquish ultrahigh-performance liquid chromatograph. We analyzed the samples two times: once with positive electrospray ionization (ESI, +) and once in negative ESI (ESI, −). For both ionization modes, we used the same chromatography scheme. Sample components were separated on a Thermo Hypersil Gold Vanquish C18 column ($100 \times 2.1 \times 1.9$ μm^3) with two mobile phases. Mobile phase A consisted of 2 mM ammonium acetate in 5% acetonitrile, and Mobile phase B consisted of 2 mM ammonium acetate in 100% acetonitrile. The total run time was 15 min. Data was acquired using data-dependent (dd) MS^2 acquisition. Fragmentation was performed in an HCD collision cell filled with N_2 (produced by a Peak Scientific Nitrogen Generator, Genius NM32LA). All spectral data files were saved in the. RAW file format, and NTA/SSA was performed in the Thermo Compound Discoverer (CD) 3.3 software. For ESI,+, and ESI,− data, peaks were detected with 10 ppm mass tolerance, 1,000,000 minimum peak intensity, and a signal-to-noise threshold of 1.5. Compounds were grouped with 5 ppm mass tolerance and 0.1 min retention time tolerance. Compound annotations were assigned using the following data sources in ranked order: mzCloud search, mzVault search, mass list search, predicted compositions, and ChemSpider search. The peak area for each putatively identified compound detected was exported to Microsoft Excel after processing the raw data and prior to data filtering.

Compounds that were identified to have a minimum of 70% spectral match, as determined by mzCloud, were summed to determine the relative complexity of each element. We found that complexity varied significantly between each library element (Figure S1). We discovered that two samples contained several compounds belonging to the PFAS family, including perfluoro-1-octanesulfonic acid (PFOS), perfluoro-1-butanesulfonic acid (PFBS), perfluoro-1-hexanesulfonic acid (PFHxS), and perfluorononanoic acid (PFNA) (all Level 1 annotations on the Schymanksi Scale),[34] which are highly stable fluorinated hydrocarbon compounds that are persistent environmental contaminants. Moreover, we identified that three of the library elements' most abundant peaks came from an identical compound, and in some elements, only 2 compounds were unique to that mixture.

Data Plate Analysis. Spectra from the FT-ICR mass spectrometer were exported for analysis in Python. The dedicated training spots were used to train 7 binary classifiers, one per library element. Logistic regression models or random forest models were trained as described in the main text. At different levels of complexity, the models used between 1 and 20,000 spectral features as inputs. After training binary classifiers on the 128 training spots, the models were applied to predict the presence or absence of the 7 library elements in the remaining 1393 spots. To recover the digital image payload, the classifiers' prediction matrix was unpermuted, reshaped, and Turbo decoded using the CommPy toolkit.[24]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c09234.

Untargeted analysis of the composition of the library elements; data recovery accuracy of each algorithm across each library element; and data recovery accuracy as a function of the number of compounds present per mixture (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jacob K. Rosenstein − *Brown University, Providence, Rhode Island 02912, United States;* orcid.org/0000-0001-9791-704X; Email: jacob_rosenstein@brown.edu

Authors

Selahaddin Gumus − *Brown University, Providence, Rhode Island 02912, United States*

Dana Biechele-Speziale − *Brown University, Providence, Rhode Island 02912, United States*

Katherine E. Manz − *Brown University, Providence, Rhode Island 02912, United States; University of Michigan, Ann Arbor, Michigan 48109, United States*

Kurt D. Pennell − *Brown University, Providence, Rhode Island 02912, United States;* orcid.org/0000-0002-5788-6397

Brenda M. Rubenstein − *Brown University, Providence, Rhode Island 02912, United States;* orcid.org/0000-0003-1643-0358

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c09234

Notes

The authors declare the following competing financial interest(s): S.G., D.B-S., B.M.R., and J.K.R. are co-founders of AtomICs Inc, an early-stage company working on topics related to the contents of this article.

■ ACKNOWLEDGMENTS

The authors are grateful to Jessica Plavicki, Ming Xian, Steven Lindhal, Xiang Ni, and Shi Xu for providing samples used in this work and Alan Bidart for algorithm optimization. S.G. acknowledges the Open Graduate Education Fellowship at Brown University. D.B.-S. acknowledges a National Science Foundation Graduate Research Fellowship under Grant No. 2018266952. This research was supported in part by funding from the Defense Advanced Research Projects Agency (DARPA W911NF-18-2-0031). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. This work was also supported in part by the National Science Foundation under Grant No. 1941344.

■ REFERENCES

(1) Ceze, L.; Nivala, J.; Strauss, K. Molecular Digital Data Storage Using DNA. *Nat. Rev. Genet.* **2019**, *20* (8), 456−466.

(2) Lee, H. H.; Kalhor, R.; Goela, N.; Bolot, J.; Church, G. M. Terminator-Free Template-Independent Enzymatic DNA Synthesis for Digital Information Storage. *Nat. Commun.* **2019**, *10* (1), No. 2383.

(3) Rutten, M. G. T. A.; Vaandrager, F. W.; Elemans, J. A. A. W.; Nolte, R. J. M. Encoding Information into Polymers. *Nat. Rev. Chem.* **2018**, *2* (11), 365−381.
(4) Nagarkar, A. A.; Root, S. E.; Fink, M. J.; Ten, A. S.; Cafferty, B. J.; Richardson, D. S.; Mrksich, M.; Whitesides, G. M. Storing and Reading Information in Mixtures of Fluorescent Molecules. *ACS Cent. Sci.* **2021**, *7* (10), 1728−1735.
(5) Cafferty, B. J.; Ten, A. S.; Fink, M. J.; Morey, S.; Preston, D. J.; Mrksich, M.; Whitesides, G. M. Storage of Information Using Small Organic Molecules. *ACS Cent. Sci.* **2019**, *5* (5), 911−916.
(6) Arcadia, C. E.; Kennedy, E.; Geiser, J.; Dombroski, A.; Oakley, K.; Chen, S.-L.; Sprague, L.; Ozmen, M.; Sello, J.; Weber, P. M.; Reda, S.; Rose, C.; Kim, E.; Rubenstein, B. M.; Rosenstein, J. K. Multicomponent Molecular Memory. *Nat. Commun.* **2020**, *11* (1), No. 691.
(7) ATDBio - Solid-Phase Oligonucleotide Synthesis. https://atdbio.com/nucleic-acids-book/Solid-phase-oligonucleotide-synthesis.
(8) Sanghvi, Y. S.; Ravikumar, V. T.; Scozzari, A. N.; Cole, D. L. Applications of Green Chemistry in the Manufacture of Oligonucleotide Drugs. *Pure Appl. Chem.* **2001**, *73* (1), 175−180.
(9) Branzei, D.; Foiani, M. Regulation of DNA Repair throughout the Cell Cycle. *Nat. Rev. Mol. Cell Biol.* **2008**, *9* (4), 297−308.
(10) Friedberg, E. C. DNA Damage and Repair. *Nature* **2003**, *421* (6921), 436−440.
(11) Matange, K.; Tuck, J. M.; Keung, A. J. DNA Stability: A Central Design Consideration for DNA Data Storage Systems. *Nat. Commun.* **2021**, *12* (1), No. 1358.
(12) Lutz, J.-F. Coding Macromolecules: Inputting Information in Polymers Using Monomer-Based Alphabets. *Macromolecules* **2015**, *48* (14), 4759−4767.
(13) Al Ouahabi, A.; Amalian, J.-A.; Charles, L.; Lutz, J.-F. Mass Spectrometry Sequencing of Long Digital Polymers Facilitated by Programmed Inter-Byte Fragmentation. *Nat. Commun.* **2017**, *8* (1), No. 967.
(14) Amalian, J.-A.; Trinh, T. T.; Lutz, J.-F.; Charles, L. MS/MS Digital Readout: Analysis of Binary Information Encoded in the Monomer Sequences of Poly(Triazole Amide)s. *Anal. Chem.* **2016**, *88* (7), 3715−3722.
(15) Barnes, J. C. Reading and Writing Data by Using Self-Immolative, Sequence-Defined Oligourethanes. *Chem* **2021**, *7* (6), 1417−1419.
(16) Al Ouahabi, A.; Charles, L.; Lutz, J.-F. Synthesis of Non-Natural Sequence-Encoded Polymers Using Phosphoramidite Chemistry. *J. Am. Chem. Soc.* **2015**, *137* (16), 5629−5635.
(17) Roy, R. K.; Meszynska, A.; Laure, C.; Charles, L.; Verchin, C.; Lutz, J.-F. Design and Synthesis of Digitally Encoded Polymers That Can Be Decoded and Erased. *Nat. Commun.* **2015**, *6* (1), No. 7237.
(18) Lee, J. M.; Kwon, J.; Lee, S. J.; Jang, H.; Kim, D.; Song, J.; Kim, K. T. Semiautomated Synthesis of Sequence-Defined Polymers for Information Storage. *Sci. Adv.* **2022**, *8* (10), No. eabl8614.
(19) Kennedy, E.; Arcadia, C. E.; Geiser, J.; Weber, P. M.; Rose, C.; Rubenstein, B. M.; Rosenstein, J. K. Encoding Information in Synthetic Metabolomes. *PLoS One* **2019**, *14* (7), No. e0217364.
(20) Rosenstein, J. K.; Rose, C.; Reda, S.; Weber, P. M.; Kim, E.; Sello, J.; Geiser, J.; Kennedy, E.; Arcadia, C.; Dombroski, A.; Oakley, K.; Chen, S. L.; Tann, H.; Rubenstein, B. M. Principles of Information Storage in Small-Molecule Mixtures. *IEEE Trans. NanoBiosci.* **2020**, *19* (3), 378−384.
(21) Kennedy, E.; Geiser, J.; Arcadia, C. E.; Weber, P. M.; Rose, C.; Rubenstein, B. M.; Rosenstein, J. K. Secret Messaging with Endogenous Chemistry. *Sci. Rep.* **2021**, *11* (1), No. 13960.
(22) Roy, B.; Shieh, M.; Xu, S.; Ni, X.; Xian, M. Single-Component Photo-Responsive Template for the Controlled Release of NO and H2S2. *J. Am. Chem. Soc.* **2023**, *145* (1), 277−287.
(23) Biechele-Speziale, D.; Camarillo, M.; Martin, N. R.; Biechele-Speziale, J.; Lein, P. J.; Plavicki, J. S. Assessing CaMPARI as New Approach Methodology for Evaluating Neurotoxicity. *Neurotoxicology* **2023**, *97*, 109−119.
(24) Taranalli, V. CommPy 2023. https://github.com/veeresht/CommPy. (accessed August 06, 2023).
(25) Blawat, M.; Gaedke, K.; Hütter, I.; Chen, X.-M.; Turczyk, B.; Inverso, S.; Pruitt, B. W.; Church, G. M. Forward Error Correction for DNA Data Storage. *Procedia Comput. Sci.* **2016**, *80*, 1011−1022.
(26) Grass, R. N.; Heckel, R.; Puddu, M.; Paunescu, D.; Stark, W. J. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angew. Chem., Int. Ed.* **2015**, *54* (8), 2552−2555.
(27) Erlich, Y.; Zielinski, D. DNA Fountain Enables a Robust and Efficient Storage Architecture. *Science* **2017**, *355* (6328), 950−954.
(28) Chatterjee, N.; Walker, G. C. Mechanisms of DNA Damage, Repair and Mutagenesis. *Environ. Mol. Mutagen.* **2017**, *58* (5), 235−263.
(29) Bonnet, J.; Colotte, M.; Coudy, D.; Couallier, V.; Portier, J.; Morin, B.; Tuffet, S. Chain and Conformation Stability of Solid-State DNA: Implications for Room Temperature Storage. *Nucleic Acids Res.* **2010**, *38* (5), 1531−1546.
(30) Xiao, F.; Sasi, P. C.; Yao, B.; Kubátová, A.; Golovko, S. A.; Golovko, M. Y.; Soli, D. Thermal Stability and Decomposition of Perfluoroalkyl Substances on Spent Granular Activated Carbon. *Environ. Sci. Technol. Lett.* **2020**, *7* (5), 343−350.
(31) Russell, M. H.; Berti, W. R.; Szostek, B.; Buck, R. C. Investigation of the Biodegradation Potential of a Fluoroacrylate Polymer Product in Aerobic Soils. *Environ. Sci. Technol.* **2008**, *42* (3), 800−807.
(32) Washington, J. W.; Ellington, J. J.; Jenkins, T. M.; Evans, J. J.; Yoo, H.; Hafner, S. C. Degradability of an Acrylate-Linked, Fluorotelomer Polymer in Soil. *Environ. Sci. Technol.* **2009**, *43* (17), 6617−6623.
(33) Dong, Y.; Sun, F.; Ping, Z.; Ouyang, Q.; Qian, L. DNA Storage: Research Landscape and Future Prospects. *Natl. Sci. Rev.* **2020**, *7* (6), 1092−1107.
(34) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097−2098.

https://doi.org/10.1021/acsomega.3c09234
ACS Omega 2024, 9, 19904−19910