Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Natural Coding of Language: SPECTRA is All You Need

Yi Feng
University of Science and Technology of China
Shijie Tao
University of Science and Technology of China
Xiaolong Zhang
University of Science and Technology of China
Meng Huang
University of Science and Technology of China
Donglai Zhou
University of Science and Technology of China
Linjiang Chen
University of Science and Technology of China
Daobin Liu
University of Science and Technology of China
Guokun Yang
University of Science and Technology of China
Hengyu Xiao
Hefei JiShu Quantum Technology Co. Ltd
Yang Wang
University of Science and Technology of China
Cheng Chen
University of Science and Technology of China
Pieter E S Smith
Hefei JiShu Quantum Technology Co. Ltd
Shuo Feng
sfeng18@ustc.edu.cn

University of Science and Technology of China https://orcid.org/0000-0003-0530-8381
Yan Huang
hyan@ustc.edu.cn

University of Science and Technology of China https://orcid.org/0000-0003-2512-2509

Song Wang

wsong09@ustc.edu.cn

University of Science and Technology of China https://orcid.org/0000-0003-1252-8091

Jun Jiang

jiangj1@ustc.edu.cn

University of Science and Technology of China https://orcid.org/0000-0002-6116-5605

Research Article

Keywords:

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9255583/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

In the data-driven era, the recording and preservation of information have become increasingly crucial. To surpass the binary storage capabilities of silicon-based materials, several novel storage media have been proposed in recent years, such as synthetic DNA and polymers with sequence-encoding characteristics. Herein, we present an alternative strategy that uses spectra of small molecules as a natural code for linking human language to molecular structures. By assigning words to codes derived from selected regions of experimental spectra, we demonstrate a workflow that generates spectra conditioned on language inputs and translates them into molecules. The molecules produced in this way yield computed and measured spectra that match their intended target codes, confirming the fidelity of the approach. These findings introduce spectra as a compact and interpretable channel – "SPECTRA[1] is All You Need" – for storing and expressing information, advancing data storage, cryptography, and the emerging field of chemistry-inspired information technologies.

Introduction

The rapid growth of digital data has driven an urgent search for new storage media that extend beyond conventional binary-based silicon technologies. Biological and chemical matter have emerged as promising candidates, as their structural diversity offers an immense encoding capacity[2,3]. For example, DNA-based storage systems use the sequence encoding capabilities of DNA to transcode digital data into nucleotide sequences, and achieve writing through de novo synthesis, in which nucleotides are assembled step by step in predefined orders[4-10]. Despite continuous advances in synthesis throughput and efficiency, the inherently serial nature of this process restricts writing speed, limits sequence length, and prevents substantial cost reduction at large scale. Synthetic polymers have also been explored as chemical carriers of digital information, but they share similar challenges: both DNA and polymers rely on their sequence as the code, and direct sequence writing remains expensive and time-consuming[11-19].

Commercially available small molecules offer a compelling alternative due to their structural diversity. Similar to movable type printing, the arrangement and combination of small molecules hold promise for achieving high-throughput parallel writing capabilities[20-22]. Storing data using small molecules requires converting molecular structures into sequence codes, and spectral techniques provide the optimal solution[23]. By capturing the fundamental interactions that define molecular identity, spectra constitute a natural, physically grounded code. If small molecules can be serialized through their spectral signatures, they could combine the synthetic accessibility of chemistry with the coding capacity required for large-scale data storage.

Herein, we present a framework that treat spectra as an information-bearing code linking language to molecules (Figure 1a). We introduce discrete "spectral codes" that bridge linguistic inputs and spectral representations, enabling words to be mapped onto specific spectral patterns. A generative model produces spectra based on these codes while ensuring spectral fidelity. Then, a language-inspired neural architecture translates spectra into molecular structures and guarantee molecular chemical validity.

Reinforcement learning and quantum-chemical validation ensure that generated molecules faithfully reflect their intended codes while remaining chemically plausible. As a proof of principle, we selected the paper title “The Natural Code of Nature: SPECTRA is All You Need” as a case study (Figure 1b), showing that words from a text can be converted into spectral codes, used to generate molecules, and validated through both computation and experiment. This work establishes spectra as a central representational layer linking language, spectra, and molecular design, positioning spectroscopy as a new paradigm for chemistry-inspired information storage.

Spectral Code Construction and Workflow Design

To construct the spectral code, we firstly assembled an experimental dataset of 21,633 infrared vibrational spectra measured in-house (Figure S1, detail in the Supplementary Information). Each spectrum was mapped to a uniform 0–4,000 cm^{-1} wavenumber grid at 1 cm^{-1} resolution (4,000 dimensions). The resulting vector was normalized to unit sum, partitioned into ten pre-specified spectral regions, and integrated within each region to obtain ten aggregates (details in the Supplementary Information). These aggregates were scaled to integers in [0,9] with the total constrained to 9, yielding the code used throughout this work. Figure 2a illustrates the procedure for C1N=COC_2=C1OC=N_2, whose spectrum encodes to 0142200000. This compact coding preserves key spectral characteristics while substantially reducing complexity, enabling word-code mapping for downstream generation.

The first step in our workflow, W2C establishes a mapping between English words and spectral codes (Figure 2a). Given that the value range of spectral codes reaches 10^{10}, sufficient to cover all English words, we developed a codebook that maps each English word to a specific spectral code (detail in the Supplementary Information). Then, the key challenge is reconstructing full spectra curves from spectral codes due to the high dimensionality and inherent complexity of the full spectral curves. We developed a Conditional Variational Autoencoder (CVAE) based on AI-generated content (AIGC)[24], referred to as the C2S model, specifically designed to handle this complexity. The C2S model was constructed by training an encoder-decoder architecture using both spectral codes and vibrational spectra, with the spectral codes serving as conditional inputs for the decoder. Once trained, the decoder was employed to generate spectra based on given spectral codes (Figure 2c, detailed architecture of the model in the Supplementary Information). The well-trained C2S model successfully generates spectra that align with the constraints of the input spectral codes, bridging the gap between the simplified spectral codes and the high-fidelity spectra. Figure 2b and Figure S2 illustrate examples of spectra generated from spectral codes. These results validate the effectiveness of the C2S model in managing the complexity of vibrational spectra and achieving high-precision conversion between spectral codes and full spectral curves.

The next step in our workflow is to generate molecular structures from spectral information. For this purpose, the self-referencing embedded strings (SELFIES) encoding method[25] was used to represent chemical structures, as detailed in the Supplementary Information. SELFIES leverages inherent syntax control mechanisms to ensure the chemical validity of generated molecules (Figure 2e). Additional

explanations and examples are provided in Figure S3. Infrared spectra were mapped to molecular SELFIES encodings using a transformer model based on the BART architecture[26]. The S2M model integrates a bidirectional encoder and an auto-regressive decoder to effectively capture the complex relationships between spectral features and molecular structures. During the decoding, the model generates SELFIES encodings incrementally, ensuring both chemical validity and structural coherence. To further enhance prediction accuracy, we employed beam search inference[27], which is an effective method for optimizing sequence generation[28] that evaluates all potential next tokens at each step, ranks them by their softmax-calculated probabilities, and selects the top candidates for subsequent steps. This process generates multiple sequences, where the probability of each sequence is the product of its token probabilities. The top 10 candidate molecules with the highest probabilities are retained for evaluation. We applied this technique to a test set comprising 21,633 infrared spectra for molecular structure prediction. The model achieved top-1 accuracy of 76% and top-10 accuracy of 94% in predicting molecular structures (Figure 2f). These results demonstrate the S2M model's capability to accurately predict viable molecular structures from infrared spectra.

Additionally, to assess the model's generative capability, we compared the chemical space of molecules originally selected from the QM9 database with that of the generated molecules (Figure S4a). Using the RDKit library[29], we computed Morgan molecular fingerprints with a radius of 3 and a bit length of 2048. The fingerprints were then reduced to two dimensions using the t-SNE method from the Scikit-learn library with default parameter settings. The results show that the generated molecules cover the original set of molecules' chemical space and extend into a broader chemical space. A total of 51,145 distinct molecules were generated, among which 15,810 new molecules lie outside the molecules originally selected from the QM9 dataset. Some of the molecules beyond those selected from the QM9 dataset are presented in the Figure S4b.

Information Storage Fidelity Verification

We verify that spectral codes act as reliable information carriers across the W2C → C2S → S2M pipeline. For each target word, W2C yields a code that conditions C2S to generate an infrared spectrum, and S2M proposes molecular structures. We then compute each candidate's DFT spectrum and convert it back to a code using the S2C tool. Agreement between this reconstructed code and the target code serves as a pass/fail criterion prior to any optional measurements. Under this baseline, per-segment predictions closely follow the identity line (MAE = 0.169, RMSE = 0.234, R^2 = 0.954), and residuals are centered near zero (μ = 0.004, σ = 0.138; Figure S5).

To improve accuracy further, we introduce a lightweight RL loop that iteratively refines a proxy model guiding S2M selection (Figure 3a). Figure 3b shows the relative prediction error as a percentage across iterations. The error steadily decreases, starting above 10% in the early iterations and dropping to approximately 2% by the final iteration. This consistent downward trend underscores the improvement in the model's predictive accuracy as reinforcement learning refines the workflow. Figure 3c illustrates the distribution of the mean absolute prediction error across 10 iterations. The violin plots show a gradual

reduction in error variability over successive iterations, with black vertical lines indicating the interquartile range and red dots representing the mean values. Initially, the prediction error spans a broad range, but by the 10th iteration, the error is focused within a much smaller range, reflecting the iterative improvement achieved by the proxy model as it is refined through reinforcement learning. After conducting RL loop, we conduct a large-scale closed-loop theoretical verification on 20,000 words traversing W2C → C2S → S2M → DFT → S2C. The end-to-end results concentrate tightly around the identity, with a mean relative error of 3.27% (Figure 3d), confirming improved workflow accuracy after RL and demonstrating code-level fidelity suitable for large-scale screening. Overall, these results demonstrate that RL can significantly optimize the workflow by enhancing prediction performance.

To move beyond computational verification, we selected 20 internationally recognized scientists' given names (single-token words). Each name was mapped to its corresponding code, which is then used by C2S model to generate a spectral curve. This spectral curve was subsequently processed by S2M model to generate a molecule. After that, we measured Fourier-transform infrared (FTIR) spectra of them, integrated the ten predefined regions to obtain measured codes, and compared them with the initial codes. As shown in Figure 4, an agreement was achieved between the initial codes and the codes derived from experimental spectral. It provides a direct experimental validation of the whole workflow (language-to-code-to-spectra-to-molecule).

Data storage of 2D images

Beyond 1D texts, we also store 2D images as ordered molecular sequences using a row-wise code → spectrum → molecule pipeline. Taking a smiley emoji and a Chinese character as examples (Figure 5), each image is represented as a 10×10 pixel grid (10 rows × 10 columns). For each row, the 10 pixel intensities are linearly normalized to non-negative values with a sum of 1, then multiplied by 9 and rounded to numbers in the range of {0,…,9}. If rounding changes the total, we apply minimal ±1 adjustments so the digit sum returns to 9. Rows that are entirely blank are labeled N/A (no spectrum or molecule generated).

Each non-blank row code is then sent to the C2S model to synthesize a 1D FTIR spectra, and the spectrum is converted to a molecule by the S2M model. Concatenating the molecules in row order yields a compact, addressable molecular representation of the image. Decoding inverts the steps: measure spectra from the stored molecules → recover the 10-digit row codes → reconstruct the 10 rows. The discrete code space is the same one defined in our W2C step and, together with the two-stage C2S/S2M pipeline, provides robustness to noise/quantization while keeping a natural row address. The results in Figure 5 demonstrate that our framework exhibits excellent scalability in terms of information storage dimensions.

Discussion

In this work, we present a novel information storage approach to leveraging molecular vibrational spectra as a medium for encoding language into chemical molecules. Vibrational spectra are converted into 10

integers sequences, termed spectral codes, which can be precisely mapped to English words. By using a dual-model framework, incorporating a CVAE-based C2S model and a BART Transformer-based S2M model, molecular structures are generated to meet specific English word requirements for information storage. Reinforcement learning further optimizes molecular generation conditions and prioritizes molecule selection, significantly enhancing workflow efficiency and accuracy. The entire workflow, spanning from words to spectral codes, spectra, and molecular structures, exhibits robust performance validated by DFT calculations and experimental FTIR measurements. These findings underscore the potential of vibrational spectra as tools for molecular innovation, improving the precision and efficiency of molecular generation, providing new directions for language-encoded molecular design, and opening new avenues for data storage technologies based on spectroscopically intelligent molecular generation.

Methods

Spectral Data Processing and Normalization

Each spectrum is represented as a 4000-dimensional vector. To simplify the data, the spectrum is divided into intervals of 400 dimensions, and the sum of each interval is calculated:

$$I_i = \sum_{j=(i-1)\times 400+1}^{i\times 400} S_j \quad (1)$$

where I_i is the integral value for segment i and S_j is the spectral value at position j. The next step involves normalizing this 10-dimensional vector such that its components are integers between 0 and 9, with the sum equal to 9. The normalized vector is computed as follows:

$$C_i = round\left(\frac{I_i \times 9}{\sum_{i=1}^{10} I_i}\right) \quad (2)$$

where C_i is the normalized spectral code.

C2S Model Architecture

The C2S model employs a CVAE to generate vibrational spectra from spectral codes. The encoder processes a 4000-dimensional vibrational spectrum concatenated with a 10-dimensional spectral code. It uses two Transformer layers with attention mechanisms to capture the relationships in the data and generate latent variables. The decoder then reconstructs the vibrational spectrum from the latent vector and spectral code, ensuring the output is within the desired range. The model is trained with a

combination of Mean Squared Error for reconstruction loss and Kullback-Leibler divergence to shape the latent space. For more details, refer to the Supplementary Information.

S2M Model Architecture

The S2M model uses a sequence-to-sequence architecture based on the BART framework to generate molecular structures from vibrational spectra. The input spectrum is first encoded using a Multilayer Perceptron network that reduces its dimensionality. This encoded vector is then passed into the BART model, which decodes it into valid SMILES strings representing molecular structures. Beam search is applied during inference to explore multiple candidate molecules. The final molecules are ranked based on their spectral similarity to the input using metrics such as Spearman correlation. For additional information, see the Supplementary Information.

Integration of Reinforcement Learning into the Workflow

Reinforcement learning for discrepancy minimization. We incorporate reinforcement learning to reduce the gap between the S2C-predicted code (from the DFT spectrum of each S2M-generated SMILES) and the initial target code. At every iteration, S2M proposes candidates; we compute their DFT spectra, obtain S2C codes, and define a reward from the code discrepancy (e.g., negative L1). A proxy model—taking RDKit fingerprints, model scores, and conditioning offsets as features—prioritizes candidates. The top 30% lowest-discrepancy molecules are retained; their conditioning offsets are updated and used to seed the next round (Figure S5). This iterative scheme accelerates convergence toward code-consistent molecules and improves time-to-target without altering the core workflow.

Relative Error Calculation for Spectral Code Matching

To assess the accuracy of the generated spectra, we implement a method for fine-tuning integral values of spectral segments to achieve exact matches with target spectral codes, while preserving the sum constraint. The relative error is calculated as the total adjustment percentage required for each segment compared to the original integral values, reflecting the deviation from the S2C-predicted code (obtained by applying S2C to the DFT-calculated spectrum) to the target code. Additional information is available in the Supplementary Information.

Declarations

Data and Code availability

All data and code of this study are openly accessible at https://github.com/taoshijie111/Spectra-Code.

Supporting Information

The Supporting Information is available free of charge at online

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (22573100, 12227901, 22303088, 22303091, 22393892, 22025304, 22033007), the Innovation Program for Quantum Science and Technology (2021ZD0303303), the Fundamental Research Funds for the Central Universities (WK2060000094, WK9990000129, WK9990000130), and the National Key Research and Development Program of China (2023YFA1508200). It was also supported by the Hefei Comprehensive National Science Center. Numerical calculations were performed at the Hefei Advanced Computing Center and Supercomputing Center of the University of Science and Technology of China. This work was also supported by the National Natural Science Foundation of China (Grant No. 22201277). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

1. SPECTRA (2025) Spectroscopic Programming Molecule for Coding Texts with Rule-embedded AI. https://aisci.ustc.edu.cn/genspectra
2. Ceze L, Nivala J, Strauss K (2019) Molecular digital data storage using. DNA Nat Rev Genet 20:456–466
3. Service RF (2017) (News, 2 Mar 2017) DNA could store all of the world's data in one room. *Science*
4. Church GM, Gao Y, Kosuri S (2012) Next-generation digital information storage in DNA. Science 337:1628
5. Goldman N, Bertone P, Chen S, Dessimoz C, LeProust EM, Sipos B, Birney E (2013) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494:77–80
6. Erlich Y, Zielinski DDNA (2017) Fountain enables a robust and efficient storage architecture. Science 355:950–954
7. Organick L, Ang SD, Chen Y-J et al (2018) Random access in large-scale DNA data storage. Nat Biotechnol 36:242–248
8. Banal JL, Shepherd TR, Berleant J et al (2021) Random access DNA memory using Boolean search in an archival file storage system. Nat Mater 20:1272–1280
9. Grass RN, Heckel R, Puddu M, Paunescu D, Stark WJ (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed 54:2552–2555
10. Tabatabaei SK et al (2020) DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nat Commun 11:1742

11. Hoose A et al (2023) DNA synthesis technologies to close the gene writing gap. Nat Rev Chem 7:144–161
12. Nguyen BH et al (2021) Scaling DNA data storage with nanoscale electrode wells. Sci Adv 7:eabi6714
13. Zhang C et al (2024) Parallel molecular data storage by printing epigenetic bits on DNA. Nature 634:824–832
14. Colquhoun H, Lutz J-F (2014) Information-containing macromolecules. Nat Chem 6:455–456
15. Lutz J-F (2013) Writing on polymer chains. Acc Chem Res 46:2696–2705
16. Chan-Seng D, Zamfir M, Lutz J-F (2012) Polymer-chain encoding: Synthesis of highly complex monomer sequence patterns by using automated protocols. Angew Chem Int Ed 51:12254–12257
17. Al Ouahabi A et al (2015) Synthesis of non-natural sequence-encoded polymers using phosphoramidite chemistry. J Am Chem Soc 137:5629–5635
18. Cavallo G, Al Ouahabi A, Oswald L, Charles L, Lutz J-F (2016) Orthogonal synthesis of easy-to-read information-containing polymers using phosphoramidite and radical coupling steps. J Am Chem Soc 138:9417–9420
19. Laurent E, Amalian J-A, Ren J-M, Lutz J-F (2020) High-capacity digital polymers: Storing images in single molecules. Macromolecules 53:4022–4029
20. Cafferty BJ et al (2019) Storage of information using small organic molecules. ACS Cent Sci 5:911–916
21. Arcadia CE et al (2020) Multicomponent molecular memory. Nat Commun 11:691
22. Bohn P et al (2022) Molecular data storage with zero synthetic effort and simple read-out. Sci Rep 12:13878
23. Dahlhauser SD et al (2022) Molecular encryption and steganography using mixtures of simultaneously sequenced, sequence-defined oligourethanes. ACS Cent Sci 8:1125–1133
24. Liang C, Du H, Sun Y et al (2024) Generative AI-driven semantic communication networks: Architecture, technologies and applications. IEEE Trans Cogn Commun Netw
25. Ross J, Belgodere B, Chenthamarakshan V et al (2022) Large-scale chemical language representations capture molecular structure and properties. Nat Mach Intell 4:1256–1264
26. Born J, Manica M (2023) Regression Transformer enables concurrent sequence regression and generation for molecular language modelling. Nat Mach Intell 5:432–444
27. Pierangeli D, Conti C (2023) Single-shot polarimetry of vector beams by supervised learning. Nat Commun 14:1831
28. Sadeghnejad-Barkousaraie A, Bohara G, Jiang S et al (2021) A reinforcement learning application of a guided Monte Carlo Tree Search algorithm for beam orientation selection in radiation therapy. Mach Learn Sci Technol 2:035013
29. Landrum G (2019) RDKit. www.rdkit.org

Figures

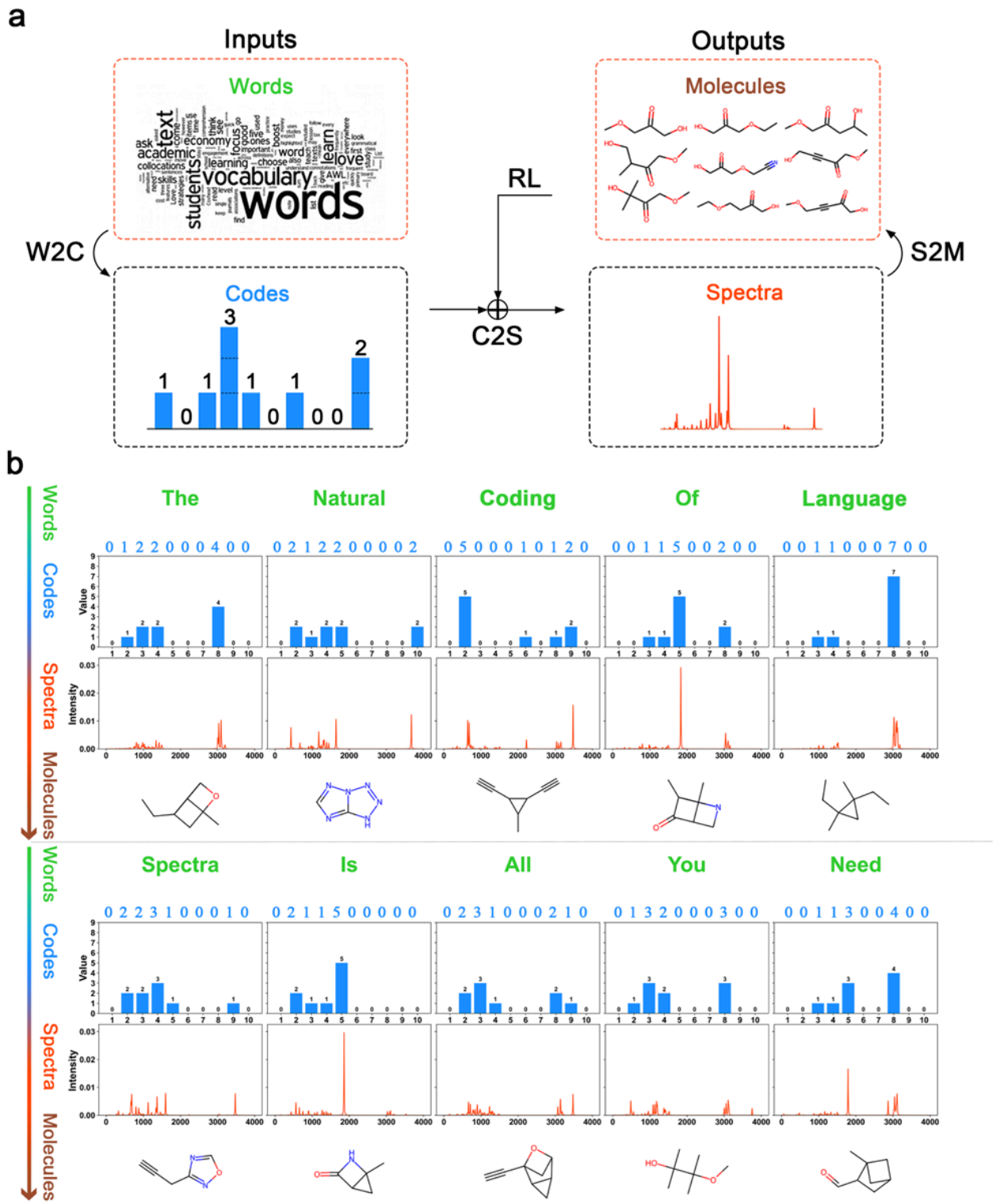


Figure 1

Overview of the workflow whereby words are converted to molecular structures with a dual-model AI-generated content framework. The process integrates spectral encoding, spectrum generation, molecular generation, and reinforcement learning to optimize conditions and molecule selection. **a**

Overall process from words to molecules, including Words-to-Spectra (W2S), Codes-to-Spectra (C2S), Spectrum-to-Molecule (S2M), and reinforcement learning (RL) for iterative optimization. **b** Stepwise conversion of the paper title into spectral codes, then generated spectra, and then the resulting molecular structures.

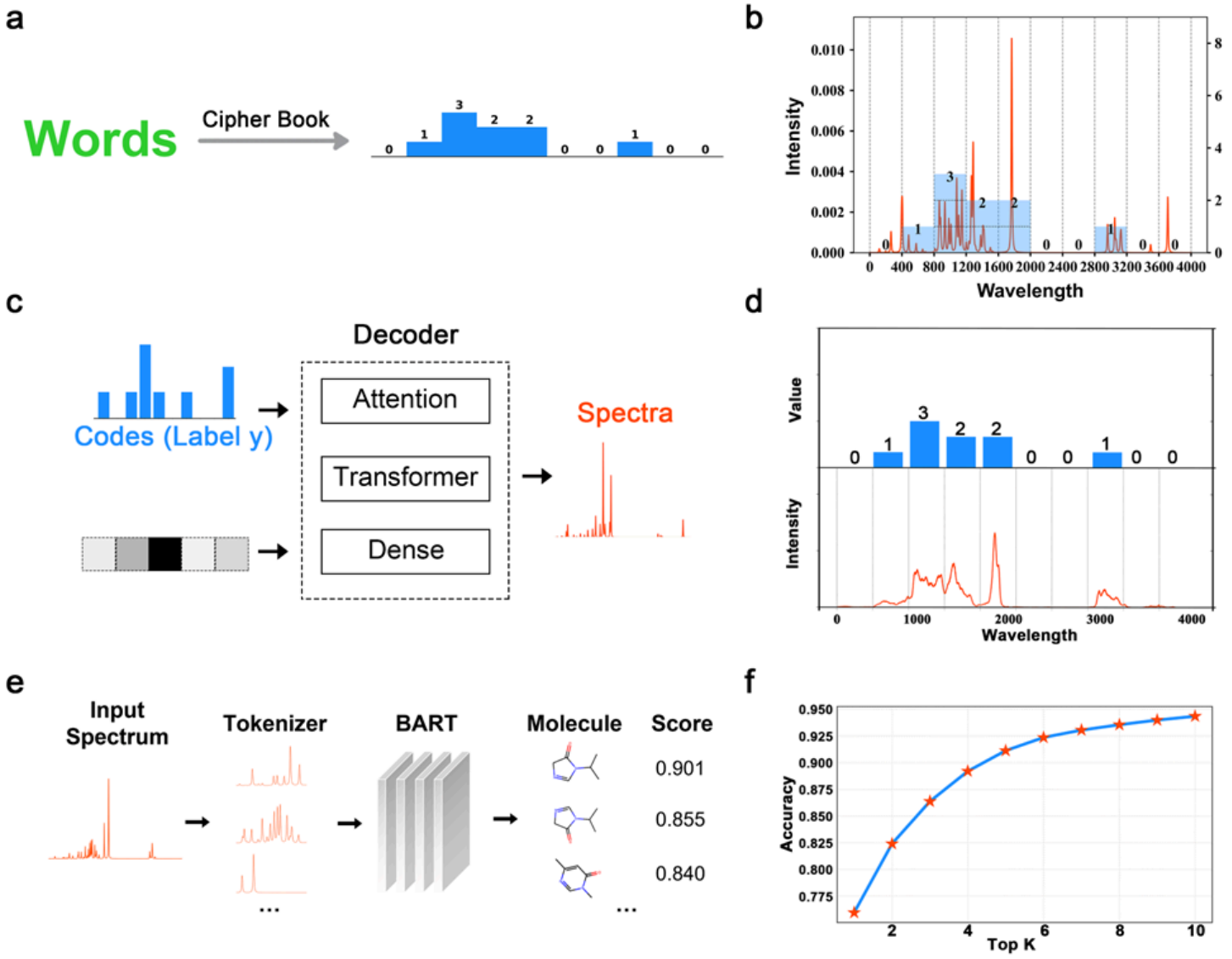


Figure 2

Integrated spectral–language–molecule workflow (W2C→C2S→S2M) and evaluation under the AI-generated content (AIGC) paradigm. **a** W2C: an English word is mapped via a cipher book to a ten-digit spectral code. **b** Example infrared spectrum with bin boundaries (dashed lines) and integrated bin heights (blue), yielding the normalized code (digits above bars). **c** C2S: a conditional decoder with Attention-Transformer-Dense blocks generates spectra from code labels y. **d**Representative spectra generated by C2S for two example codes (e.g., 0132200100 and 0111200400); the blue bars indicate the conditioning code. **e**Spectrum-to-Molecule (S2M): the input spectrum is tokenized and decoded by a BART model; beam search returns candidate molecules with scores. **f**Test-set top-k accuracy of S2M, showing increasing recall as k grows.

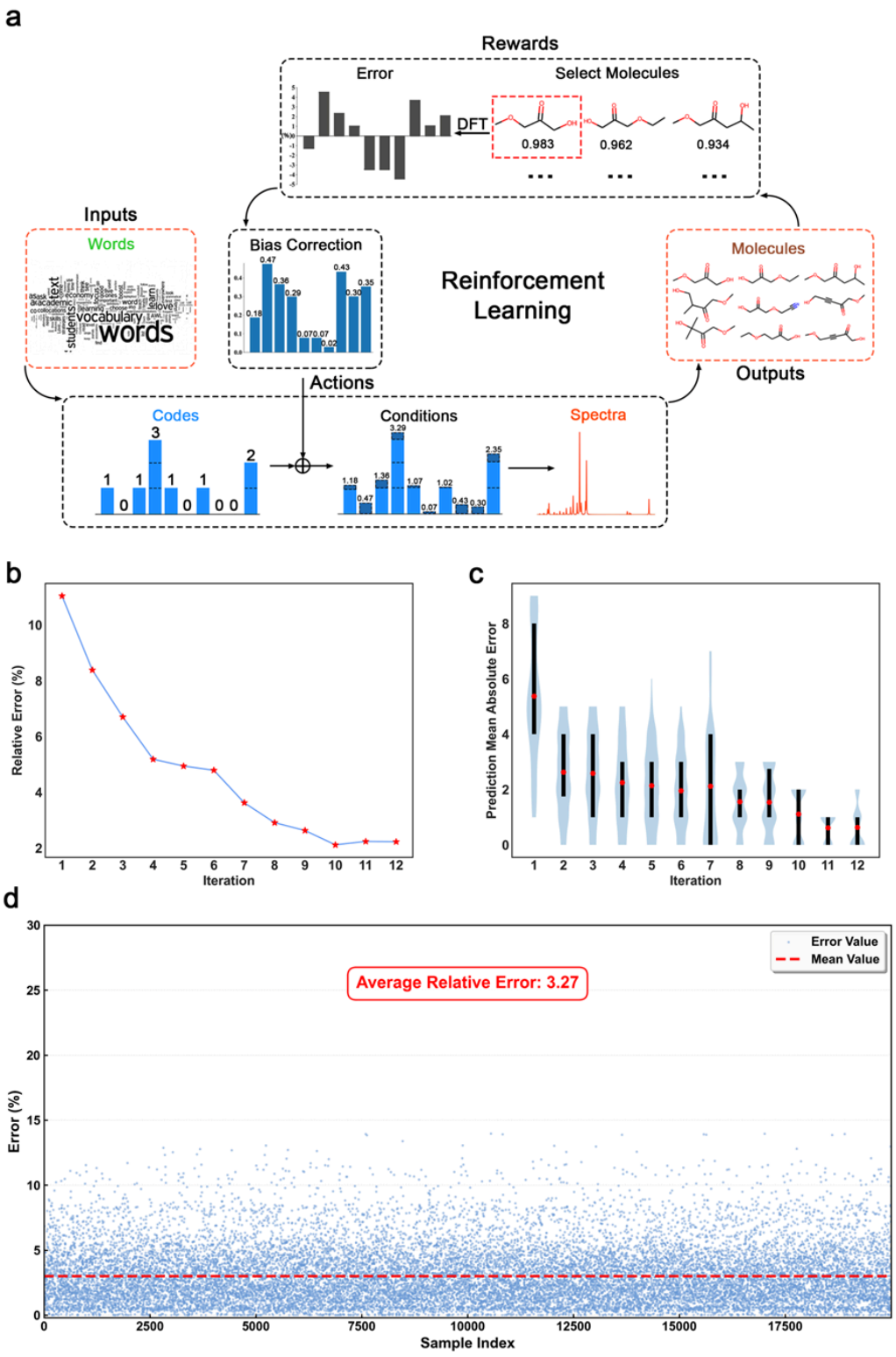


Figure 3

Reinforcement learning–enhanced molecular generation. **a** Workflow of the reinforcement learning loop. Words are mapped to spectral codes, which are combined with randomly initialized bias correction values to form conditions for the C2S model, generating vibrational spectra that constitute the Environment. The spectra are then input into the S2M model to generate molecular structures as States. The reinforcement learning agent estimates the discrepancies between the spectral codes derived from

molecules and the target spectral codes, and the top 30% of molecules with the smallest predicted errors are selected. Their DFT spectral discrepancies serve as Rewards to update the bias correction values and conditions for the next iteration. **b** Violin plots of mean absolute prediction error across RL iterations; black bars indicate the interquartile range and red dots the means, showing progressive error contraction. **c** Convergence of relative prediction error (%) versus iteration, decreasing from >10% in early iterations to ~2% at the final iteration. **d**Large-scale closed-loop verification on 20,000 words (W2C →C2S → S2M → DFT → S2C); distribution of sample-wise relative errors with the mean marked (3.27%), demonstrating high code-level agreement after RL.



Figure 4

Experimental validation with measured FTIR spectra. For 20 single-token scientist names (panel titles), each target code is shown (blue, top), followed by the S2M-generated molecule (structure) with its measured FTIR spectra (green), and the code obtained by integrating the measured spectrum (purple). The measured codes agree with the targets across cases, providing direct laboratory validation of the workflow.

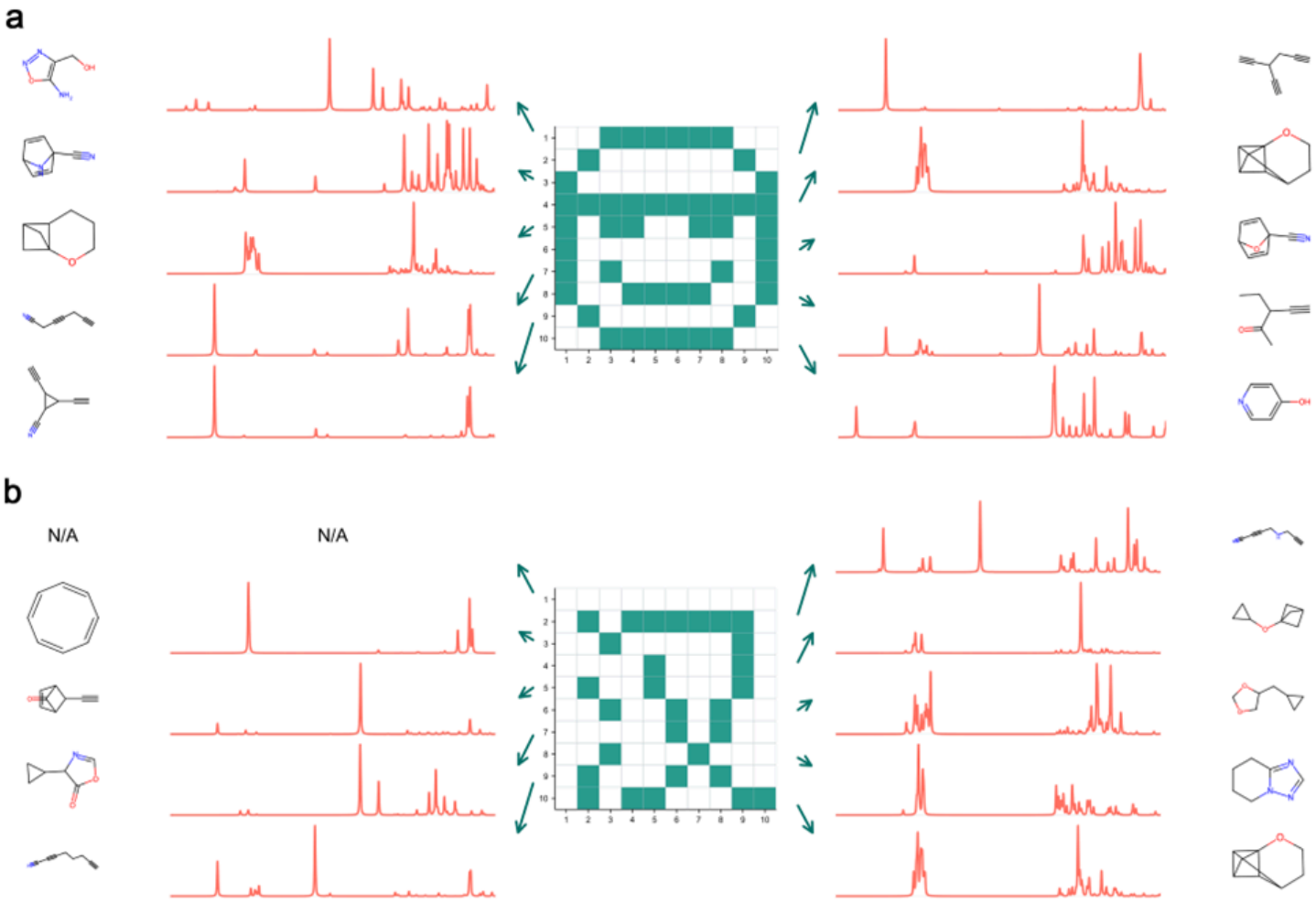


Figure 5

Row-wise storage of 2D images as molecules using the C2S/S2M pipeline. **a** The 10×10 “smiley”. For each row, 10 pixel values are normalized, scaled by 9, rounded to digits, and minimally corrected so the digit sum equals 9. The code drives C2S to generate an FTIR spectra (coral), which S2M converts into a molecule (side panels). **b** The Chinese character “”.The first row is blank and therefore shown as N/A for both spectrum and molecule. Arrows indicate image → spectrum → molecule. Colors follow the teal/coral/indigo scheme for center tiles, spectra, and arrows; molecular drawings use RDKit defaults.

Supplementary Files

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

- SI.docx