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.

ChemCAPTCHA: A domain-specific challenge–response framework for controlling access to chemical data

Zeyu Deng
South China University of Technology
Peng Jia
InsightChemy Co., Ltd
Hongyu Zhang
InsightChemy Co., Ltd
Jingxin Zhai
China Pharmaceutical University
Yang Tian
tiany21@mails.tsinghua.edu.cn

InsightChemy Co., Ltd

Research Article

Keywords: CAPTCHA, cheminformatics, chemical data access control, molecular-structure recognition, behavioural risk analysis

Posted Date: October 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11364217/v1

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

Additional Declarations: No competing interests reported.

ChemCAPTCHA: A domain-specific challenge − response framework for controlling access to chemical data

Zeyu Deng[1], Peng Jia[2], Hongyu Zhang[2], Jingxin Zhai[4], Yang Tian[2,3*]

[1]School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, P. R. China.

[2]InsightChemy Co., Ltd., Jinan 250000, P. R. China.

[3]School of Pharmaceutical Sciences, Tsinghua University, Beijing 100084, P. R. China.

[4]School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing 211198, P. R. China.

*Corresponding author: Yang Tian (tiany21@mails.tsinghua.edu.cn)

Abstract

Chemical data platforms are vulnerable to automated bulk collection and low-cost human-solving services, yet general-purpose completely automated public Turing tests to tell computers and humans apart (CAPTCHAs) do not exploit users' chemical expertise. Here we present ChemCAPTCHA, a domain-specific verification system that combines molecular-structure recognition, manual drawing and behavioural analysis. Chemically valid candidate molecules are rendered with visual perturbations that preserve the underlying molecular graph. A constrained Ketcher editor blocks structure import and records the drawing process. Two forms of within-session randomisation shuffle the element toolbar and alter selected element editing outcomes, forcing solvers to track and correct structural changes. In a test of 1,000 molecular images, MolScribe recognition accuracy decreased from 96.4% without perturbation to 0.0% with perturbation. The server verifies answers by canonical structure matching and independently assigns a behavioural risk score from interaction logs. ChemCAPTCHA therefore provides a traceable, domain-specific layer for protecting chemical databases, although its security and usability require evaluation against adaptive attacks and diverse users.

Scientific contribution

ChemCAPTCHA introduces a domain-specific access-control challenge–response framework that combines visually perturbed molecular prompts, constrained manual structure drawing, within-session randomisation, time-limited interaction challenges and behavioural analysis to raise automation costs while enabling verifiable answers and auditable risk assessment.

Keywords: CAPTCHA; cheminformatics; chemical data access control; molecular-structure recognition; behavioural risk analysis

Introduction

Chemical data platforms underpin both chemical research and industrial practice. Curated commercial resources such as CAS SciFinder, Reaxys and the Cambridge Structural Database provide validated links among compounds, structures, reactions and the literature, helping chemists assess precedent, plan experiments and interpret structure–property relationships.[1–4] Public repositories such as PubChem and ZINC broaden access to chemical identities, properties and purchasable molecules, enabling large-scale screening, computational modelling and candidate selection.[5–8] Open reaction databases provide reusable evidence for route design and for training and evaluating reaction-prediction and retrosynthesis models,[9,12–17] whereas electronic laboratory notebook (ELN) records preserve procedures, observations and provenance needed for reproducibility and knowledge transfer across research and production organisations.[10,11] Chemical-reagent catalogues connect digital knowledge to procurement and manufacturing by documenting material availability, specifications and suppliers.

These platforms are therefore not merely search tools: they are high-value scientific infrastructure and, for many providers and organisations, strategic commercial assets. Their reliability depends on sustained investment in curation, manual validation, standardisation, storage and maintenance, making responsible access and protection essential. However, exposing data through websites and application programming interfaces also creates opportunities for persistent automated querying, bulk traversal and centralised downloading. At scale, such activity can enable unauthorised copying or aggregation, circumvent intended access controls and consume disproportionate bandwidth and computing resources; bursts of frequent requests may overload servers, slow response times and reduce availability for legitimate users.[18,19] Chemical data platforms must therefore distinguish legitimate scientific

use from unauthorised automated collection while preserving practical and efficient access for bona fide users.

General-purpose CAPTCHAs use text, image or audio recognition, puzzles, visual reasoning, and behavioural signals such as pointer trajectories.[20] Text CAPTCHAs are inexpensive, but security often comes at the expense of readability. Deep learning has also reduced the difficulty of automated solving.[21,22] Image and visual-reasoning CAPTCHAs are intuitive, yet specialised recognition models and attack frameworks have exposed their security limits.[23,24] Audio CAPTCHAs support users with visual impairments, but recognition performance, noise and usability constrain their effectiveness.[25,26] Puzzle, slider and behavioural schemes increase interaction costs, although trajectories can be simulated and classification depends on devices, habits and risk thresholds.[27] Large user studies have also revealed substantial differences in completion time, success rate and perceived usability among CAPTCHA types.[28,29] Multimodal large language models have now been evaluated as general-purpose CAPTCHA solvers, further intensifying the trade-off between security and usability.[30-32]

For chemical data platforms, general-purpose CAPTCHAs have an additional limitation. They test broad perceptual abilities rather than differences in domain knowledge and practical experience. Chemists can generally recognise and draw simple molecular structures. Automated programs and low-skill human-solving services may instead require specialised parsing tools or additional training. A domain-specific CAPTCHA could exploit this difference to raise the cost of bulk solving without imposing an excessive burden on legitimate users.

Several engineering efforts have incorporated chemical knowledge into CAPTCHAs. One open-source chiral-carbon CAPTCHA presents tiled molecular images and asks users to

identify tiles containing chiral carbon atoms.[33,34] Although this task requires stereochemical knowledge, it relies largely on static image recognition and click selection. It verifies the final selection but captures little information about the response process. Answers may also depend on image segmentation when a target atom lies at a tile boundary.[34] Chemical structure drawing offers a way to combine domain knowledge with analysis of the interaction process. However, optical chemical structure recognition tools, including DECIMER,[35] DECIMER.ai,[36] MolScribe,[37] MolMiner[38] and MolGrapher,[39] can recover atoms, bonds and some stereochemical information from static images. Verification based only on static recognition may therefore remain vulnerable to automated solving.

Here we present ChemCAPTCHA, a domain-specific challenge–response framework for controlling access to chemical data. The system targets automated programs and low-skill human-solving services while assuming that legitimate users possess basic molecular-recognition and drawing skills. It generates simple target molecules, applies visual perturbations that preserve molecular-graph semantics and requires users to reconstruct each structure manually in a constrained Ketcher environment. By disabling structure-import shortcuts, including simplified molecular input line entry system (SMILES) pasting, ChemCAPTCHA enforces stepwise construction and combines image interpretation, structure drawing and continuous interaction within a single response.[40] Within-session element-toolbar randomisation disrupts fixed-coordinate scripts, whereas element-editing randomisation requires solvers to inspect and correct the evolving structure after each operation. Throughout this process, the system records pointer trajectories, tool selections, structural edits, corrections and delays. Canonical structure matching verifies the final answer, while the interaction record yields a separate behavioural risk score. Together, these components position ChemCAPTCHA as one layer of protection for chemical databases, literature portals, ELNs and reagent-procurement platforms.

Results

System architecture

ChemCAPTCHA is designed for controlled-access chemical data services whose intended users possess basic chemical expertise. Its design follows three principles: controllable challenges, auditable interactions and verifiable answers (Fig. 1). A request for a protected resource initiates a workflow comprising three connected modules. These modules generate and process the challenge, support interactive structure drawing and behaviour recording, and verify the response before an access decision.

The challenge-generation and image-processing module produces chemically valid candidate molecules and renders them as CAPTCHA prompts. A controlled depth-first search generates molecular scaffolds, followed by chemical-validity checks and canonical deduplication. Visual perturbations then increase the difficulty of static image recognition without changing the underlying molecular graph.

The interactive structure-drawing module presents a constrained Ketcher editor. It blocks input routes that could bypass manual drawing, including structure import, pasting and drag-and-drop. Within-session uncertainty is introduced by randomising the element-toolbar order and selected element editing outcomes. As users construct the target structure, the behaviour-recording module captures pointer trajectories, tool selections, edits, corrections and inter-event intervals. Selected structural keyframes provide a temporal record from which the response process can be reconstructed.

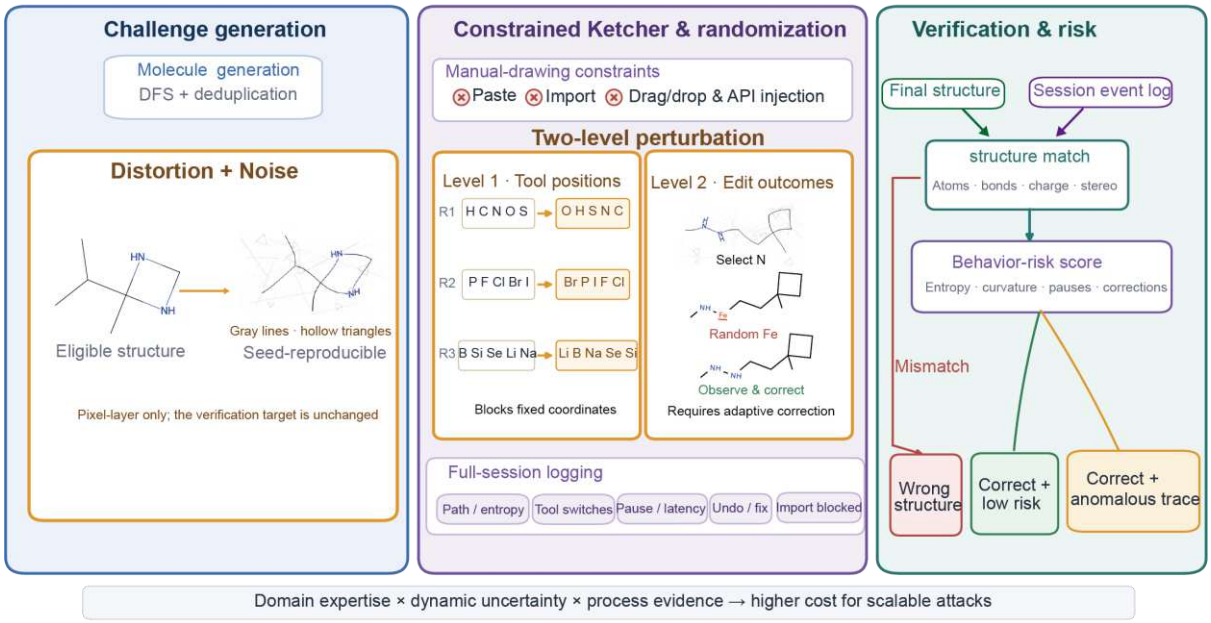


Fig. 1 | Overall architecture and verification workflow of ChemCAPTCHA.

Molecule generation

To limit repeated use of a fixed challenge library, we implemented a small-molecule generator based on molecular-graph enumeration and chemical-rule validation. Inputs specify the target number of heavy atoms, the maximum number of heteroatoms and whether one ring is permitted. The method enumerates all-carbon scaffolds, substitutes heteroatoms, validates the resulting structures and removes duplicates. It returns a set of canonical SMILES strings.

Scaffold generation begins with one carbon atom and adds carbon atoms incrementally by depth-first search (DFS). Each new atom forms a single bond to an existing carbon whose degree remains no greater than four. Once the scaffold reaches the target heavy-atom count, the structure is retained as a candidate. After each expansion, the system converts the intermediate structure to canonical SMILES and prunes topologically equivalent branches. When rings are permitted, one single bond is added between eligible unconnected atoms. RDKit then evaluates chemical validity and ring count, retaining acyclic or simple monocyclic scaffolds.[41]

After scaffold generation, the system identifies substitutable CH_2 sites. Eligible sites are non-aromatic carbon atoms with degree two and two total hydrogens. The program retains the original scaffold and enumerates combinations of CH_2 sites up to the specified limit. Selected atoms are replaced with O, S or NH. RDKit evaluates each substitution, and candidates with invalid valence states or parsing failures are discarded.

All candidates undergo a common standardisation and deduplication procedure. Structures that pass RDKit validation and satisfy the ring-count constraint have their stereochemical annotations removed. They are then converted to non-isomeric canonical SMILES. This

string serves as a unique structural identifier for duplicate removal and yields the molecular set used for image generation (Fig. 2).

Combinatorial molecular-graph enumeration can produce a large challenge space. Under fixed constraints on heavy atoms, elements and bonding, the number of candidates is finite but increases rapidly with scope. GDB-13 enumerated about 977 million molecules containing up to 13 heavy atoms under chemical stability and synthetic feasibility rules. GDB-17 extended the limit to 17 heavy atoms and contains about 166.4 billion structures.[42,43] These databases illustrate the structural diversity available in small-molecule chemical space, rather than the size of the present ChemCAPTCHA library.

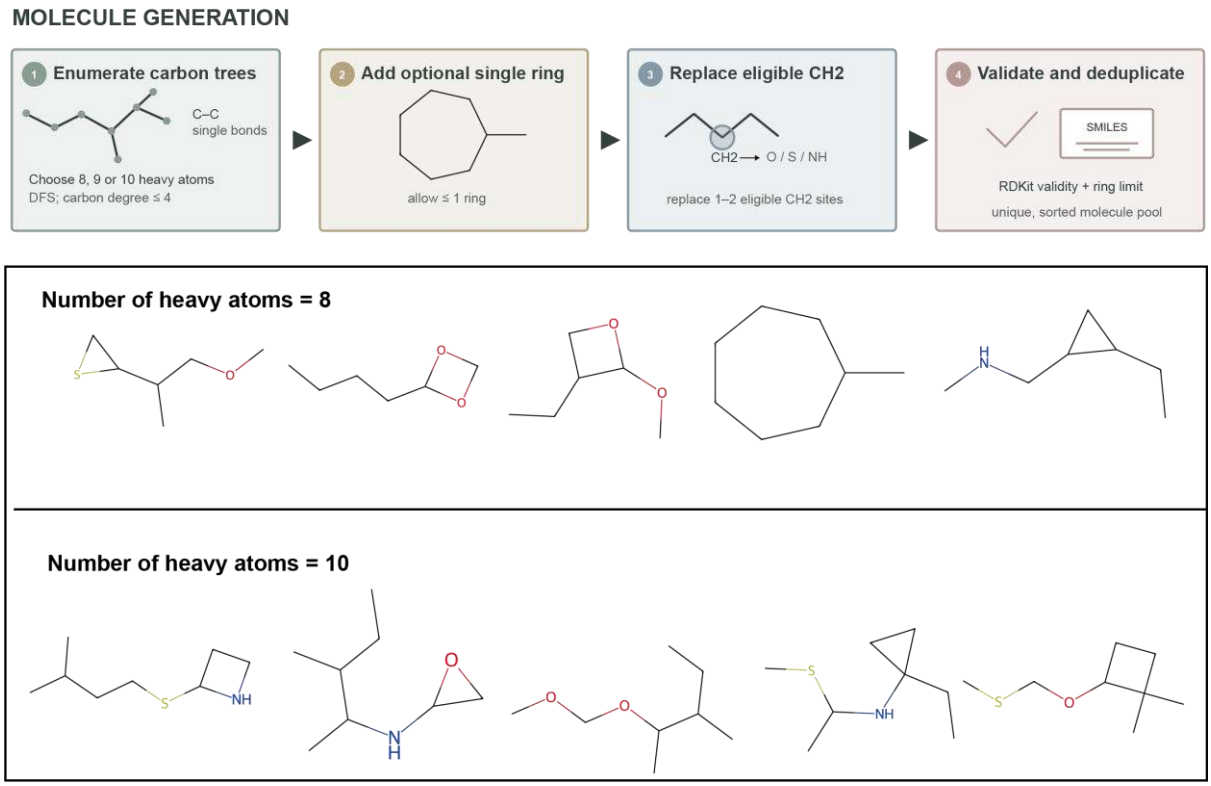


Fig. 2 | Molecular-generation workflow and representative output structures.

Visual perturbation of molecular prompts

ChemCAPTCHA applies visual perturbations to molecular prompts to increase the difficulty of recovering molecular graphs by optical chemical structure recognition. The perturbations comprise short bond-like line segments, triangles and local geometric deformation.

The system first uses RDKit to render the target molecule as a 720 × 420-pixel two-dimensional structure. It then applies bidirectional sinusoidal warping at the pixel level. For image coordinates (x, y), the sampling positions are periodically displaced in the horizontal and vertical directions (Fig. 3a):

$$x' = x + A\sin(\frac{2\pi y}{\lambda} + \varphi_x)$$

$$y' = y + A\sin(\frac{2\pi x}{\lambda} + \varphi_y)$$

Here, A = 18 denotes the wave amplitude, λ = 300 the wavelength, and φx and φy phases generated from a random seed. The transformation introduces continuous wave-like deformation into bonds and element labels while preserving the overall image structure.

After coordinate warping, the system adds a random intensity offset from −6 to 6 to each pixel and applies a Gaussian blur with a radius of 0.35. It then overlays 14 random greyscale line segments and eight hollow equilateral triangles (Fig. 3b). Segment endpoints, greyscale values and widths are sampled randomly, with opacity fixed at 0.10. Triangle side lengths range from 24 to 72 pixels, with random positions and rotations and an opacity of 0.16. Because the perturbations are applied after rendering, they do not alter the molecular graph, canonical SMILES or verification rules. A fixed seed reproduces the perturbation, whereas different seeds generate distinct challenge images.Separating image processing from the chemical structure prevents perturbations from changing bond orders or element labels.

We evaluated 1,000 molecular structure images with MolScribe. Recognition accuracy was 96.4% for unperturbed diagrams and 0.0% after the specified perturbations. Representative recognition results are shown in Fig. 3c.

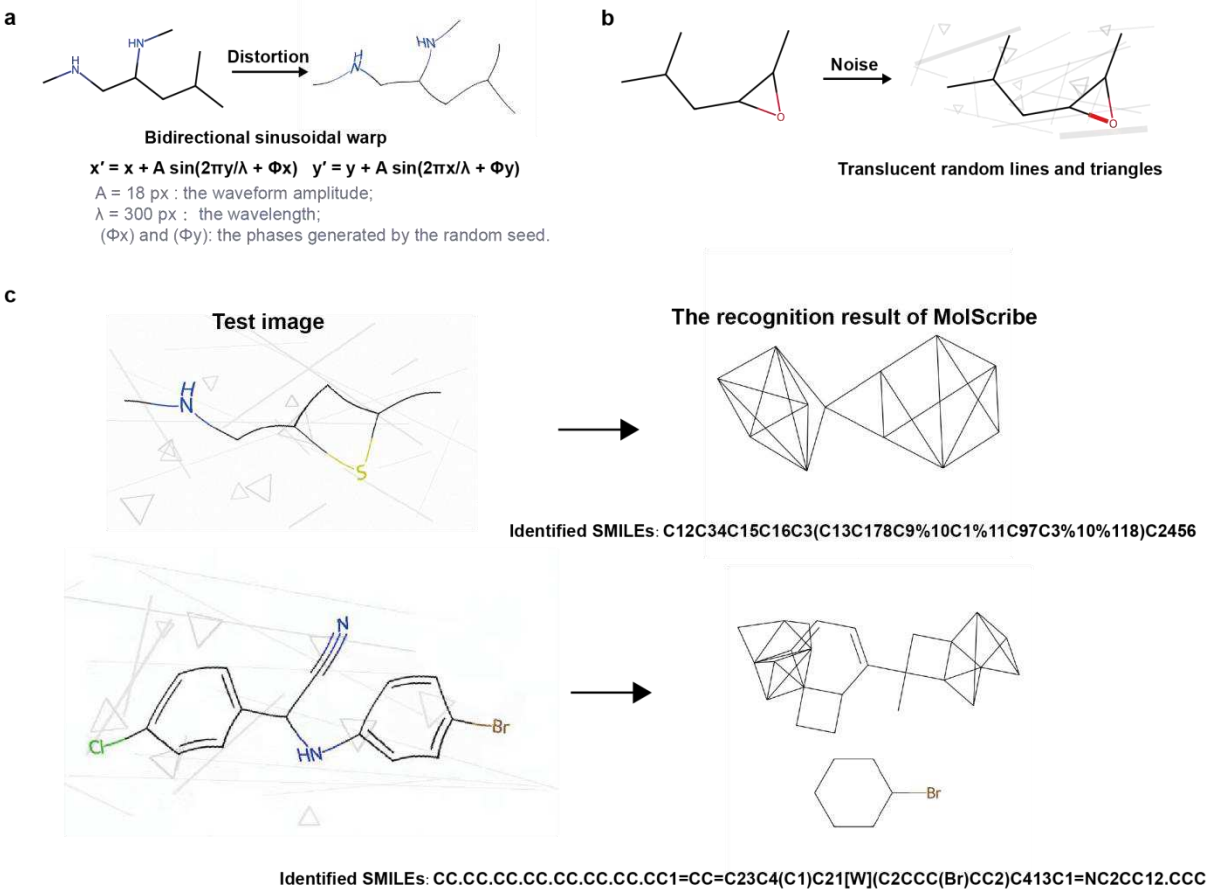


Fig. 3 | Visual perturbations and MolScribe recognition outcomes. a, Sinusoidal image deformation. b, Background lines and triangles added after warping. c, MolScribe recognition outcomes for perturbed molecular structure images.

Constrained Ketcher editor and manual-drawing requirements

ChemCAPTCHA combines a chemical-knowledge task with interaction analysis to provide complementary evidence for distinguishing ordinary users from automated programs. Users manually reproduce the target in a chemical structure editor. Each response is represented as a temporal sequence of pointer movements, tool selections, structural edits and inter-event

intervals. The final structure determines chemical correctness, while the drawing process provides supplementary evidence for interaction-risk assessment.

The structure-drawing module is based on the Ketcher chemical structure editor. External input routes that could bypass manual drawing are restricted to preserve task validity and the analytical value of recorded trajectories. The system hides or disables file opening, structure import, clipboard pasting, text and file drag-and-drop, and public structure-injection interfaces. It also intercepts Ctrl+O, Ctrl/Command+V and Shift+Insert. These restrictions prevent one-step entry of complete structures encoded as SMILES, Molfile or other Ketcher-compatible formats. Users must therefore recognise and construct the target structure manually (Fig. 4).

These restrictions do not alter Ketcher's core editing functions. Users can still draw molecules with atom, bond, ring, selection, erase and undo tools. Controlled internal interfaces remain available for clearing the canvas and replaying behaviour, but structure injection is unavailable during normal responses. Attempts to open files, paste content, drag and drop data, or inject structures through an application programming interface are recorded as import_blocked events. These events provide supplementary risk signals for violations of the manual-drawing requirement.

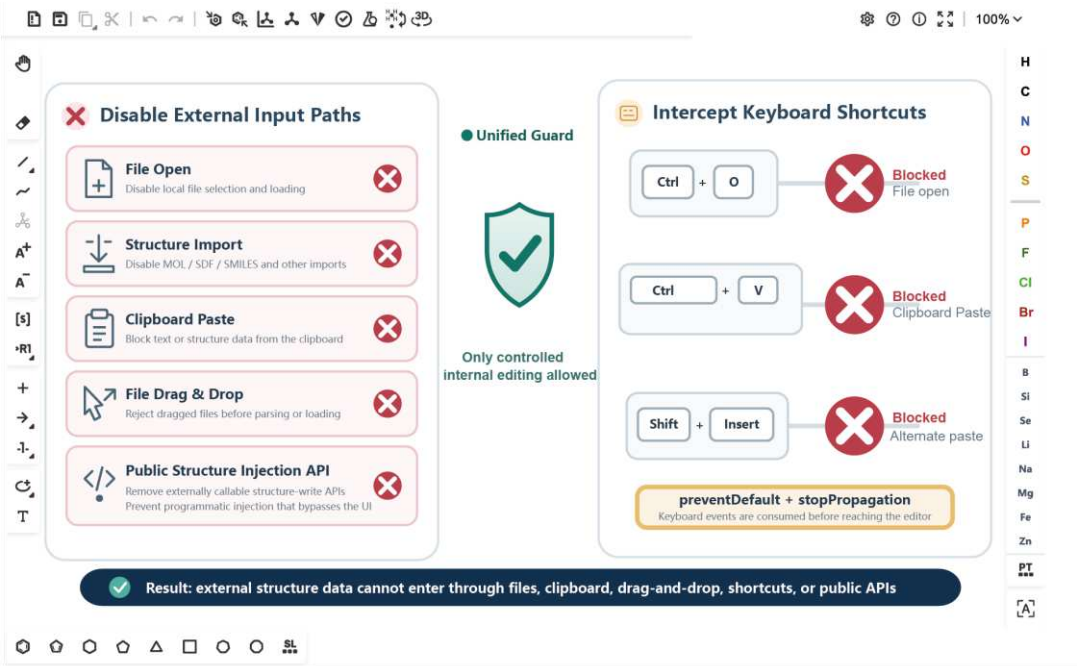


Fig. 4 | Constrained Ketcher editor and blocked structure-import routes.

Within-session randomisation

We introduced two forms of within-session randomisation into Ketcher to increase the adaptation cost for fixed-coordinate clicking, predefined action sequences and automated scripts. These mechanisms randomize element-toolbar order and selected element editing outcomes. Neither mechanism changes the target structure. Instead, both add uncertainty to the drawing process and require users to respond to the current interface state.

The first mechanism randomizes element-toolbar order. The right-hand element toolbar is divided into three groups: H/C/N/O/S, P/F/Cl/Br/I, and B/Si/Se/Li/Na/Mg/Fe/Zn. After each element selection, the system independently applies a Fisher – Yates shuffle within all three groups, but elements cannot move between groups. Random values are generated through the browser cryptographic interface, crypto.getRandomValues(), to reduce predictability from conventional pseudorandom-number generators. The mechanism preserves the available elements, button functions and native visual style, changing only their positions within each

group. It thereby reduces the reusability of automated scripts that depend on fixed screen coordinates or click sequences (Fig. 5a).

The second mechanism randomises selected element editing outcomes (Fig. 5b). It operates after a user selects a non-carbon element from the right-hand toolbar. Eligible actions either replace an existing carbon atom or add an atom by dragging a new bond from an existing carbon. The system uses a dynamic probability to replace the intended edit with another candidate element. If n denotes consecutive eligible operations without a trigger, the trigger probability is

$$p_n = \min(0.40 + 0.1n, 1.0).$$

The initial probability is 40% and increases by 10 percentage points after each non-triggering operation, to a maximum of 100%.

The randomised outcome is selected from all non-carbon elements in the right-hand toolbar, excluding the current selection and the previous randomised outcome. Candidates are maintained in a shuffled bag to limit short-term repetition. The outcome appears immediately on the canvas without changing the active element tool, allowing users to inspect and correct the structure (Fig. 5b).

The mechanisms introduce uncertainty at complementary levels. Element-toolbar randomisation disrupts fixed-coordinate clicking, while element editing randomisation requires continuous inspection and correction. Target-structure correctness remains the primary verification criterion. Response time and corrective behaviour during randomised events are included among the supplementary behavioural features.

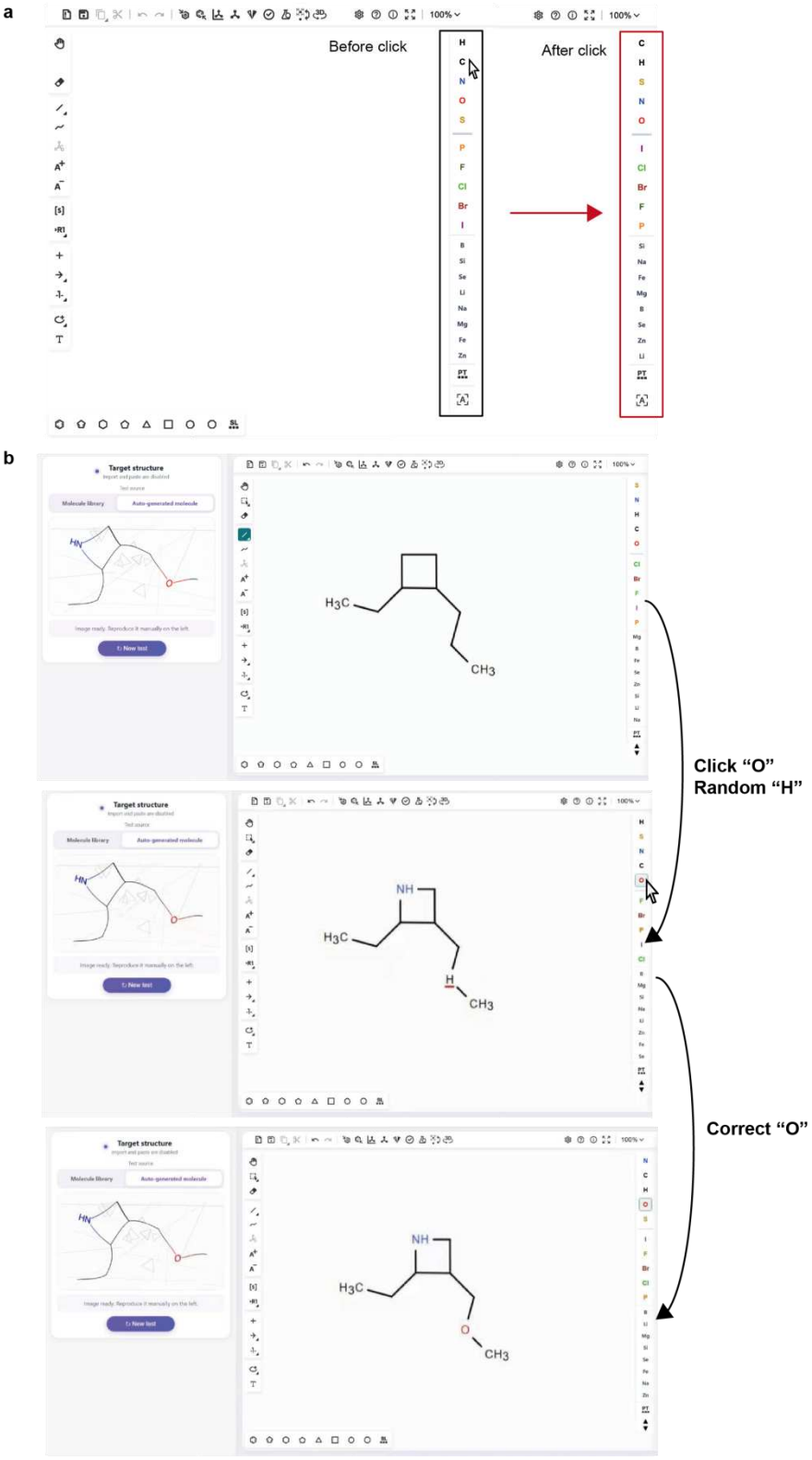


Fig. 5 | Two forms of within-session randomisation. a, Element-toolbar order. b, Element editing outcomes.

Randomly timed click task

ChemCAPTCHA adds temporal uncertainty to manual structure drawing through a randomly timed click task. At an unpredictable point during construction, a prompt identifies an interface button and requires the user to click it within 5 s (Fig. 6). The interruption is independent of the sequence of molecular edits, so its timing cannot be embedded reliably in a fixed action script.

A solver must therefore monitor the interface while drawing, detect the prompt, locate the indicated button and respond before the deadline. Automated agents must coordinate prompt detection with structure editing, while human-solving services must maintain attention throughout the response. The task does not alter the target molecule or the final canonical-structure comparison. It increases variation between sessions and raises the operational cost of automated or outsourced solving.

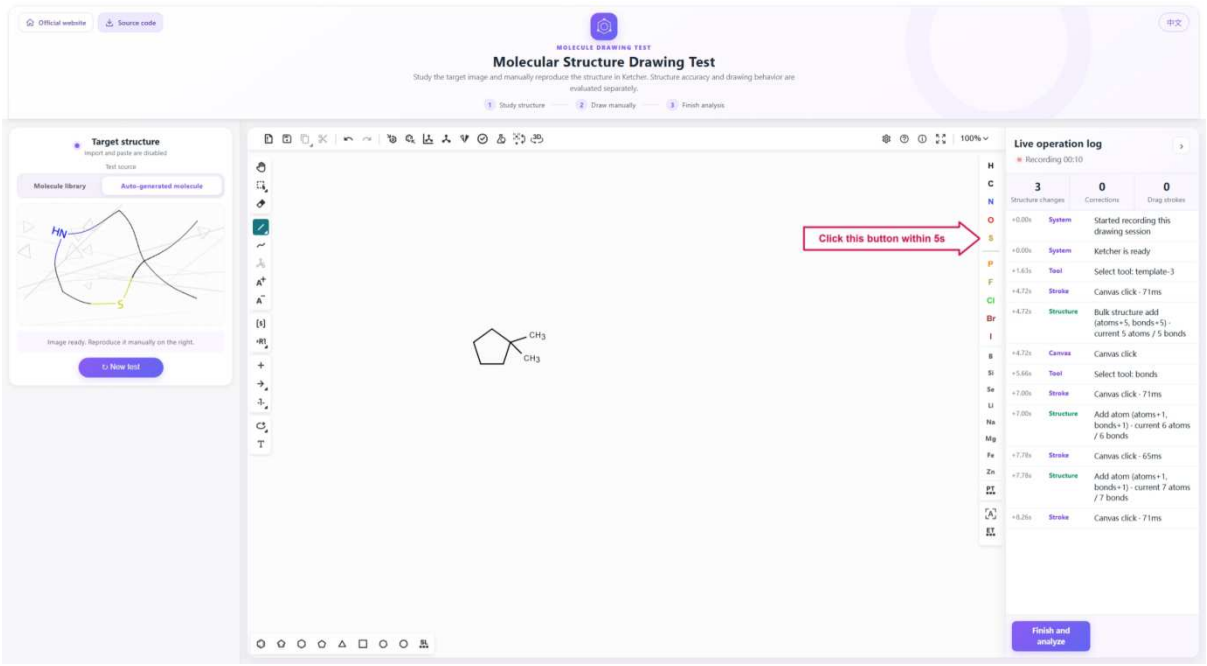


Fig. 6 | Randomly timed click task during manual structure drawing.

Structure verification and combined decision process

After submission, the server parses both the target and response as molecular structures. It generates canonical SMILES that retain bond orders and charges information. The comparison is recorded as a match, mismatch or invalid response. Structural verification and behavioural risk are reported separately. The former determines whether the submitted molecule satisfies the challenge, whereas the latter provides supplementary evidence about the response process.

Behavioural scoring covers structure construction, pointer trajectories, interaction timing and input integrity (Fig. S1). Construction risk increases when a final molecule has at least ten atoms, the response takes less than 5 s and the canvas receives fewer than four operations. Trajectory risk increases when repeated structural changes lack corresponding canvas trajectories. It also increases when straight paths constitute at least 0.8 of four or more drags and the speed coefficient of variation is below 0.1. Curved trajectories with variable speed can reduce the score. The timing dimension evaluates intervals between canvas operations, and an interval exceeding 2 s between structural edits can reduce risk. Input-integrity risk increases after blocked import, paste, drag-and-drop or interface-injection attempts, or when raw evidence duplicates a previous report. A structural change within 3.5 s after undo, redo or deletion is recorded as a correction and can reduce the score.

The four dimensions are summed to produce a heuristic behavioural risk score from 0 to 100. Evidence is classified as insufficient when the final structure is empty, fewer than two structural changes occur, or the canvas receives fewer than three operations. Responses shorter than 2 s are also considered insufficient. Remaining sessions are categorised at thresholds of 30 and 60 as ‘more human-like’, ‘inconclusive evidence’ or ‘suspected automation’ (Fig. 7 and Fig. S2).

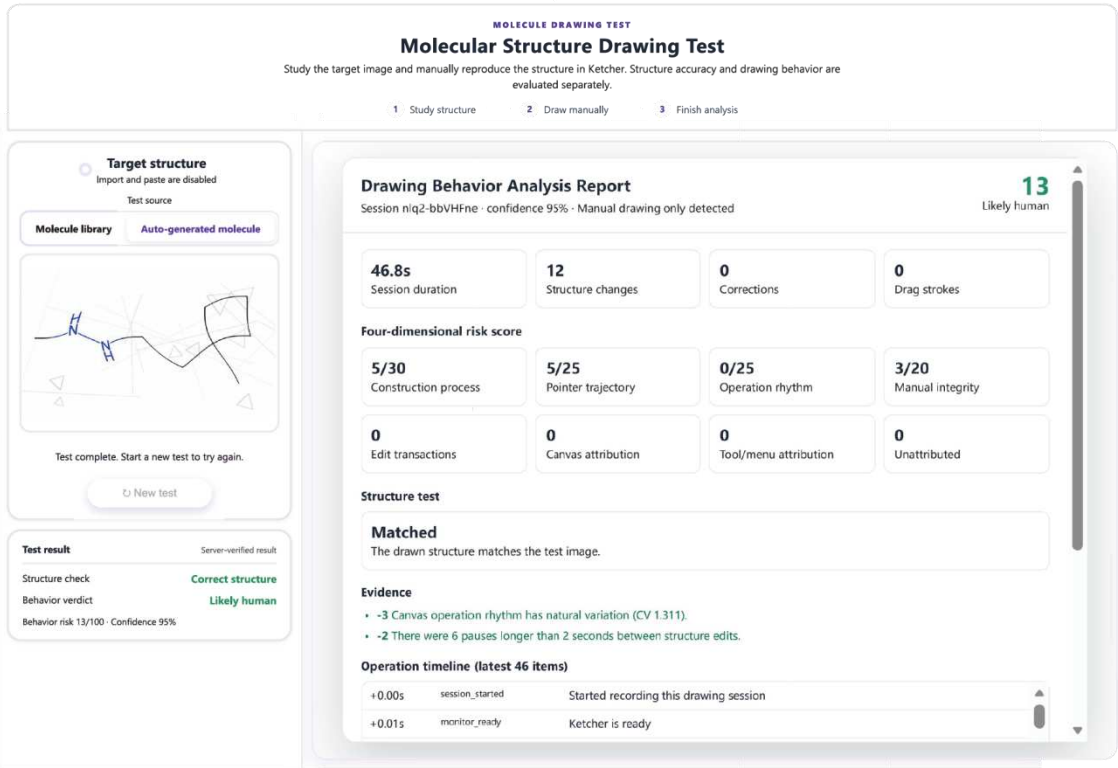


Fig. 7 | Structural-verification result and behavioural risk score.

Discussion

ChemCAPTCHA integrates molecular generation, visual perturbation, constrained Ketcher drawing, behaviour recording, two forms of within-session randomisation and a randomly timed click task. The task extends static prompt recognition to structure interpretation, manual construction and dynamic correction. It primarily targets automated programs and low-skill human-solving services. Blocking structure-import routes forces automated solvers to interact with the editor, while interface and editing randomisation disrupt fixed coordinates and predefined action sequences. The time-limited click task adds an unpredictable interruption that requires continuous monitoring during drawing. Human-solving services must additionally recognise and draw chemical structures. These mechanisms are intended to raise the cost of bulk solving, but their effects must be quantified against attacker capability, attack cost and legitimate-user burden.

The system separates answer correctness from a supplementary risk score derived from trajectories, tool selections, timing, corrections and anomalous input events. Structural verification establishes whether the response matches the target, whereas behavioural analysis characterises the response process. Multimodal language-model attacks must coordinate structure recognition, interface manipulation and dynamic correction. This design may increase their operational cost, but the magnitude of that increase has not yet been measured.

ChemCAPTCHA has several limitations. First, the current molecular generator is constrained by heavy-atom count, heteroatom-substitution rules and ring count. The broad potential chemical space therefore does not represent the challenge library available in the present implementation. Future work should quantify candidate counts, duplication rates and structural-complexity distributions across generation settings. These measurements could guide expansion of the library while controlling the manual-drawing burden.

Second, domain-knowledge requirements, randomised perturbations and the time-limited click task may burden legitimate users. Users can differ in chemical training, professional role, device conditions, structure-recognition ability and editing efficiency.

Conclusion

ChemCAPTCHA's central contribution is the integration of chemical domain knowledge into access verification for chemical data. It requires users to interpret and manually reconstruct molecular structures, while interaction analysis provides a supplementary behavioural risk signal. A constrained editor, within-session randomisation and a time-limited click task increase the adaptation required from fixed scripts and low-skill human solvers. However, chemically trained solvers and adaptive systems that combine structure recognition with interface control may still complete the challenge. ChemCAPTCHA should therefore

complement authentication, authorisation, rate limiting and anomalous-query detection rather than serve as a standalone defence. Attack evaluation, ablation experiments and user studies are required to establish its security benefits and usability limits.

Acknowledgements

We thank Mr. Zhong Tang for his investment in InsightChemy Co., Ltd.

Authors' contributions

All authors contributed to the conception and development of the project. Z.D.: conceptualisation (lead), data curation, methodology, software, writing (original draft, lead), and writing (review and editing, lead). P.J. and H.Z.: conceptualisation, validation and visualisation. J.Z.: figure preparation and writing. Y.T.: conceptualisation, funding acquisition, project administration, resources, supervision, writing (original draft), and writing (review and editing). All authors reviewed and approved the submitted manuscript.

Funding

Y.T. was supported by the 2024 'Haiyou Project' Industry Leading Talent Program.

Data availability

All code and processed data required to reproduce the analyses reported in this study are publicly available on the company's website https://insightchemy.com/safety/molecule-captcha/.

Competing interests

The authors declare no competing interests.

References

1. Goodman J (2009) Computer Software Review: Reaxys. J Chem Inf Model 49:2897–2898. https://doi.org/10.1021/ci900437n
2. Lawson AJ, Swienty-Busch J, Géoui T, Evans D (2014) The Making of Reaxys: Towards Unobstructed Access to Relevant Chemistry Information. In: McEwen LR, Buntrock RE (eds) The Future of the History of Chemical Information. ACS Symposium Series, vol 1164. American Chemical Society, Washington, DC, pp 127–148. https://doi.org/10.1021/bk-2014-1164.ch008
3. Groom CR, Bruno IJ, Lightfoot MP, Ward SC (2016) The Cambridge Structural Database. Acta Crystallogr B Struct Sci Cryst Eng Mater 72:171–179. https://doi.org/10.1107/S2052520616003954
4. Sabe VT, Ntombela T, Jhamba LA, et al (2021) Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review. Eur J Med Chem 224:113705. https://doi.org/10.1016/j.ejmech.2021.113705
5. Kim S, Chen J, Cheng T, et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102–D1109. https://doi.org/10.1093/nar/gky1033
6. Kim S, Chen J, Cheng T, et al (2023) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380. https://doi.org/10.1093/nar/gkac956
7. Kim S, Chen J, Cheng T, et al (2025) PubChem 2025 update. Nucleic Acids Res 53:D1516–D1525. https://doi.org/10.1093/nar/gkae1059
8. Irwin JJ, Tang KG, Young J, et al (2020) ZINC20: A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model 60:6065–6073. https://doi.org/10.1021/acs.jcim.0c00675

9. Kearnes SM, Maser MR, Wleklinski M, et al (2021) The Open Reaction Database. J Am Chem Soc 143:18820–18826. https://doi.org/10.1021/jacs.1c09820
10. Bird CL, Willoughby C, Frey JG (2013) Laboratory notebooks in the digital era: the role of ELNs in record keeping for chemistry and other sciences. Chem Soc Rev 42:8157–8175. https://doi.org/10.1039/C3CS60122F
11. Walsh E, Cho I (2013) Using Evernote as an Electronic Lab Notebook in a Translational Science Laboratory. SLAS Technol 18:229–234. https://doi.org/10.1177/2211068212471834
12. Schneider N, Lowe DM, Sayle RA, et al (2016) Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. J Med Chem 59:4385–4402. https://doi.org/10.1021/acs.jmedchem.6b00153
13. Schneider N, Stiefl N, Landrum GA (2016) What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. J Chem Inf Model 56:2336–2346. https://doi.org/10.1021/acs.jcim.6b00564
14. Bai F, Ritter A, Madrid P, et al (2022) SynKB: Semantic Search for Synthetic Procedures. arXiv:2208.07400 [Preprint]. https://doi.org/10.48550/arXiv.2208.07400
15. Schwaller P, Gaudin T, Lányi D, et al (2018) "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci 9:6091–6098. https://doi.org/10.1039/C8SC02339E
16. Jin W, Coley CW, Barzilay R, Jaakkola T (2017) Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html

17. Dai H, Li C, Coley CW, et al (2019) Retrosynthesis Prediction with Conditional Graph Logic Network. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html
18. Krotov V, Johnson L, Silva L (2020) Legality and Ethics of Web Scraping. Commun Assoc Inf Syst 47:539–563. https://doi.org/10.17705/1CAIS.04724
19. Thelwall M, Stuart D (2006) Web crawling ethics revisited: Cost, privacy, and denial of service. J Am Soc Inf Sci Technol 57:1771–1779. https://doi.org/10.1002/asi.20388
20. Guerar M, Verderame L, Migliardi M, et al (2022) Gotta CAPTCHA 'Em All: A Survey of 20 Years of the Human-or-computer Dilemma. ACM Comput Surv 54:1–33. https://doi.org/10.1145/3477142
21. Bursztein E, Martin M, Mitchell J (2011) Text-based CAPTCHA strengths and weaknesses. In: Proceedings of the 18th ACM Conference on Computer and Communications Security. ACM, Chicago, IL, pp 125–138. https://doi.org/10.1145/2046707.2046724
22. Tang M, Gao H, Zhang Y, et al (2018) Research on Deep Learning Techniques in Breaking Text-Based Captchas and Designing Image-Based Captcha. IEEE Trans Inf Forensics Secur 13:2522–2537. https://doi.org/10.1109/TIFS.2018.2821096
23. Weng H, Zhao B, Ji S, et al (2019) Towards understanding the security of modern image captchas and underground captcha-solving services. Big Data Min Anal 2:118–144. https://doi.org/10.26599/BDMA.2019.9020001
24. Gao Y, Gao H, Luo S, et al (2021) Research on the Security of Visual Reasoning CAPTCHA. In: 30th USENIX Security Symposium (USENIX Security 21). USENIX

Association, pp 3291–3308. https://www.usenix.org/conference/usenixsecurity21/presentation/gao

25. Bigham JP, Cavender AC (2009) Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Boston, MA, pp 1829–1838. https://doi.org/10.1145/1518701.1518983
26. Bursztein E, Beauxis R, Paskov H, et al (2011) The Failure of Noise-Based Non-continuous Audio Captchas. In: 2011 IEEE Symposium on Security and Privacy. IEEE, Oakland, CA, pp 19–31. https://doi.org/10.1109/SP.2011.14
27. Chang G, Gao H, Pei G, et al (2024) The robustness of behavior-verification-based slider CAPTCHAs. J Inf Secur Appl 81:103711. https://doi.org/10.1016/j.jisa.2024.103711
28. Bursztein E, Bethard S, Fabry C, et al (2010) How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation. In: 2010 IEEE Symposium on Security and Privacy. IEEE, Oakland, CA, pp 399–413. https://doi.org/10.1109/SP.2010.31
29. Searles A, Nakatsuka Y, Ozturk E, et al (2023) An Empirical Study & Evaluation of Modern CAPTCHAs. In: 32nd USENIX Security Symposium (USENIX Security 23). USENIX Association, Anaheim, CA, pp 3081–3097. https://www.usenix.org/conference/usenixsecurity23/presentation/searles
30. Luo Y, Li Z, Liu J, et al (2025) Open CaptchaWorld: A Comprehensive Web-based Platform for Testing and Benchmarking Multimodal LLM Agents. Adv Neural Inf Process Syst 38:99835–99860. https://doi.org/10.52202/085713-3006
31. Teoh X, Lin Y, Li S, et al (2025) Are CAPTCHAs Still Bot-hard? Generalized Visual CAPTCHA Solving with Agentic Vision Language Model. In: 34th USENIX Security

Symposium (USENIX Security 25). USENIX Association, Seattle, WA, pp 3747–3766. https://www.usenix.org/conference/usenixsecurity25/presentation/teoh

32. Wang J, Zhu C, Zhou Y, et al (2026) COGNITION: From Evaluation to Defense against Multimodal LLM CAPTCHA Solvers. arXiv:2512.02318 [Preprint]. https://doi.org/10.48550/arXiv.2512.02318
33. leafLeaf9. chiral-carbon-captcha: Chiral carbon atom CAPTCHA API. GitHub. https://github.com/leafLeaf9/chiral-carbon-captcha. Accessed 4 Oct 2026
34. CGT Software (2025) CAPTCHA service: making the simple complex, a new experience [in Chinese]. https://www.cgtsoft.com/archives/25/. Accessed 4 Oct 2026
35. Rajan K, Zielesny A, Steinbeck C (2020) DECIMER: towards deep learning for chemical image recognition. J Cheminform 12:65. https://doi.org/10.1186/s13321-020-00469-w
36. Rajan K, Brinkhaus HO, Agea MI, et al (2023) DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nat Commun 14:5045. https://doi.org/10.1038/s41467-023-40782-0
37. Qian Y, Guo J, Tu Z, et al (2023) MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. J Chem Inf Model 63:1925–1934. https://doi.org/10.1021/acs.jcim.2c01480
38. Xu Y, Xiao J, Chou C-H, et al (2022) MolMiner: You Only Look Once for Chemical Structure Recognition. J Chem Inf Model 62:5321–5328. https://doi.org/10.1021/acs.jcim.2c00733
39. Morin L, Danelljan M, Agea MI, et al (2023) MolGrapher: Graph-based Visual Recognition of Chemical Structures. In: 2023 IEEE/CVF International Conference on

Computer Vision (ICCV). IEEE, Paris, pp 19495–19504. https://doi.org/10.1109/ICCV51070.2023.01791

40. Karulin B, Kozhevnikov M (2011) Ketcher: web-based chemical structure editor. J Cheminform 3(Suppl 1):P3. https://doi.org/10.1186/1758-2946-3-S1-P3
41. Landrum G (2016) RDKit: Open-source cheminformatics software. https://www.rdkit.org/. Accessed 4 Oct 2026
42. Blum LC, Reymond J-L (2009) 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. J Am Chem Soc 131:8732–8733. https://doi.org/10.1021/ja902302h
43. Reymond J-L (2015) The Chemical Space Project. Acc Chem Res 48:722–730. https://doi.org/10.1021/ar500432k

Supplementary Files

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

- SI.docx