Journal of Cheminformatics

DATA NOTE Open Access

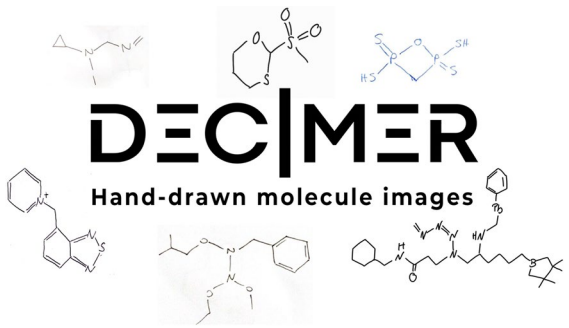
DECIMER—hand-drawn molecule images dataset

Henning Otto Brinkhaus[1], Achim Zielesny[2], Christoph Steinbeck[1] and Kohulan Rajan[1*]

Abstract

The translation of images of chemical structures into machine-readable representations of the depicted molecules is known as optical chemical structure recognition (OCSR). There has been a lot of progress over the last three decades in this field, but the development of systems for the recognition of complex hand-drawn structure depictions is still at the beginning. Currently, there is no data for the systematic evaluation of OCSR methods on hand-drawn structures available. Here we present *DECIMER — Hand-drawn molecule images*, a standardised, openly available benchmark dataset of 5088 hand-drawn depictions of diversely picked chemical structures. Every structure depiction in the dataset is mapped to a machine-readable representation of the underlying molecule. The dataset is openly available and published under the CC-BY 4.0 licence which applies very few limitations. We hope that it will contribute to the further development of the field.

Graphical Abstract



Objective

Most chemical information is published in text and images in the primary scientific literature. The automated conversion of these unstructured, human-readable data formats into structured, machine-readable representations is essential to make the information available in publicly accessible databases. The reliable extraction of information from the depictions of the chemical structures is an ongoing challenge that still has not been fully solved yet. Chemical structure depictions are converted into computer-readable representations using optical chemical structure recognition (OCSR) systems [1].

The field of OCSR has developed significantly over the last 30 years. Most OCSR tools follow a hard-coded set of rules to assemble the underlying molecule based on the

*Correspondence: kohulan.rajan@uni-jena.de

[1] Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07743 Jena, Germany
Full list of author information is available at the end of the article

elements in the vectorised image [2–11]. By 2020 several deep learning-based solutions are available [12–18].

In order to evaluate the performance of the available OCSR tools, realistic benchmark datasets are necessary. At present, there are four real-world datasets available [1, 9, 19] that contain chemical structure depictions that were collected and curated from publications and patents. The evaluation of the performance on realistic data is crucial to demonstrate whether the tools are robust enough to be used in an automated chemical literature mining process.

The resolution of hand-drawn chemical structures is a more challenging task than the resolution of automatically generated depictions. In addition to the varying depiction features which are present anyway, the individual, unique way of drawing the structure adds an increased level of complexity. In 2021, the deep learning-based OCSR tool ChemPix [15] demonstrated its capability to interpret simple hand-drawn hydrocarbon structures with high accuracy. There also are a few closed-source methods and commercial systems available that claim to be capable of resolving hand-drawn chemical structures [20–22]. The authors of the deep-learning-based OCSR tool img2mol demonstrated the capability of their tool to recognise some hand-drawn chemical structures that they had picked themselves and noted the lack of a standardised benchmark set [14].

With the development of more OCSR tools that focus on the resolution of hand-drawn chemical structure depictions, there is a need for a standardised dataset to evaluate their performance. Here we present *DECIMER — Hand-drawn molecule images*, a set of 5088 hand-drawn chemical structures depictions. Every image is mapped to a machine-readable representation of the underlying molecule. The diversely picked molecules represent a wide variety of small molecules. The dataset was created to facilitate the ongoing development in the field of OCSR and is openly accessible.

Data description

The dataset consists of 5088 PNG images of unique hand-drawn chemical structure depictions (Fig. 1) which are mapped to their corresponding SMILES [23] string as well as an SD file. The structures have been drawn by 24 volunteers from the Westphalian University of Applied Sciences, Campus Recklinghausen, Germany, who have graciously offered to use their free time to contribute to the generation of this dataset.

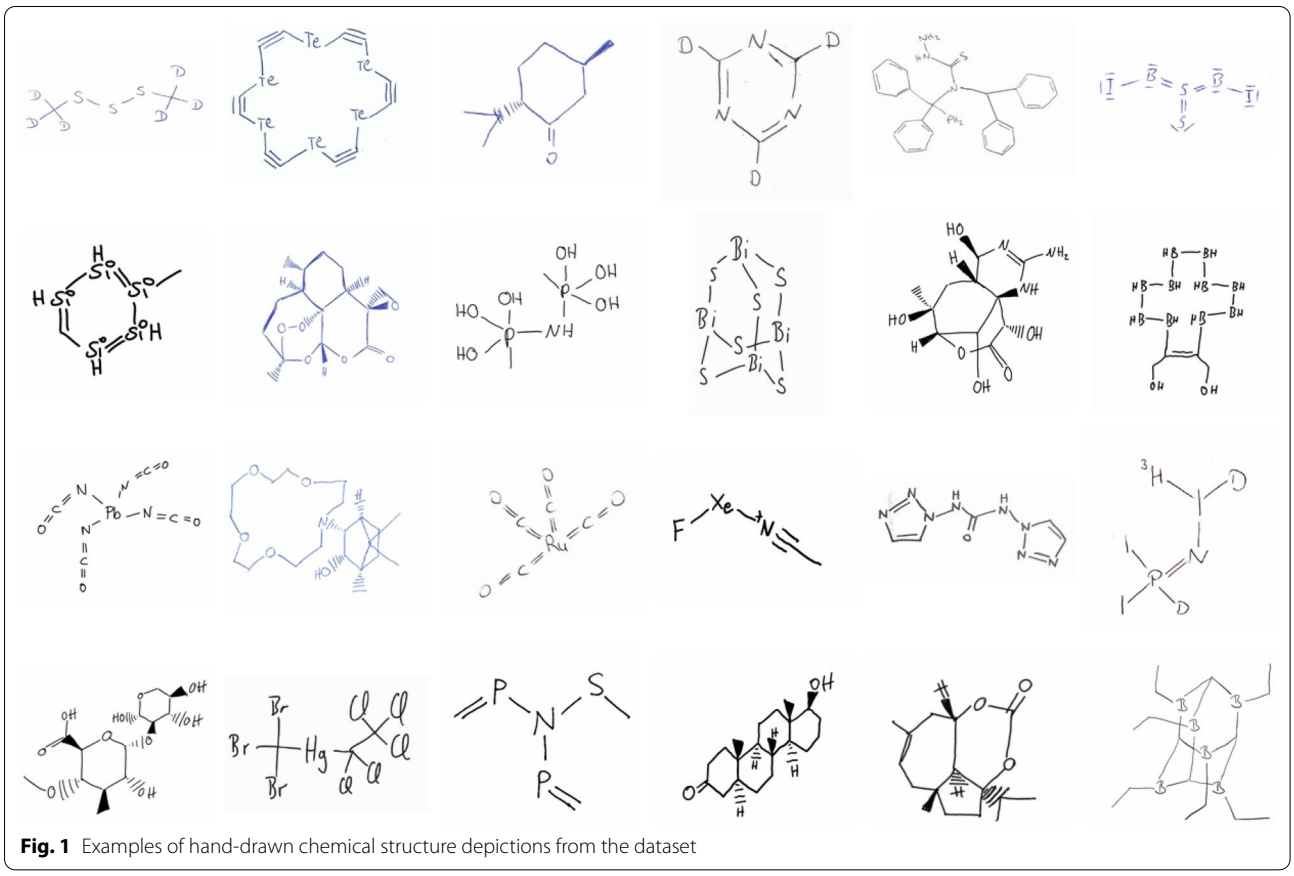
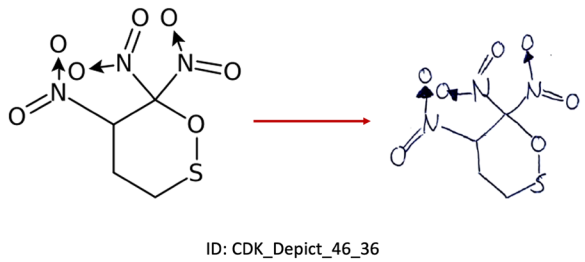
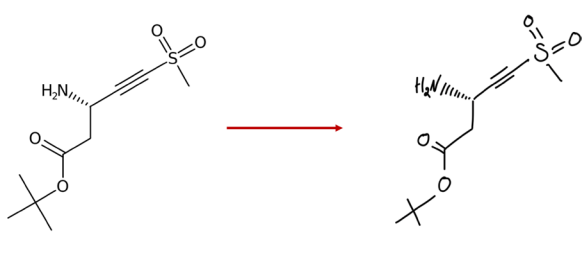


Fig. 1 Examples of hand-drawn chemical structure depictions from the dataset



ID: CDK_Depict_46_36

Fig. 2 A chemical structure depiction generated by CDK, sketched on a sheet of paper and scanned as an image file



ID: CDK_Depict_45_18

Fig. 3 A chemical structure depiction generated by CDK, sketched on a tablet and saved as an image file

The molecules have been picked from all structures in PubChem [24] using RDKit's implementation of the MaxMin algorithm [25] based on Morgan fingerprints [26] to ensure a diverse coverage of the chemical space. The only filtering rule that has been applied is a molecular weight maximum of 1500 Da. As a consequence, features like stereochemical information, charged groups as well as different types of isotopes are present in the dataset.

There are two categories of images:

Drawn on a piece of white paper and scanned (Fig. 2)
Drawn using a mobile device or tablet and directly saved as an image (Fig. 3).

Curation

In total, 6000 diverse molecules were selected from PubChem using RDKit's implementation of the MaxMin algorithm based on Morgan fingerprints. Subsequently, CDK Depict [27], a structure depiction generator based on the Chemistry Development Kit (CDK) [28], was used to create production-quality 2D images in batches. Each batch of images was then converted into PDF files and they were distributed among the volunteers. Using the chemical structure depictions generated by CDK as a visual template, each volunteer drew the structures on a piece of paper using a black or blue pen or on their tablet using an input device.

Each volunteer sent back the scanned images or the images generated using their device after completing a batch. The curators reviewed the drawings, manually confirmed the correctness of the molecules, cropped the scanned images and stored them in separate image files. As part of the curation, structures that weren't correct due to human error were discarded. A total of 568 images out of 6000 were rejected due to issues with the depicted structure. Another 344 structures were not returned by the volunteers. This resulted in the final dataset of 5088 images in total.

An identifier was assigned to each image, and the same identifier was used to label the SD file which was generated using the CDK. Additionally, the dataset contains a file containing a table of the identifiers and corresponding SMILES representations.

FAIR-ification

The following steps were taken in order to make the dataset findable, accessible, interoperable and reusable (FAIR) [29]. The dataset was deposited in a publicly accessible data repository, in this case, Zenodo. This ensures that the dataset is easily findable. Furthermore, Zenodo provides a digital object identifier (DOI) that can be used to locate the dataset and it can also easily be integrated into Github as well. With Zenodo being an open, public repository, the dataset can be accessed from any part of the globe. To make it as interoperable as possible, the generated images use PNG as the final image format, which can be used across a variety of operating systems. Additionally, SMILES and SDF are representations of chemical structures which can be read by every cheminformatics toolkit. The dataset has been published under the CC-BY 4.0 licence. This licence includes that every user can redistribute or change the data as much as they want as long as they refer to the original authors when publishing results based on it. It is possible to use the data for non-commercial or commercial purposes without further obligations.

Limitation

No restrictions or limitations apply to using and reusing the dataset. Everyone can use this dataset as a standard benchmark set for the evaluation of the performance of their OCSR tools. The dataset includes a wide range of chemical structures and represents a much larger chemical space. The structures were drawn by various individuals to ensure the diversity of drawing styles. The main limitation is caused by the molecular weight filter (< 1500Da) as it excludes certain molecules like big macrocycles, proteins or artificial polymers. Additionally, Markush structures are not represented.

Due to the limited number of images in this dataset, we do not recommend attempting to train a deep learning model using this dataset. We highly recommended using it exclusively for benchmarking instead of fitting the tools to the dataset.

Abbreviations

CDK: Chemistry development kit; CC: Creative commons; DOI: Digital object identifier; FAIR: Findable, accessible, interoperable, and reusable; OCSR: Optical chemical structure recognition; PDF: Portable document format; PNG: Portable network graphics; SDF: Structural data file; SDG: Structure diagram generator; SMILES: Simplified molecular-input line-entry system.

Acknowledgements

The authors would like to acknowledge the following students of the Westphalian University of Applied Sciences for their contribution to the project:Muammer Ates, Samuel Behr, Thilo Bredtmann, Michele Cassano, Saaruky Chanthirakanthan, Zeynep Dagtekin, Dustin Finke, Janina Gleißenberger, Safiye Sarah Kantar, Irina Krützner, Lilly Maddox, Fabian Matten, Timo van Meegdenburg, Alisa Muminovic, Paulina Paszko, Josefine Reiser, Maximilian Rottmann, Lisa Scharfenberg, Patricia Schoof, Betül Sevindik, Fatma Zehra Sevindik, Kaan Sipahi and Julia Zielinski.

Author contributions

KR and HOB initiated the project and curated the final dataset. CS and AZ conceived the project and supervised the work. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Carl-Zeiss-Foundation and by the German Research Foundation within the framework CRC1127 ChemBioSys.

Availability of data and materials

The dataset is openly available at ZENODO: https://doi.org/10.5281/zenodo.6456306.

Declarations

Competing interests

AZ is co-founder of GNWI—Gesellschaft für naturwissenschaftliche Informatik mbH, Dortmund, Germany.

Author details

[1]Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University Jena, Lessingstr. 8, 07743 Jena, Germany. [2]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, 45665 Recklinghausen, Germany.

Received: 20 April 2022 Accepted: 25 May 2022
Published online: 09 June 2022

References

1. Rajan K, Brinkhaus HO, Zielesny A, Steinbeck C (2020) A review of optical chemical structure recognition tools. J Cheminform 12:60
2. McDaniel JR, Balmuth JR (1992) Kekule: OCR-optical chemical (structure) recognition. J Chem Inf Comput Sci 32:373–378
3. Casey R, Boyer S, Healey P, Miller A, Oudot B, Zilles K (1993) Optical recognition of chemical graphics. In: Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93), pp 627–631
4. Ibison P, Jacquot M, Kam F, Neville AG, Simpson RW, Tonnelier C, Venczel T, Johnson AP (1993) Chemical literature data extraction: the CLiDE project. J Chem Inf Comput Sci 33:338–344
5. Valko AT, Johnson AP (2009) CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. J Chem Inf Model 49:780–787
6. Zimmermann M (2011) Chemical structure reconstruction with chemoCR. In: The Twentieth Text REtrieval conference (TREC 2011) Proceedings
7. Filippov IV, Nicklaus MC (2009) Optical structure recognition software to recover chemical information: OSRA, an open-source solution. J Chem Inf Model 49:740–743
8. Park J, Rosania GR, Shedden KA, Nguyen M, Lyu N, Saitou K (2009) Automated extraction of chemical structure information from digital raster images. Chem Cent J 3:4
9. Sadawi N (2009) Recognising chemical formulas from molecule depictions. In: Pre-proceedings of the 8th IAPR international workshop on graphics recognition (GREC 2009). pp 167–175
10. Tharatipyakul A, Numnark S, Wichadakul D, Ingsriswang S (2012) ChemEx: information extraction system for chemical data curation. BMC Bioinformatics 13(Suppl 17):S9
11. Beard EJ, Cole JM (2020) Chemschematicresolver: a toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. J Chem Inf Model 60:2059–2072
12. Rajan K, Zielesny A, Steinbeck C (2021) DECIMER 1.0: deep learning for chemical image recognition using transformers. J Cheminform 13:61
13. Rajan K, Zielesny A, Steinbeck C (2020) DECIMER: towards deep learning for chemical image recognition. J Cheminform 12:65
14. Clevert D-A, Le T, Winter R, Montanari F (2021) Img2Mol—accurate SMILES recognition from molecular graphical depictions. Chem Sci. https://doi.org/10.1039/D1SC01839F
15. Weir H, Thompson K, Woodward A, Choi B, Braun A, Martínez TJ (2021) ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning. Chem Sci 12:10622–10633
16. Oldenhof M, Arany A, Moreau Y, Simm J (2020) Chemgrapher: optical graph recognition of chemical compounds by deep learning. J Chem Inf Model 60:4506–4517
17. Zhang X-C, Yi J-C, Yang G-P, Wu C-K, Hou T-J, Cao D-S (2022) ABC-Net: a divide-and-conquer based deep learning architecture for SMILES recognition from molecular images. Brief Bioinform. https://doi.org/10.1093/bib/bbac033
18. Khokhlov I, Krasnov L, Fedorov MV, Sosnin S (2022) Image2SMILES: transformer-based molecular optical recognition engine. Chem Methods. https://doi.org/10.1002/cmtd.202100069
19. Osra (2022) https://sourceforge.net/p/osra/wiki/Validation/. Accessed 30 Mar 2022
20. Ouyang TY, Davis R (2007) Recognition of hand drawn chemical diagrams. AAAI 7:846–851
21. Ramel J-Y, Boissier G, Emptoz H (1999) Automatic reading of handwritten chemical formulas from a structural representation of the image. In: Proceedings of the 5th International Conference on Document Analysis and Recognition, ICDAR '99 (Cat. No.PR00318), pp 83–86
22. Vision Arcanum: InkToMolecule online. https://visionarcanum.com/ink2mol/. Accessed 30 Mar 2022
23. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
24. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388–D1395
25. Ashton M, Barnard J, Casset F, Charlton M, Downs G, Gorse D, Holliday J, Lahana R, Willett P (2002) Identification of diverse database subsets using property-based and fragment-based molecular descriptions. Quant struct-act relatsh 21:598–604
26. Morgan HL (1965) The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service. J Chem Doc 5:107–113
27. Mayfield J, Swain M, Willighagen E (2022) CDK Depict. In: GitHub. https://github.com/cdk/depict. Accessed 4 Mar 2022
28. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43:493–500
29. Jacobsen A, de Miranda AR, Juty N et al (2020) FAIR principles: Interpretations and implementation considerations. Data Intelligence 2:10–29

Publisher's Note