Testing programs for optical recognition of molecular structures from scientific articles

Ilya S. Tonkii [1] **Alina M. Luzanova** [1] **Rodion P. Golovinsky** [1] **Denis A. Chistyakov** [1] **Elena A. Shcherbakova** [1] **Timur A. Aliev** [1] **Dmitrii O. Shkil** [2] **Maksim V. Fedorov** [3] **Michael G. Medvedev** [1] **Ekaterina V. Skorb** [1]

[1]*Infochemistry Scientific Center, ITMO University, 9 Lomonosova Street, Saint-Petersburg 191002, Russia* [2]*Syntelly LLC,Bolshoy Boulevard 30, bld. 1, 121205 Moscow, Russia* [3]*Kharkevich Institute for Information Transmission Problems of RAS RU, Bolshoy Karetniy per. 19, Moscow, Russia 127994*. Correspondence to: Ekaterina V. Skorb skorb@itmo.ru.

1. Introduction

The advancement of optical recognition technologies for chemical structures has increased interest in automated analysis and processing of chemical data. These systems are applicable in pharmaceutical, biological, and chemical industries, enabling rapid analysis of large volumes of visual data and accelerating the discovery of new compounds [1, 2, 3]. This paper presents an extensive review and testing of current software solutions for the optical recognition of 2D chemical structure images, including both rule-based algorithms and neural network approaches.

2. Methods

To evaluate the performance of optical recognition tools, a unique dataset of 2405 molecule images was manually extracted from scientific publications. Each molecule was redrawn in ChemDraw to generate corresponding SMILES strings. Additionally, synthetic images were generated using the RDKit library [4] to compare algorithm performance on real and artificial data (Fig. 1).

The study tested two main categories of recognition tools: rule-based algorithms and neural networks. Rule-based tools included OSRA [5, 6], Imago [7, 8, ?], and Molvec [9], which rely on predefined image-processing rules to extract molecular structures. Neural network-based tools, such as ChemGrapher [10], DECIMER [11, 12, 13, 14], Img2Mol [15], Img2SMILES [16], MolScribe [17], SwinOCSR [18], MolMiner [19], AutoChemplete [20], utilized deep learning models trained on large datasets.

Performance was assessed by comparing recognized SMILES strings with ground truth values. Two key evaluation metrics were used: (1) SMILES string accuracy, which measures the exact match between recognized and expected SMILES, and (2) Tanimoto coefficient, which quantifies structural similarity between predicted and actual molecular fingerprints. Testing was conducted separately for molecules with and without chirality to assess the impact of stereochemical information on recognition accuracy.

To analyze performance variations, molecule images were categorized by atom count and structural complexity. Recognition difficulties were identified by testing challenging cases such as organometallic compounds, polycyclic structures, and molecules with unconventional representations. The study also examined the effects of image resolution and quality on recognition accuracy.

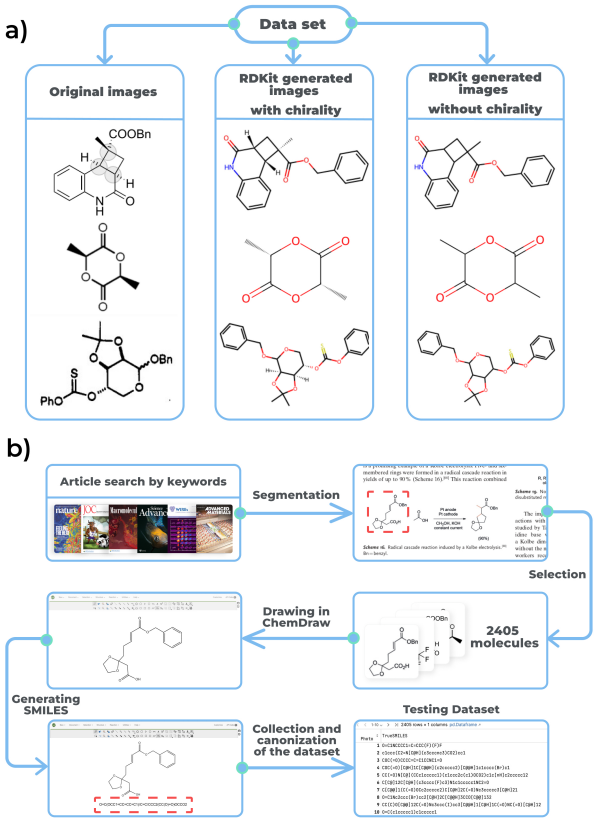

Fig. 1: a) Datasets the programs were tested on; b) steps of building a dataset of pictures of molecules with their corresponding SMILES.

3. Results

Comparative testing (Fig. 2) demonstrated that rule-based algorithms are less sensitive to image quality compared to neural networks. Among rule-based tools, OSRA performed the best with relatively stable recognition rates across different image types, while Imago and Molvec showed higher error rates for low-resolution images.

For neural network-based methods, DECIMER 2.0 exhibited the highest accuracy on synthetic RDKit-generated images (88.98%), followed by MolScribe (65.53%). On real molecule images extracted from scientific articles, MolScribe demonstrated the best performance (58.13%), though a significant drop

in accuracy was observed compared to synthetic data. This drop highlights the limitations of current datasets in training models to handle real-world variations in molecular depictions.

The study also revealed a strong correlation between recognition accuracy and molecule complexity. As molecule size increased (measured by atom count), accuracy declined across all tested tools (Appendix A). Large and highly branched molecules were particularly challenging for both rule-based and neural network methods. Additionally, the presence of chiral centers reduced recognition accuracy, with DECIMER 2.0 experiencing a decline of 7.4% when chirality was considered.

Recognition challenges were most pronounced for specific molecular features (Fig. 3), such as organometallic complexes, polycyclic systems, and structures containing charged fragments. Programs also struggled with non-standard representations, including superatoms and condensed structural formulas.

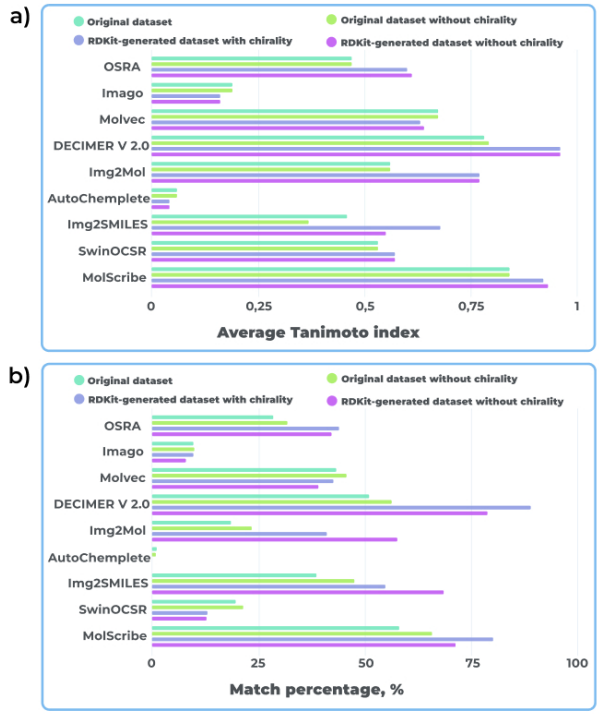

Fig. 2: (a) Tanimoto index of molecule image recognition results and (b) accuracy (in percent) of molecule image recognition results.

4. Discussion

Neural network-based recognition tools performed significantly worse on real molecular images than synthetic ones, confirming the limitations of current training datasets. The reliance on RDKit-generated structures fails to account for real-world variations, reducing performance, particularly for complex structures such as polycyclic systems, organometallic compounds, and chiral molecules. Larger molecules (over 60 atoms) also showed lower recognition accuracy, with errors increasing as structural complexity grew.

To improve recognition accuracy, future datasets should incorporate more real molecular images from scientific literature and apply augmentation techniques to simulate real-world distortions. Additionally, refining neural network architectures and integrating rule-based preprocessing could enhance generalization across different molecular representations.

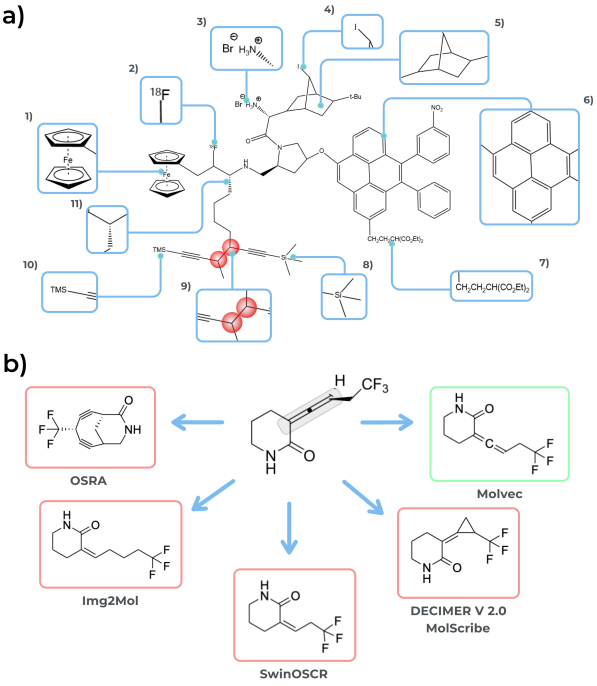

Fig. 3: (a) Manually generated molecule reflecting the main recognition problems: 1) organometallic complexes; 2) isotopes; 3) charged particles; 4) single-symbol chemical elements; 5) bicyclic structures; 5) polyaromatic systems; 6) condensed structural format; 7) two-symbol elements; 9) context highlights; 10) letter abbreviation of chemical compounds and fragments; 11) chirality; (b) ways of the same molecule recognition by various instruments.

5. Conclusion

This study highlighted the importance of creating diverse and high-quality datasets with molecule images for training neural network-based recognition systems. Current algorithms have limitations in dealing with complex and non-standard chemical structures. Nevertheless, ongoing developments and enhancements in optical recognition technologies will significantly simplify and accelerate the digitization of chemical data, enhancing research effectiveness and facilitating broader access to chemical information.

Acknowledgments

The authors thank the students of school 619 in St. Petersburg, Russia, and the students who attended the Infochemistry program at the Sirius Educational Center in April 2023 for their help in collecting the original dataset of molecules. RSF grant 23-16-00224 is acknowledged for financial support. Priority 2030 Federal Academic Leadership Program is acknowledged for infrastructural support.

References

[1] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery, 2018.

[2] Zachary J Baum, Xiang Yu, Philippe Y Ayala, Yanan Zhao, Steven P Watkins, and Qiongqiong Zhou. Artificial intelligence in chemistry: current trends and future directions, 2021.

[3] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T. Chayes, and Omar M. Yaghi. Chatgpt chemistry assistant for text mining and the prediction of mof synthesis, 2023.

[4] Greg Landrum. Rdkit documentation, 2013.

[5] Igor Filippov and Marc Nicklaus. Optical structure recognition software to recover chemical information: Osra, an open source solution, 04 2009.

[6] Edbeard. Python wrapper for osra. supports r-group logic and integration with chemschematicresolver. https://github.com/edbeard/pyosra, 2020.

[7] Imagoocr. https://lifescience.opensource.epam.com/imago/index.html, 2023.

[8] Imago. Imago - open-source toolkit for 2d chemical structure image recognition. https://lifescience.opensource.epam.com/imago/index.html#commercial-availability, 2025. Accessed: 15-Mar-2025.

[9] Tylerperyea, Dkatzel-ncats, Caodac, and Dan2097. A feeble attempt at molecular recognition (in the literal sense). https://github.com/ncats/molvec, 2022.

[10] Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. Chemgrapher: optical graph recognition of chemical compounds by deep learning, 2020.

[11] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer 1.0: deep learning for chemical image recognition using transformers, 2021.

[12] Kohulan Rajan, Henning Otto Brinkhaus, M Isabel Agea, Achim Zielesny, and Christoph Steinbeck. Decimer. ai-an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14:5045, 2023.

[13] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. DECIMER-V2, May 2022.

[14] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):1–9, 2020.

[15] Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2mol–accurate smiles recognition from molecular graphical depictions, 2021.

[16] Ivan Khokhlov, Lev Krasnov, Maxim V Fedorov, and Sergey Sosnin. Image2smiles: Transformer-based molecular optical recognition engine, 2022.

[17] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. Molscribe: Robust molecular structure recognition with image-to-graph generation, 2023.

[18] Zhanpeng Xu, Jianhua Li, Zhaopeng Yang, Shiliang Li, and Honglin Li. Swinocsr: end-to-end optical chemical structure recognition using a swin transformer, 2022.

[19] Youjun Xu, Jinchuan Xiao, Chia-Han Chou, Jianhang Zhang, Jintao Zhu, Qiwan Hu, Hemin Li, Ningsheng Han, Bingyu Liu, Shuaipeng Zhang, et al. Molminer: You only look once for chemical structure recognition, 2022.

[20] Merlin Knaeble, Gabriel Sailer, Zihan Chen, Thorsten Schwarz, Kailun Yang, Mario Nadj, Rainer Stiefelhagen, and Alexander Maedche. Autochemplete - making chemical structural formulas accessible, 04 2023.

Appendix A. orrelation between the number of atoms and the quality of recognition

There is a correlation between the number of atoms and the quality of recognition.

The larger the molecule, the more difficult it is for its image to transform into SMILES. However, what is this correlation in numbers? To confirm this assumption, we have posted diagrams that can be used to give a certain estimate of recognition accuracy (Fig. A1).

Appendix B. Availability of a unique dataset for testing

The collected dataset, as well as images of structures generated with the RDKit library are available in the GitHub repository at `https://github.com/RodionGolovinsky/dataset_testing_recognition_tools`. Also in this repository there is a complete table with the results of testing optical recognition tools on the described datasets.

Fig. A1: (a) Comparison of molecule image recognition results depending on the amount of atoms on RDKit-generated images not considering chirality; (b) comparison of molecule image recognition results depending on the amount of atoms on RDKit-generated images considering chirality; (c) comparison of molecule image recognition results depending on the amount of atoms on images from original dataset not considering chirality; (d) comparison of molecule image recognition results depending on the amount of atoms on images from original dataset considering chirality.