Optical Graph Recognition of Chemical Compounds by Deep Learning

Martijn Oldenhof [*1] **Adam Arany** [*1] **Yves Moreau** [1] **Jaak Simm** [1]

1. Motivation

In drug discovery, knowledge of the graph structure of chemical compounds is essential. Many thousands of scientific articles in chemistry and pharmaceutical sciences have investigated chemical compounds, but in cases the details of the structure of these chemical compounds is published only as images. We develop a deep neural network model for optical compound recognition.

2. Related Work

A few tools for recognizing graph structures from chemical compound images are available, such as OSRA [1] and others like ChemReader [4], Kekule [3], CLiDE Pro [6], and the work of M. Sadawi et al. [2]. However, we observe that, using these tools, bond multiplicity and stereochemical information are sometimes lost. Those tools are mainly expert systems using different techniques, such as image processing, optical character recognition or hand-coded rules. Modifying or further improving these tools requires a lot of effort. A tool based on machine learning, which learns directly from training data, would be most valuable.

In the work of Staker et al. [5], another deep learning model was also proposed. However, there the output is only a text-sequence representing the graph. In our approach, we focus on directly predicting the graph structure (*i.e.*, identifying all the nodes and the edges and their labels). The nodes and edges in the resulting graph are annotated with the positions in the original image of the chemical structure.

3. Proposed Model

The deep learning solution presented here consists of an image segmentation model, followed by three classification models that predict atom locations, bonds and charges. Furthermore, this model not only predicts the graph structure of the molecule but also produces all information necessary to relate each component of the resulting graph to the source image. This solution is scalable and could rapidly process thousands of images.

[*]Equal contribution [1]ESAT, KU Leuven, Leuven, Belgium. Correspondence to: Martijn Oldenhof <martijn.oldenhof@esat.kuleuven.be>.

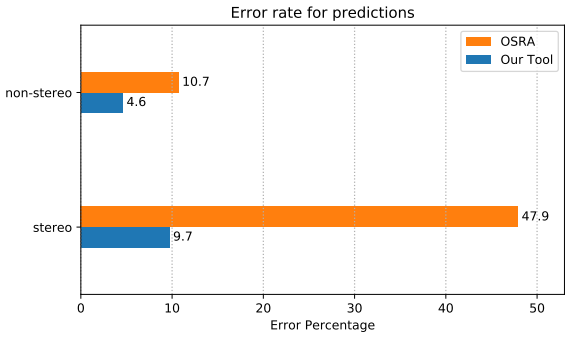

Figure 1. Comparison of error rate between our tool and OSRA [1]: our tool has an improvement of 4.94X for compounds with stereo chemical information and an improvement of 2.33X for compounds with no stereo chemical information.

4. Results

We ran two experiments to compare our tool with OSRA [1]. First we compare the performance of prediction of both tools on images with no stereo chemical information encoded. There we obtain an error rate of 10.7% for OSRA and an error rate of 4.6% for our tool. If we do the same with images with stereo chemical information encoded, OSRA performs much worse with an error rate of 47.9% and our tool only has an error rate of 9.7%. The results are summarized in Figure 1.

5. Conclusions

We presented a method to recognize the graph structure of molecules from 2D images of chemical structures using deep learning. This method learns a model directly from data. The performance is also clearly better than the well known tool OSRA [1] and also provides us more detailed information about the resulting graph. Finally, we would stress on the fact that in the context of drug discovery, such tools are important. In fact, in general we see that machine learning is gaining importance in this area and that it contributes to improving the quality of the drug discovery process.

6. Acknowledgments

MO, AA, YM and JS are funded by (1) Research Council KU Leuven: C14/18/092 SymBioSys3; CELSA-HIDUCTION, (2) Innovative Medicines Initiative: MELLODY, (3) Flemish Government (ELIXIR Belgium, IWT: PhD grants, FWO 06260) and (4) Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. Computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI. We also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] Igor V. Filippov and Marc C. Nicklaus. "Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution". In: *Journal of Chemical Information and Modeling* 49.3 (2009), pp. 740–743. URL: https://doi.org/10.1021/ci800067r.

[2] Noureddin M. Sadawi, Alan Sexton, and Volker Sorge. "Chemical Structure Recognition: A Rule Based Approach". In: *Proceedings of SPIE - The International Society for Optical Engineering* 8297 (Jan. 2012), pp. 32–. URL: https://doi.org/10.1117/12.912185.

[3] Joe R. McDaniel and Jason R. Balmuth. "Kekule: OCR-optical chemical (structure) recognition". In: *Journal of Chemical Information and Computer Sciences* 32.4 (1992), pp. 373–378. URL: https://doi.org/10.1021/ci00008a018.

[4] Jungkap Park et al. "Automated Extraction of Chemical Structure Information from Digital Raster Images". In: *Chemistry Central journal* 3 (Mar. 2009), p. 4. URL: https://doi.org/10.1186/1752-153X-3-4.

[5] Joshua Staker et al. "Molecular Structure Extraction from Documents Using Deep Learning". In: *Journal of Chemical Information and Modeling* 59.3 (2019), pp. 1017–1029. URL: https://doi.org/10.1021/acs.jcim.8b00669.

[6] Aniko T. Valko and A. Peter Johnson. "CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition". In: *Journal of Chemical Information and Modeling* 49.4 (2009), pp. 780–787. URL: https://doi.org/10.1021/ci800449t.