Accessible Chemical Structural Formulas through Interactive Document Labeling

Merlin Knaeble[1], Zihan Chen[2], Thorsten Schwarz[2], Gabriel Sailer[1], Kailun Yang[2], Rainer Stiefelhagen[2], and Alexander Maedche[1]

[1] Research Group Information Systems I,
[2] Center for Digital Accessibility and Assistive Technology,
Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Abstract. Despite a number of advances in the accessibility of STEM education, there is a lack of advanced tool support for authors and educators seeking to make corresponding documents accessible. We propose an interactive labeling method that combines an AI with user input to create accessible chemical structural formulas and incrementally improve the model. The model is a deep learning method based on a convolutional neural network and a transformer-based encoder-decoder. We implement this in a tool that enables graphical labeling of structural formulas and supports the user by performing a similarity search to suggest matches. Our approach aims to improve both the efficiency and effectiveness of labeling chemical structural formulas for accessibility purposes.

Keywords: accessibility · STEM · chemistry · structural formulas · interactive labeling

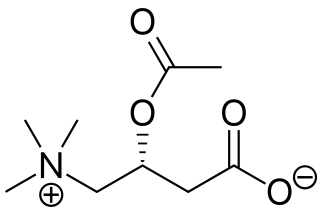
1 Introduction

Making learning materials accessible to visually impaired people is of utmost importance, especially in schools and higher education. The lack of accessible content was cited as one of the reasons why few blind students choose STEM subjects [18]. Learning chemistry, e.g., seems almost impossible, without its molecules visualized in structural formulas [7]. Thus, making such structural formulas accessible demands priority for STEM education. The issue is exacerbated when neither the authors are familiar with accessibility requirements, nor those who seek accessibility have the required STEM background.

Recently, in the research field of automatic image description generation promising deep-learning-based approaches for the recognition of mathematical formulas have been proposed [2]. However, the recognition of chemical structural formulas is more challenging. Often, it is based on low-level image processing approaches that require copious amounts of fine-tuning [9, 11, 16]. Further, none of the previous approaches integrate well into the established semi-automated approach in which STEM content is usually being made accessible. Currently either content authors themselves, or third parties like educators, try to make documents as a whole accessible. Thereby fully autonomous approaches while solving the issue of the process being time-consuming, do not tackle the challenge of

quality control. Neither fully automated approaches, due to lack of oversight, nor fully manual approaches, because of the error-proneness of the linearized label input, satisfy this. As such, there is a lack of research on how to facilitate interactive labeling of structural formulas efficiently and effectively.

In this paper we propose such an interactive approach and show its feasibility with a prototype. Firstly, we generated a dataset of images of structural formulas and their linear representations. We then trained a deep learning model using this dataset. To label new structural formulas, the model prediction is being fed into an interactive labeling interface that supports the user in either deciding that the prediction was accurate and thus accepting it, or correcting the linear representation of the input image. With this user input the deep learning model can be further improved, while also forming the basis to generate suitable alternative text representations for structural formulas that can be read and further used by blind chemists. We thus demonstrate the feasibility of interactive labeling approaches for structured image contents in document accessibility.



Colloquial name
Acetylcarnitine, Acetyl-L-carnitine, ALCAR, ALC

Sum formula
C9H17NO4

IUPAC name
(R)-3-Acetyloxy-4-trimethylammonio-butanoate

SMILES
[O-]C(=O)C[C@@H](OC(=O)C)C[N+](C)(C)C

InChI
InChI=1S/C9H17NO4/c1-7(11)14-8(5-9(12)13)6-10(2,3)4/h8H,5-6H2,1-4H3/t8-/m1/s1

Fig. 1: Acetylcarnitine as visualized in its structural formula (`https://en.wikipedia.org/wiki/Acetylcarnitine` and different text based (linear) representations of it.

2 Related Work

2.1 Accessibility of Structural Formulas

Research focusing on the accessibility of chemical structural formulas has undertaken several avenues. Via image segmentation and subsequent rule-based recognition, researchers have managed to make a first step towards automatic extraction of linear representations from an image [14]. Their approach initially translates the bitmap image into geometrical forms in a vector graphic. Such vector graphics are more easily accessible for the blind, especially with guidance and further labeling. However, slight errors during the translation of the image into geometric components make a rule-based detection of the underlying molecule all but unfeasible, as the resulting vector representations could be chemically impossible. Other approaches such as [11] build upon similar rules to extract geometric components. Practical applications, e.g. for teaching blind students chemistry in secondary education [18], underline how important not only vector representations, but also semantic enrichment of them are. However, such approaches require a certain domain expertise of the labeler.

Recent literature has been focusing largely on a single linearized representation called SMILES. This is a clear disadvantage because different standards

may be helpful to different readers of the document. There is a variety of textual representation formats for molecule structures including SMILES, InChI, colloquial name, IUPAC name, and sum formula [6], as illustrated in Figure 1. Among those standards, SMILES [19] and InChI [3] are compact, unique, and able to represent the entire molecule including its structure. SMILES [19] was designed to be read and written by humans, and it is therefore relatively straightforward to interpret, provided that the user knows a few basic principles of the format and thus usable for the people with visual impairments. InChI does not fulfill the human readability requirement (c.f. Figure 1). Finally, other representations such as aforementioned sum formulas, names or browseable vector graphics can be generated from this. This allows for a range of options regarding the output, such as generating a list of representations in the alt-text of a PDF, or using the multiple-rendition feature of EPUB3 to provide all alternatives simultaneously and selectable by user-choice [12]. In an educational setting, further requirements like exam fairness come to mind. Hereby giving blind students the IUPAC name would provide them with an unfair advantage as compared to their seeing peers, as such a name could already contain partial solutions to exam questions. Hereby, the person making the exam document accessible must make choices regarding the availability of certain representations.

2.2 Learning-based Detection and Labeling of Structural Formulas

The task of identifying molecular structure images in documents is challenging due to various reasons, especially because of complexity of molecule structures and the diversity of image formats and styles. Conventionally, low-level image processing techniques (scanning, vectorization, etc.) are used in conjunction with high-level rules to organize components into their respective structures [9,11,16]. Each stage must be fine-tuned independently as well as in relation to the other parts, leading to a time-consuming process of incorporating new elements requiring high amounts of human intervention. On the other hand, deep learning approaches using sequence-to-sequence models involve training a complex learning system represented by a single artificial neural network that embodies a complete target system. This system does not include explicit intermediate stages usually present in the traditional pipeline approach. Recently, deep learning has been applied to a variety of problems ranging from image-to-text and text-to-text tasks, many of which have demonstrated strong performances with sequence-to-sequence models [17]. Therefore, it is intuitive to apply sequence-to-sequence prediction models like transformers to this problem given the suitability of the image inputs and the sequential outputs. [15] presented deep learning solutions to predict SMILES encodings from bitmaps. They showed that deep learning can learn to predict images of molecules from literature at reasonably high accuracy. [10] used transformer models to predict SMILES encodings of chemical structure depictions with about 90% accuracy.

Outside of accessibility research, the domain of chemistry and specifically cheminformatics, have developed graphical labeling and search tools. A prominent example of such a tool is *kekule.js* [4]. Among other features it provides a

graphical user interface supporting the entry of structural formulas. As further aids for chemists there exist a series of databases of already discovered molecules, searchable by e.g., SMILES, InChI, IUPAC or name. Existing work has already incorporated graphical editors, foremost *kekule.js*, as input for such databases, e.g., [13]. Combined with a similarity search this allows for the retrieval of either structurally related molecules or a certain haziness in the users search input.

3 Interactive Labeling of Structural Formulas

Our method and tool are designed to be used either in batch processing of a large set of pre-segmented images of structural formulas or as a component of a document accessibility platform that is being invoked when there is an image containing such formulas in a document. Thereby, we consider our work as complimentary to existing research, e.g. [14], that focus largely on automated extraction and means of comprehensibility for blind users. The encoder-decoder concept was first designed to handle machine translation tasks in which both the input and output modalities are textual sequences. However, this concept is also known to be highly flexible such that the infrastructure of the encoder or decoder can be changed according to the required modality. In our specific context, it is intuitive to use a convolutional architecture to accommodate the imagery input, and a sequential auto-regressive model to handle the sequential generation nature of the decoder. Our overall architecture is a widely-used image captioning model [20], using a CNN-based encoder for feature extraction and an attention-augmented RNN decoder for generating outputs, namely SMILES and InChI in a multi-task learning setup.

Our work investigated different architectures, from simple to sophisticated, in order to show a progressive observation over the effectiveness of these architectures. We acquired a training dataset of about 4 million images from the PubChem database with matching SMILES labels of length less than 100. In our encoder-decoder architecture, for the encoder, we leverage 8 different EfficientNet variations B0, B3, and B7 with adding a transformer encoder layer [17] as well as ResNet152 for feature extraction. From a $3\times256\times256$ input image, we extract the last convolutional feature of a CNN-based encoder, which downsamples the image yet adds extra channels, as a $1024\times20\times20$ input for the decoder. From the point of view of sequence generation with attention, this can be viewed as an input with 400 memory states for querying in attention. On the other hand, the simplest solution for the decoder is an auto-regressive (without bi-directional) LSTM recurrent network that processes the text sequences in characters while using the attention mechanism [1] to align the generated characters with the input states in the encoder. In addition to such Convolution-Recurrent architecture, we take advantage of the power of Transformers to increase the modeling capacity and thus enhance learning performance in the next step. Transformer [17] originally replaces recurrence or convolution with self-attention and is theoretically able to represent both local and global connectivities, which CNN and RNN excel at, respectively. This model is also highly parallelizable and can utilize

the computational power of devices such as GPUs more efficiently than other networks. Using Transformers leads to adding self-attention layers on top of the convolutional features of the encoder and replacing the LSTM decoders with self-attentional decoders.

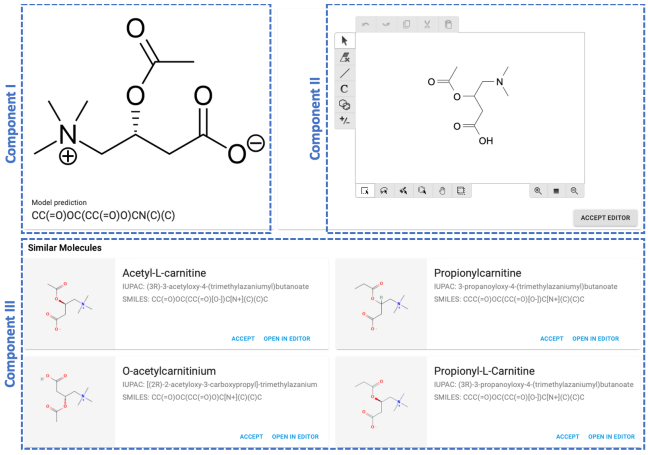


Fig. 2: User interface of our tool

Figure 2 shows the user interface (UI) of the interactive labeling tool. We have designed this interface with three main components to effectively support the creation of accessible chemical structural formulas. The focus is to automate the process as much as the state-of-the-art allows, while simultaneously using the input to collect more training data in order to further improve the model. We hereby follow established approaches concerning the development of interactive labeling systems [8]. Situated in the top left corner of the UI, **component I** depicts the original input image for reference, as well as the SMILES string that our model predicted. In **component II**, on the top right side, we integrated the established *kekule.js* molecule editor. Upon invocation of the interface, the editor is pre-loaded with an automatically generated graphical representation of the SMILES string the model predicted. The user can now compare this representation with the original input image to identify discrepancies - or to accept the current rendering in the editor. Furthermore, errors can be corrected via the graphical editor. This can be either small changes to the existing structure or a completely new one may be entered. Finally, **component III** shows similar molecules and is situated along the bottom. Specifically, it shows the structural formulas, names, IUPAC and SMILES representations of four molecules retrieved from the PubChem[3] database via a similarity search [6]. With this feature we provide advanced support for the user beyond the graphical editor. It is reasonable to assume that the molecules shown in literature are real existing molecules which are recorded in the PubChem database. This database is not only searchable via SMILES strings, but also allows for a similarity matching of structurally

[3] `https://pubchemdocs.ncbi.nlm.nih.gov`

kindred molecules. This search is initialized with the model prediction, but updates live as the user makes changes in the graphical editor. The user has the option to open any suggestion in the editor to make changes, or to accept it as is. Our labeling interface potentially allows for incorporating users without any chemical domain knowledge in the process of making PDFs accessible. Specifically, no knowledge about the output format of SMILES, is required, as the input image can be matched by visual comparison only. For a well performing model and similarity search engine, even the amount of manual graphical input is limited to edge cases, as predictions could be correct out of the box or the matching molecule may be contained in the suggestions in component III.

4 Evaluation

The two best performing model configurations were EfficientNet B2 and B7, each with 2 transformer encoder layers, respectively. Validation set accuracies for exact identifications are similar among them at 98.22% for B2+2 and 98.87% for B7+2. However, on yet unseen test data the results differ between them. To evaluate this, we use the mean Tanimoto similarity (T), as well as mean Levenshtein distances (L). The former refers to an established measure in chemistry, comparing molecule similarity along functional groups and the molecules effects in chemical reactions. Also known as edit distance, the latter is a general-purpose string distance measure, better representing the amount of changes required in the interactive labeling procedure. Hereby we report values of $T_{B2+2} = 70.50\%$ and $L_{B2+2} = 81.69\%$, as well as $T_{B7+2} = 80.86\%$ and $L_{B7+2} = 82.66\%$.

Comparing our results to other state-of-the-art solutions also employing deep-learning methods, we find that our performance is competitive. We achieve slightly better results than the work of [5] which reported a validation accuracy (on a molecule level) of 92.80% with our model version B7+2 at 98.87%. Furthermore we do so with significantly less training data of 2 versus 10 million images. Using even more data, at 35 million images, [10] obtain Tanimoto similarities of 96.47% far surpassing our values. They do however, limit the options for input severely, by allowing only the 12 most common elements, limiting the number of bonds and the weight, disallowing stereochemistry, counterions, charged groups and isotopes, and limiting the SMILES string to a length of 40. As compared to that, we pose just a single restriction of SMILES lengths less than 100. Following suggestions from other researchers, we could improve upon our work by adding image augmentations like blur and noise [10] and extending the size of our dataset, especially including low-quality real-world data [5].

A currently still missing subsequent user evaluation of our tool may focus on different types of users. Firstly, to obtain a baseline comparison to the current approach to making chemical formulas accessible, an evaluation should involve educators to blind or vision impaired STEM students. With this target audience performance and user-related benefits against aforementioned baseline can be investigated. Further, it allows for studying implications of different representational formats on the users of the output of our system (i.e. blind students

and chemists). Students may have different needs from seasoned chemists, as they may not need an efficient input on what information is contained in the image, e.g. as an IUPAC name, but may need to first understand how such a molecule is constructed, e.g. as a SMILES string or a vector graphic. Complementary, an evaluation may also involve users not previously involved in working on document accessibility. Hereby varying levels of chemical expertise could be represented in a stratified sample. On the top end, chemists with experience in using SMILES strings could typify power users that may not even profit from graphical entry as they proficiency with writing such strings may be too high. The other end of the spectrum is represented by chemical novices, knowing little more than the fact that the graphically labeled molecule should look the same as in the input image. Such an evaluation could provide input into the required domain expertise required to effectively contribute to making STEM documents accessible. If our tool is indeed capable in lowering the entry barrier for supporting the costly and labor-intensive process of labeling [8], it may contribute significantly to the goal of making knowledge accessible for all.

5 Conclusions

In this paper we have introduced a new interactive labeling method and tool to support making chemical structural formulas accessible. Thereby, we leverage the respective strengths of humans and computers. Creating accurate textual representations such as SMILES is achievable with state-of-the-art deep learning technology. However, precision cannot be guaranteed. Our interactive labeling approach involves a human, who is strongly supported by simultaneously accelerating input and checking its credibility. Thereby a first suggestion is being made by our deep learning model, while the subsequent correction process is supported by similarity matching. Graphical, as opposed to text-based, label input enables novices to contribute. With our approach, we do not only ensure precision, but also support continuous model improvement. There is still a lot of potential for optimization in the automated recognition of structural formulas, but our proof of concept has demonstrated its feasibility. The deep learning model is considered to be able to further improve upon a multi-task learning strategy in the future. Data acquired from using the tool will support here.

Ultimately, our next steps lie within improving our classifier, evaluating our interactive labeling tool with users, and later integrating it into broader platforms to make documents with STEM content accessible in a scalable way.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Deng, Y., Kanervisto, A., Rush, A.M.: What you get is what you see: A visual markup decompiler. arXiv preprint arXiv:1609.04938 (2016)

3. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC international chemical identifier. Journal of Cheminformatics (2015)
4. Jiang, C., Jin, X., Dong, Y., Chen, M.: Kekule.js: An open source javascript chemoinformatics toolkit. Journal of Chemical Information and Modeling **56**(6), 1132–1138 (2016)
5. Khokhlov, I., Krasnov, L., Fedorov, M.V., Sosnin, S.: Image2smiles: Transformer-based molecular optical recognition engine. Chemistry-Methods **2**(1), e202100069 (2022)
6. Kim, S., Thiessen, P., Cheng, T., Yu, B., Bolton, E.: An update on PUG-REST: RESTful interface for programmatic access to PubChem. Nucleic Acids Research **46**(W1), W563–W570 (2018)
7. McGrath, M., Brown, J.: Visual learning for science and engineering. IEEE Computer Graphics and Applications **25**, 56–63 (2005)
8. Nadj, M., Knaeble, M., Li, M.X., Maedche, A.: Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning. KI - Künstliche Intelligenz **34**, 131–142 (2020)
9. Park, J., Rosania, G.R., Shedden, K.A., Nguyen, M., Lyu, N., Saitou, K.: Automated extraction of chemical structure information from digital raster images. Chemistry Central Journal **6** (2009)
10. Rajan, K., Zielesny, A., Steinbeck, C.: DECIMER 1.0: Deep learning for chemical image recognition using transformers. Journal of Cheminformatics **13**(1), 61 (2021)
11. Sadawi, N.M., Sexton, A.P., Sorge, V.: Chemical structure recognition: a rule-based approach. In: Document Recognition and Retrieval XIX. vol. 8297, pp. 101 – 109. SPIE (2012)
12. Schwarz, T., Rajgopal, S., Stiefelhagen, R.: Accessible EPUB: Making EPUB 3 Documents Universal Accessible. In: Miesenberger, K., Kouroupetroglou, G. (eds.) Computers Helping People with Special Needs. pp. 85–92. Springer, Cham (2018)
13. Shave, S., Auer, M.: SimilarityLab: Molecular similarity for SAR Exploration and target prediction on the web. Processes **9**(9) (2021)
14. Sorge, V.: Polyfilling Accessible Chemistry Diagrams, vol. 9758, pp. 43–50. Springer (2016), 15th International Conference on Computers Helping People with Special Needs (ICCHP 2016)
15. Staker, J., Marshall, K., Abel, R., McQuaw, C.M.: Molecular structure extraction from documents using deep learning. Journal of Chemical Information and Modeling **59**(3), 1017–1029 (2019)
16. Valko, A.T., Johnson, A.P.: CLiDE Pro: The latest generation of CLiDE, a tool for optical chemical structure recognition. Journal of Chemical Information and Modeling **49**(4), 780–787 (2009)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. vol. 30. Curran Associates, Inc. (2017)
18. in't Veld, D., Sorge, V.: The dutch best practice for teaching chemistry diagrams to the visually impaired. In: Miesenberger, K., Kouroupetroglou, G. (eds.) Computers Helping People with Special Needs. pp. 644–647. Springer, Cham (2018)
19. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences (1988)
20. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML). pp. 2048–2057. PMLR (2015)