doi.org/10.26434/chemrxiv.12464420.v1

DECIMER - Towards Deep Learning for Chemical Image Recognition

Kohulan Rajan, Achim Zielesny, Christoph Steinbeck

Submitted date: 11/06/2020 • Posted date: 15/06/2020

Citation information: Rajan, Kohulan; Zielesny, Achim; Steinbeck, Christoph (2020): DECIMER - Towards Deep Learning for Chemical Image Recognition. ChemRxiv. Preprint.
https://doi.org/10.26434/chemrxiv.12464420.v1

The automatic recognition of chemical structure diagrams from the literature is an indispensable component of workflows to re-discover information about chemicals and to make it available in open-access databases. Here we report preliminary findings in our development of DECIMER (Deep lEarning for Chemical ImagE Recognition), a deep learning method based on existing show-and-tell deep neural networks which makes very few assumptions about the structure of the underlying problem. The training state reported here does not yet rival the performance of existing traditional approaches, but we present evidence that our method will reach a comparable detection power with sufficient training time. Training success of DECIMER depends on the input data representation: DeepSMILES are clearly superior over SMILES and we have preliminary indication that the recently reported SELFIES outperform DeepSMILES. An extrapolation of our results towards larger training data sizes suggest that we might be able to achieve >90% accuracy with about 60 to 100 million training structures, so that training can be completed within several months on a single GPU. This work is completely based on open-source software and open data and is available to the general public for any purpose.

DECIMER - Towards Deep Learning for Chemical Image Recognition

Kohulan Rajan[1], Achim Zielesny[2], Christoph Steinbeck[1*]

[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, D-45665 Recklinghausen, Germany

Corresponding author email: christoph.steinbeck@uni-jena.de

Abstract

The automatic recognition of chemical structure diagrams from the literature is an indispensable component of workflows to re-discover information about chemicals and to make it available in open-access databases. Here we report preliminary findings in our development of DECIMER (Deep lEarning for Chemical ImagE Recognition), a deep learning method based on existing show-and-tell deep neural networks which makes very few assumptions about the structure of the underlying problem. The training state reported here does not yet rival the performance of existing traditional approaches, but we present evidence that our method will reach a comparable detection power with sufficient training time. Training success of DECIMER depends on the input data representation: DeepSMILES are clearly superior over SMILES and we have preliminary indication that the recently reported SELFIES outperform DeepSMILES. An extrapolation of our results towards larger training data sizes suggest that we might be able to achieve >90% accuracy with about 60 to 100 million training structures, so that training can be completed within several months on a single GPU. This work is completely based on open-source software and open data and is available to the general public for any purpose.

Introduction

The automatic recognition of chemical structure diagrams from the chemical literature (herein termed Optical Chemical Entity Recognition, OCER) is an indispensable component of

workflows to re-discover information about chemicals and to make it available in open-access databases. While the chemical structure is often at the heart of the findings reported in chemical articles, further information about the structure is present either in textual form or in other types of diagrams such as titration curves, spectra, etc. (Figure 1).

Previous software systems for OCER been described and were both incorporated into commercial and open source systems. These software systems include Kekulé [1-2], the Contreras system [3], the IBM system [4], CLIDE [5] as well as the open source approaches chemOCR [6-8] , ChemReader [9] and OSRA [10].

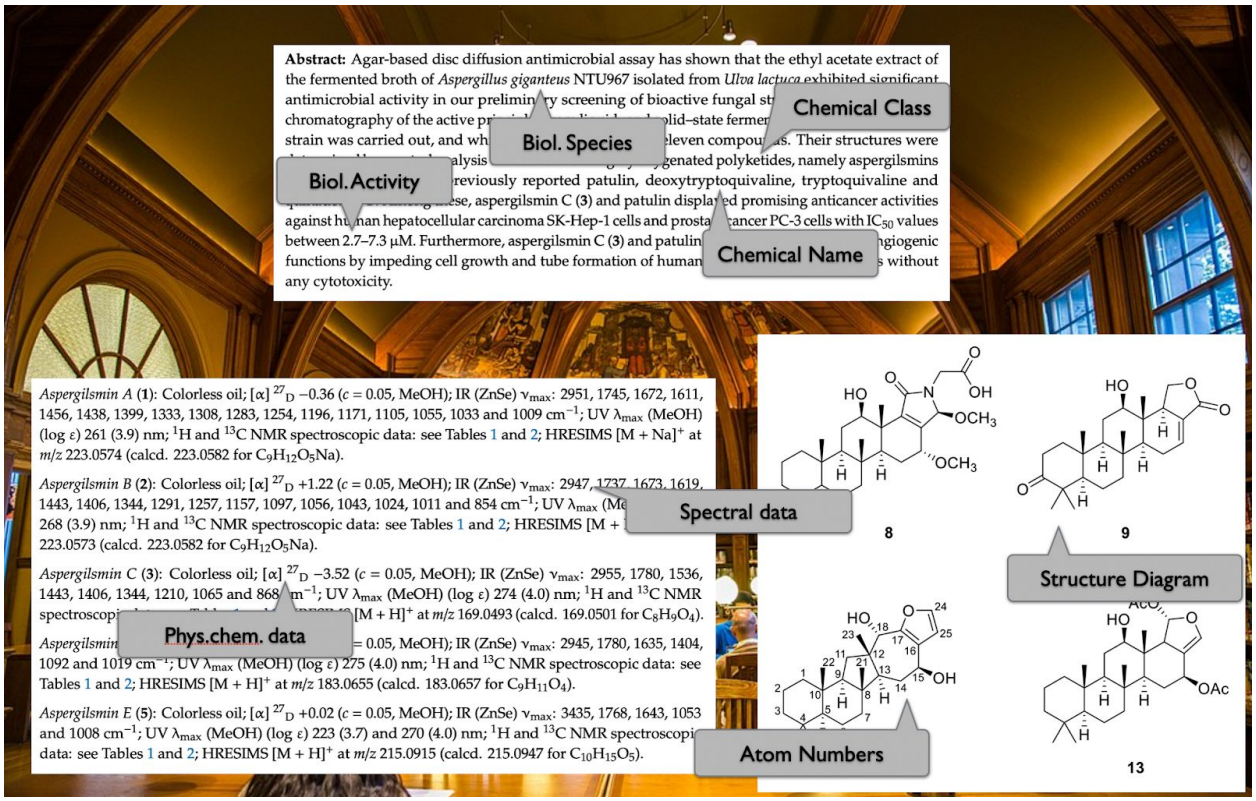


Figure 1: Information about a natural product is scattered across the various sections of an individual scientific article. Grouped around a structure and a chemical name, further information such as chemical classes, species and organism parts from which the compound was isolated, spectral and other data are listed. [1]

[1] Background image © Alina Chan, distributed under https://creativecommons.org/licenses/by-sa/4.0/deed.en, figures and text from Kwon 2020 [11]

A general approach to the problem is shared by all of these software packages, comprising steps a) scanning, b) vectorization, c) searching for dashed lines and dashed wedges, d) character recognition, e) graph compilation, f) post processing and d) display and editing.
Each of the steps in such systems need to be carefully hand-tuned both individually as well as for its interplay with the other steps. The incorporation of new image features to be detected is a laborious process.
We were recently inspired by the stunning success of AlphaGo Zero [12], a deep neural network (NN) based approach that enabled AlphaGo Zero to reach superhuman strength in the Game of Go by playing a potentially unlimited number of games against itself, starting with no more knowledge than the basic rules of the game. In this example, as well as in other prominent examples of successful deep learning, the key to success was the availability of a potentially unlimited or very large amount of training data.
The example of AlphaGo Zero made us realize that we are in a similar situation for the visual computing challenge described above. Instead of working with a necessarily small corpus of human-annotated examples from the printed literature, as has been common in the text mining and machine learning applications in chemistry in the past, we realised that we could generate training data from a practically unlimited source of structures generated by structure generators or by using the largest collections of open chemical data available to mankind.
After we started our work presented here, other attempts to use deep learning for OCER were reported. Work by the Schrödinger group [13] reports the successful extraction of machine-readable chemical structures from bitmaps but no software system available for the general public to replicate the reported results. A method called Chemgrapher [14] suggests to deal with the problem in a modular fashion by using a segmentation algorithm to segment the images containing chemical graphs to detect atoms locations, bonds and charges separately, and employ a graph building algorithm to re-generate the chemical graph.
Here we report preliminary findings of our development of DECIMER (Deep lEarning for Chemical ImagE Recognition), a deep learning method based on existing show-and-tell deep neural networks which, unlike for example Chemgrapher, makes very few assumptions about the structure of the underlying problem. The training state reported here does not yet rival the performance of existing traditional approaches, but we present evidence that, given sufficient training time, our method will reach a comparable detection power without the need of the sophisticated engineering steps of an OCER workflow.

Method

The principal idea reported here is to repurpose a show-and-tell deep NN designed for general photo annotation earlier and train it to report series of SMILES tokens when presented with bitmaps of chemical structures. The original NN reported sentences like “A giraffe standing in a forest with trees in the background” when presented with a corresponding photo.

Data

Instead of abstracting chemical diagrams from the chemical literature to generate training data, we decided to use structure diagram generators (SDG) like the one found in the Chemistry Development Kit (CDK) [15] to generate a potentially unlimited amount of training data. This type of training data can be accommodated to become more realistic and comparable to the varying picture quality in the chemical literature by using image manipulation such as blurring, adding noise, etc. As a source of input structures for the CDK SDG, we turned to PubChem [16], one of the largest databases of organic molecules. The following rules were used to curate the Pubchem data for our work presented here (in future versions of this deep NN, these rules might be relaxed):

Molecules must

- have a molecular weight of fewer than 1500 Daltons,
- not possess counter ions,
- only contain the elements C, H, O, N, P, S, F, Cl, Br, I, Se and B,
- not contain isotopes of Hydrogens (D, T),
- have 5 - 40 bonds,
- not contain any charged groups,
- only contain implicit hydrogens, except in functional groups,
- have less than 40 SMILES characters.

The generation of molecular bitmap images from chemical graphs was performed using the CDK SDG, which generates production quality 2D depictions to feed the deep learning algorithm.

The text data used here were SMILES [17] strings which were encoded into different formats, regular SMILES, DeepSMILES [18] and SELFIES [19], to test the dependency of the learning success on the data representation. These datasets were used in different training models in order to evaluate their performance for our use case. Here we report our results for DeepSMILES, where we currently have the richest set of data.

Model

For the model (Figure 2), we employed an autoencoder-based network with TensorFlow 2.0 [20] at the backend, based on the model designed by Xu et al., 2015 [21], in their work on Show, Attend and Tell, where they demonstrate a higher accuracy for an Image caption generation system with the attention mechanism. The encoder network is a convolutional NN (CNN) which consists of a single fully connected layer and a RELU activation function. The decoder network is a recurrent NN (RNN), consisting of a gated recurrent unit (GRU), 2 fully connected layers and a soft attention mechanism, which was introduced by Bahdanau, Cho, & Bengio in 2015 [22].

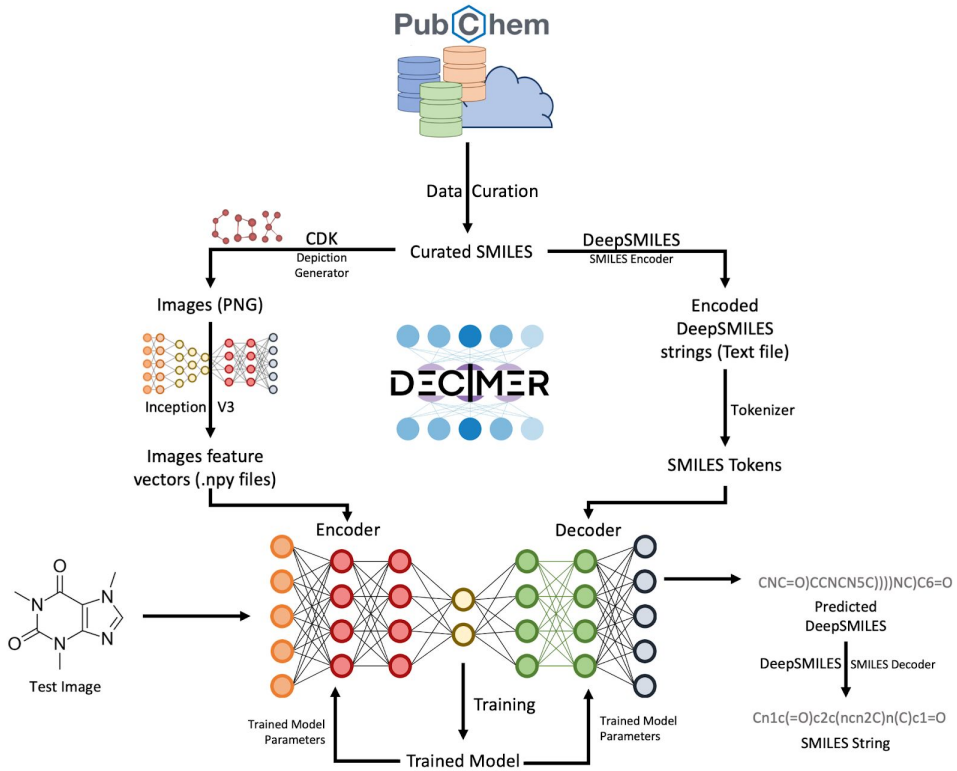


Figure 2: The complete architecture of the DECIMER model and workflow

We trained the model with DeepSMILES textual data and the corresponding bitmap of the chemical diagram. The text file is read by the model, the DeepSMILES is tokenized by the tokenizer and the unique tokens are stored. The images are converted into feature vectors by using the Inception V3 [23] model and saved as NumPy arrays.
The model accuracy is determined by the average of all the calculated Tanimoto similarity scores as well as the number of Tanimoto 1.0 hits.

Training

Initially, we trained multiple models with small training datasets to obtain the best hyper-parameters for our network. Exploration of the hyperparameter space led to 640 images per batch size, with embedding dimension size of 600 for the images which we depicted on a 299 x 299 canvas size to match the Inception V3 model. We used an Adam optimizer with a learning rate 0.0005 and Sparse Categorical Cross entropy to calculate the loss. We trained all the models for 25 epochs which typically led to convergence. Once the models converged, we started the evaluation of the test set.
The models were trained on an inhouse server equipped with an nVidia Tesla V100 Graphics Card, 384GB of RAM and two Intel(R) Xeon(R) Gold 6230 CPUs. Even though the training completely happens on the GPU, the initial dataset preparation was CPU-based.
Training time obviously scales with data size (Table 1). Model success was evaluated with an independent test data set. During the preparation of this manuscript, initial experiments with parallel training indicated that scaling was not satisfactory beyond 2 or 3 GPUs.

Results

Here we report our results for training data sizes between 54,000 and 15,000,000 structures, with the largest training data set taking 27 days to converge on the hardware reported above (Table 1).
Figure 4 shows the growth of the accuracy of predictions with increasing train data size.

Train Data Size	Test Data Size	Avrg. time/epoch (s)	Time for 25 epochs (s)
54000	6000	94.32	2358
90000	10000	159.88	3997
450000	50000	880.6	22015
900000	100000	2831.8	70795
1800000	200000	7239.28	180982
2700000	300000	11964.72	299118
4050000	450000	17495.12	437378
5850000	650000	25702	642550
7200000	800000	32926.8	823170
8969751	996639	41652.24	1041306
12600000	1400000	64909.28	1622732
15102000	1678000	91880.84	2297021

Table 1: Data sizes used in this work with computing times. The time for training the model with 15mio structures corresponds to approximately a month on a single Tesla V100 GPU.

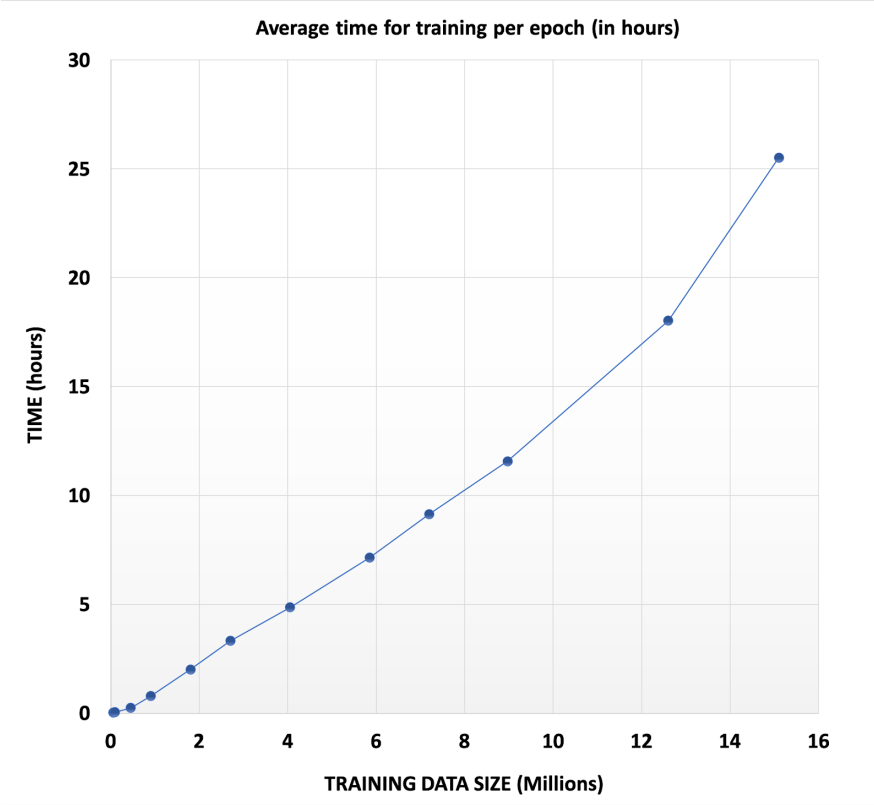


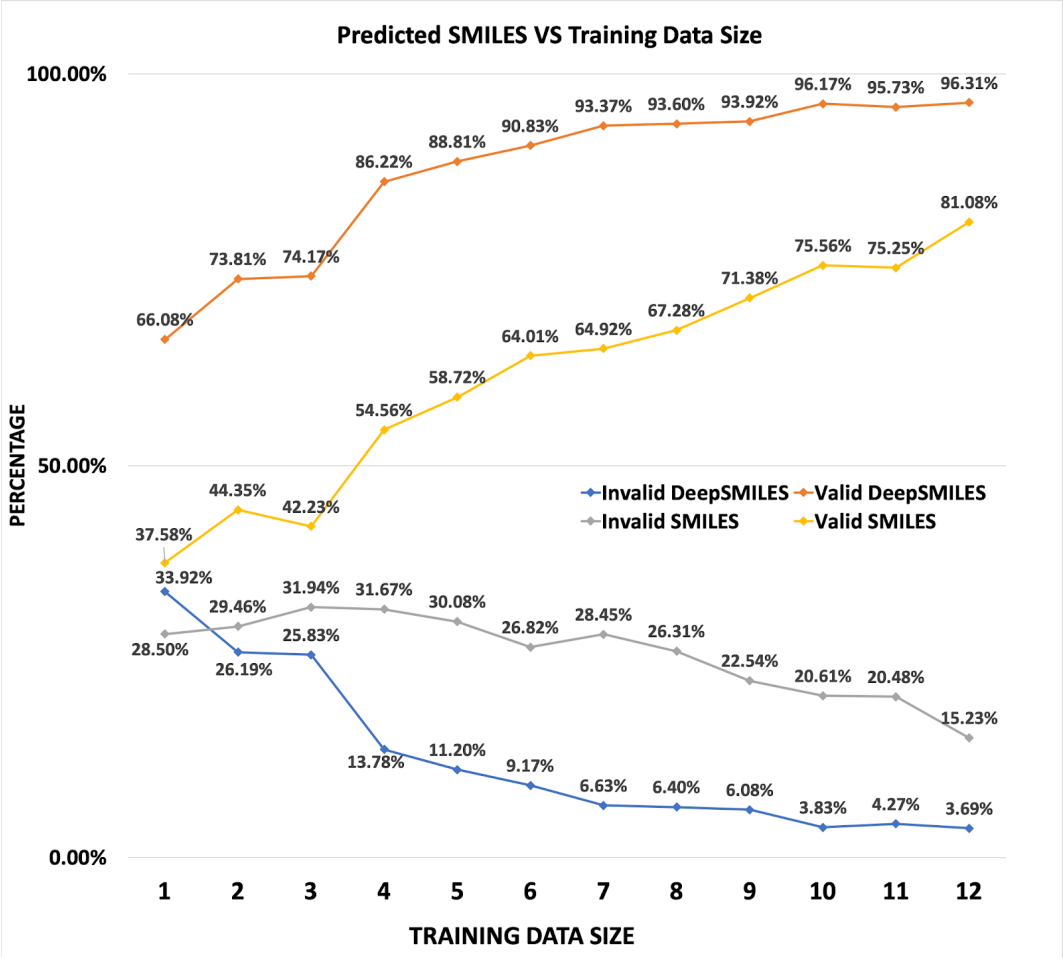
Figure 3 : Average time spent on training each epoch with increasing dataset size.

Training success was determined with a number of indicators (Figure 4), such as the percentage of Tanimoto 1.0 predictions, the average Tanimoto similarity over all predictions and the percentage of invalid SMILES produced by the model. Figure 4a demonstrates that the model's ability to produce valid SMILES and avoid invalid ones steeply increases with larger training datasets. The same can be observed for the two key parameters of this application, the average Tanimoto similarity and the Tanimoto 1.0 percentage, which indicate the fitness of the model to accurately generate a machine-readable structure from a bitmap of a chemical diagram. We further evaluated the models' success with additional descriptors such as LogP and Ring count between original and the predicted SMILES, which indicates that the model consistently produces better and better machine representations with growing training data size. The improvements do not seem to converge prematurely. More detailed results will be reported in a full paper later.

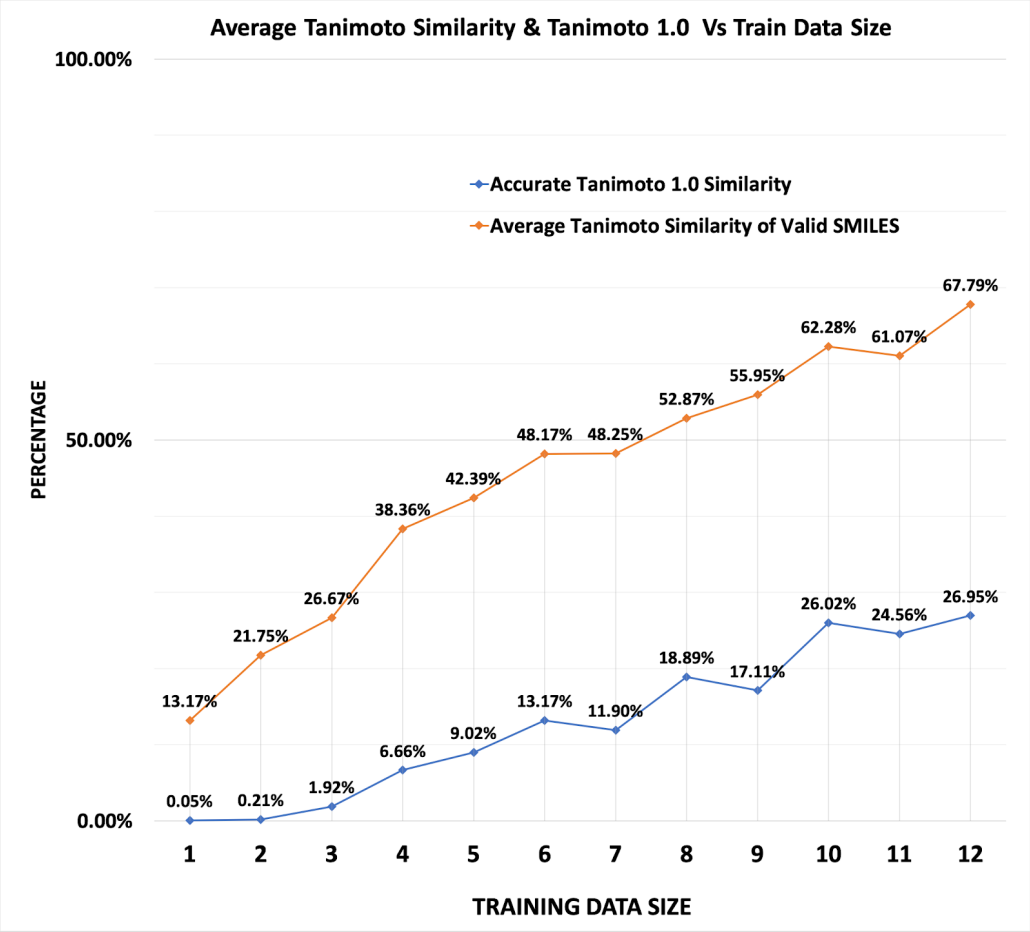
In order to assess the promise of these preliminary results, we performed an idealistic linear extrapolation of our data toward larger training data sizes, which indicate that close to perfect detection of chemical structures would require training data sizes between 50 and 60 million

structures. Such a training data volume will likely require a training time of 4 months with our setup with a single GPU..

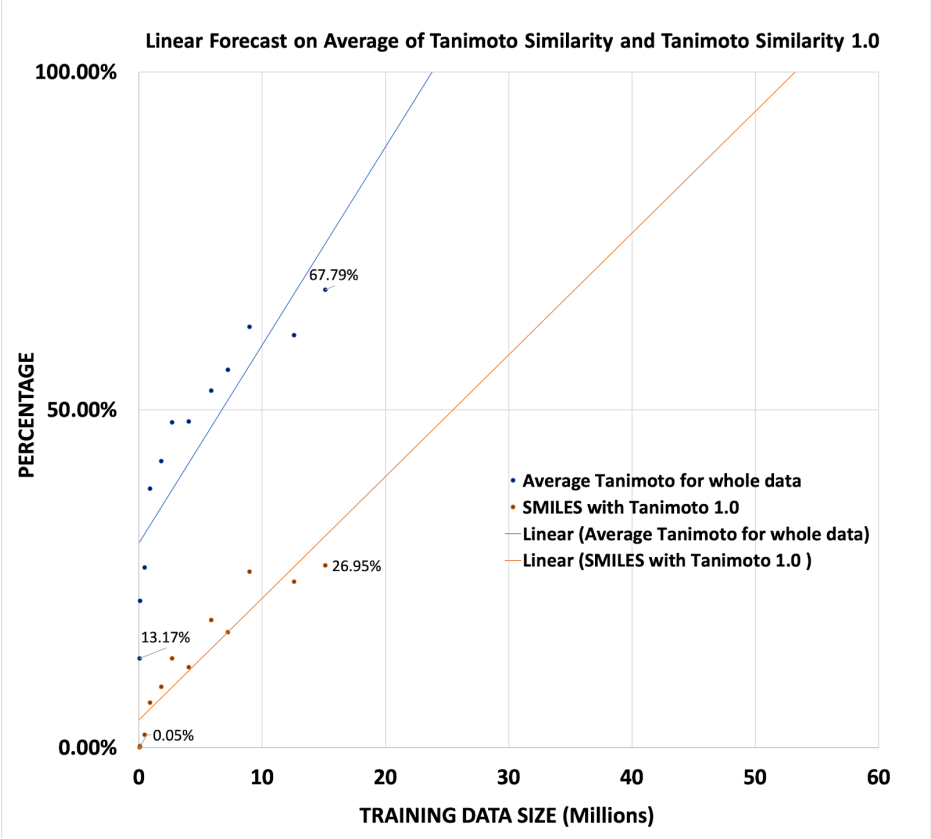
We are currently experimenting with the distributed learning solution currently available in the Tensorflow 2.0 API to significantly reduce this training time, also evaluating Google's Tensor Processing Units (TPU).



(a) Accuracy on predicted SMILES vs training data size, x axis bins correspond to the train data sizes mentioned in Table 1.

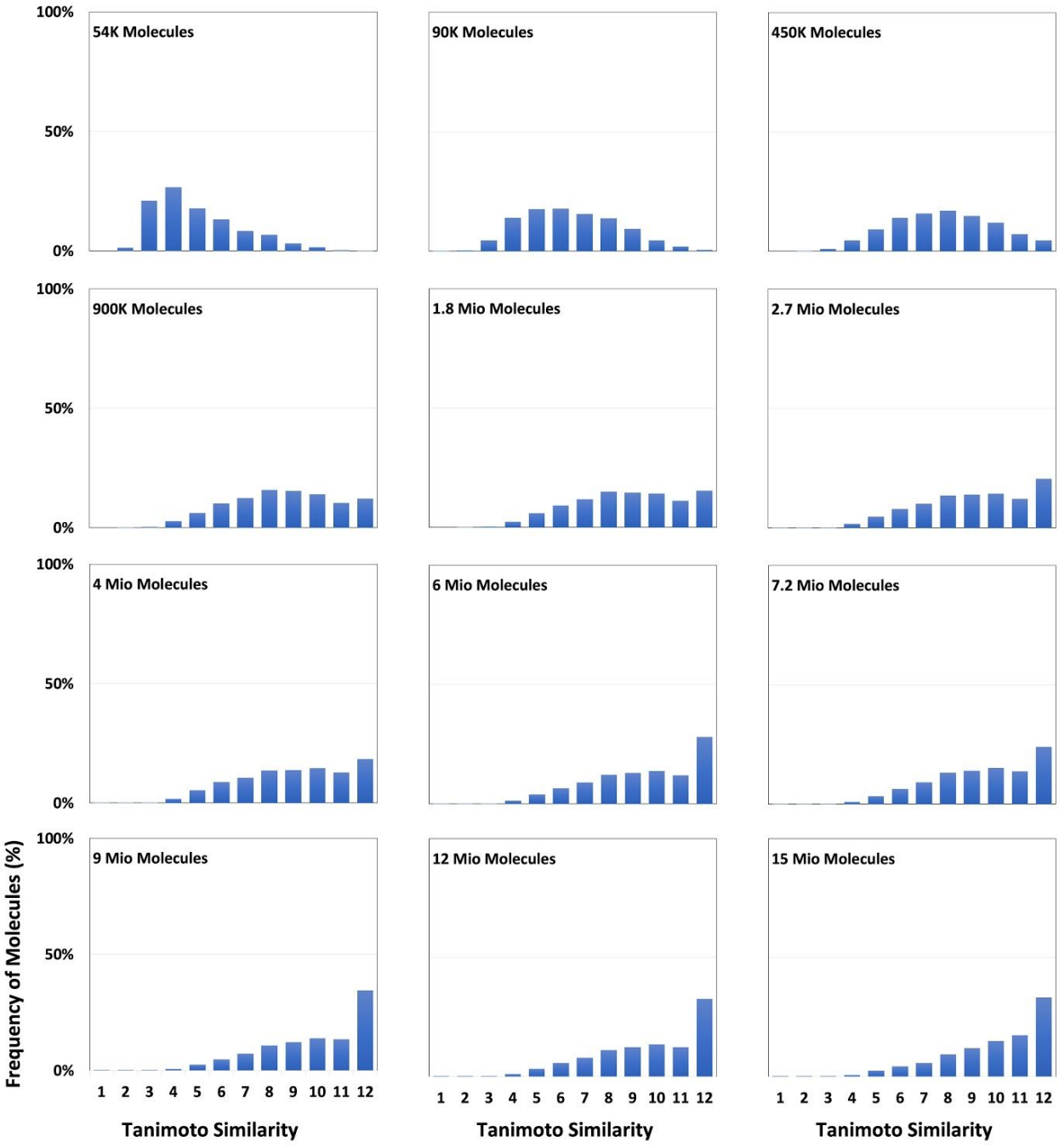


(b) *Average Tanimoto similarity and Tanimoto Similarity 1.0 on valid SMILES concerning the training data*



(c) *Linear fit on the predicted results forecasting the accuracy increases with more data.*

Figure 4: Development of training success indicators with increasing train data size



Bin values on x axis												
Bins	1	2	3	4	5	6	7	8	9	10	11	12
Values	0	0.0-0.1	0.1-0.2	0.2 - 0.3	0.3 - 0.4	0.4 -0.5	0.5 - 0.6	0.6 - 0.7	0.7 - 0.8	0.8 - 0.9	0.9 - 1.0	1

Figure 5: Distribution of Tanimoto-Similarity between the training structures and the structure actually recognised by DECIMER. Y-axis: Frequency of molecules in percentage, x-axis: Tanimoto similarity range in bins.

Conclusion

Here we have presented preliminary results indicating that a show-and-tell deep neural network setup has the potential to successfully extract a machine-readable structure representation when trained with tens of millions of examples. The training setup makes minimal assumptions about the problem. Training success depended on the input data representation. DeepSMILES were clearly superior over SMILES and we have the preliminary indication that the recently reported SELFIES outperform DeepSMILES. An extrapolation of our results towards larger training data sizes suggests that we might be able to achieve >90% accuracy with 60 to 100 million training structures. Such training can be completed in uncomfortable but feasible time spans of several months on a single GPU.

Our work is completely based on open-source software and open data and is available to the general public for any purpose.

We are currently moving towards larger training sets with the use of parallelization and more powerful hardware and hope to report the results in a full paper on this work in due time.

Availability of data and materials

The code for DECIMER is available at https://github.com/Kohulan/DECIMER , https://github.com/Kohulan/DECIMER-Image-to-SMILES

References

1. McDaniel JR, Balmuth JR. Kekule: OCR-optical chemical (structure) recognition. J Chem Inf Model [Internet]. 1992 Jul 1;32(4):373–8. Available from: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00008a018
2. Borman S. New computer program reads, interprets chemical structures. Chem Eng News [Internet]. 1992 Mar 23;70(12):17–9. Available from: http://pubs.acs.org/doi/abs/10.1021/cen-v070n012.p017
3. Contreras ML, Allendes C, Alvarez LT, Rozas R. Computational perception and recognition of digitized molecular structures. J Chem Inf Model [Internet]. 1990

Aug 1;30(3):302–7. Available from: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00067a014

4. Casey R, Boyer S, Healey P, Miller A, Oudot B, Zilles K. Optical recognition of chemical graphics. In: Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93) [Internet]. IEEE Comput. Soc. Press; 1993. p. 627–31. Available from: http://ieeexplore.ieee.org/document/395658/
5. Ibison P, Jacquot M, Kam F, Neville AG, Simpson RW, Tonnelier C, et al. Chemical literature data extraction: The CLiDE Project. J Chem Inf Model [Internet]. 1993 May 1;33(3):338–44. Available from: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00013a010
6. Zimmermann M, Bui Thi LT, Hofmann M. Combating Illiteracy in Chemistry: Towards Computer-Based Chemical Structure Reconstruction. ERCIM News [Internet]. 2005;60(60):40–1. Available from: http://www.ercim.eu/publication/Ercim_News/enw60/zimmermann.html%5Cnhttps://www.researchgate.net/publication/228766116_Combating_illiteracy_in_chemistry_towards_computer-based_chemical_structure_reconstruction
7. Algorri M-E, Zimmermann M, Friedrich CM, Akle S, Hofmann-Apitius M. Reconstruction of Chemical Molecules from Images. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society [Internet]. IEEE; 2007. p. 4609–12. Available from: http://ieeexplore.ieee.org/document/4353366/
8. Algorri M-E, Zimmermann M, Hofmann-Apitius M. Automatic Recognition of Chemical Images. In: Eighth Mexican International Conference on Current Trends in Computer Science (ENC 2007) [Internet]. IEEE; 2007. p. 41–6. Available from: http://ieeexplore.ieee.org/document/4351423/
9. Park J, Rosania GR, Shedden KA, Nguyen M, Lyu N, Saitou K. Automated extraction of chemical structure information from digital raster images. Chem Cent J [Internet]. 2009 Feb 5 [cited 2018 Dec 19];3(1):4. Available from: http://ccj.springeropen.com/articles/10.1186/1752-153X-3-4
10. Filippov I V., Nicklaus MC. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. J Chem Inf Model [Internet]. 2009 Mar 23;49(3):740–3. Available from: http://pubs.acs.org/doi/abs/10.1021/ci800067r
11. Kwon O-S, Kim D, Kim C-K, Sun J, Sim CJ, Oh D-C, et al. Cytotoxic Scalarane Sesterterpenes from the Sponge Hyrtios erectus. Mar Drugs. 2020;18(5):253.
12. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature [Internet]. 2017 Oct 18 [cited 2018 Dec 13];550(7676):354–9. Available from: http://www.nature.com/doifinder/10.1038/nature24270

13. Staker J, Marshall K, Abel R, McQuaw CM. Molecular Structure Extraction from Documents Using Deep Learning. J Chem Inf Model. 2019;59(3):1017–29.
14. Oldenhof M, Arany A, Moreau Y, Simm J. ChemGrapher: Optical Graph Recognition of Chemical Compounds by Deep Learning. 2020; Available from: http://arxiv.org/abs/2002.09914
15. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform [Internet]. 2017;9(1):33. Available from: https://doi.org/10.1186/s13321-017-0220-4
16. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: Improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102–9.
17. Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
18. O'Boyle N, Dalke A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. chemRxiv: 1026434 [Internet]. 2018;1–9. Available from: https://github.com/nextmovesoftware/deepsmiles
19. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. 2020 [cited 2020 Jun 2]; Available from: https://github.com/
20. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems [Internet]. 2016. Available from: http://arxiv.org/abs/1603.04467
21. Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. Show, attend and tell: Neural image caption generation with visual attention. 32nd Int Conf Mach Learn ICML 2015. 2015;3:2048–57.
22. Bahdanau D, Cho KH, Bengio Y. Neural machine translation by jointly learning to align and translate. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc. 2015;1–15.
23. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016. p. 2818–26.