Learning Chemistry: Exploring the Suitability of Machine Learning for the Task of Structure-based Chemical Ontology Classification

Janna Hastings (hastings@ovgu.de)
Otto von Guericke Universität Magdeburg https://orcid.org/0000-0002-3469-4923
Martin Glauer
Otto von Guericke Universität Magdeburg: Otto von Guericke Universitat Magdeburg
https://orcid.org/0000-0001-6772-1943
Adel Memariani
Otto-von-Guericke-University Magdeburg: Otto von Guericke Universitat Magdeburg
https://orcid.org/0000-0002-8368-7658
Fabian Neuhaus
Otto-von-Guericke-University Magdeburg: Otto von Guericke Universitat Magdeburg
https://orcid.org/0000-0002-1058-3102
Till Mossakowski
Otto-von-Guericke-University Magdeburg: Otto von Guericke Universitat Magdeburg
https://orcid.org/0000-0002-8938-5204

Research article

Keywords: chemical ontology, automated classification, machine learning, LSTM

Posted Date: November 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-107431/v1

Version of Record: A version of this preprint was published on March 16th, 2021. See the published version at https://doi.org/10.1186/s13321-021-00500-8.

Abstract

Chemical data is increasingly openly available in databases such as PubChem, which contains more than 110 million compound entries as of October 2020. With the availability of data at such scale, the burden has shifted to organisation, analysis and interpretation. Chemical ontologies provide structured classifications of chemical entities that can be used for navigation and filtering of the large chemical space. ChEBI is a prominent example of a chemical ontology, widely used in life science contexts. However, ChEBI is manually maintained and as such cannot easily scale to the full scope of public chemical data. There is a need for tools that are able to automatically classify chemical data into chemical ontologies, which can be framed as a hierarchical multi-class classification problem. In this paper we evaluate machine learning approaches for this task, comparing different learning frameworks including logistic regression, decision trees and long short-term memory articial neural networks, and different encoding approaches for the chemical structures, including cheminformatics fingerprints and character-based encoding from chemical line notation representations. We nd that classical learning approaches such as logistic regression perform well with sets of relatively specific, disjoint chemical classes, while the neural network is able to handle larger sets of overlapping classes but needs more examples per class to learn from, and is not able to make a class prediction for every molecule. Future work will explore hybrid and ensemble approaches, as well as alternative network architectures including neuro-symbolic approaches.

Full Text

This preprint is available for download as a PDF.

Figures

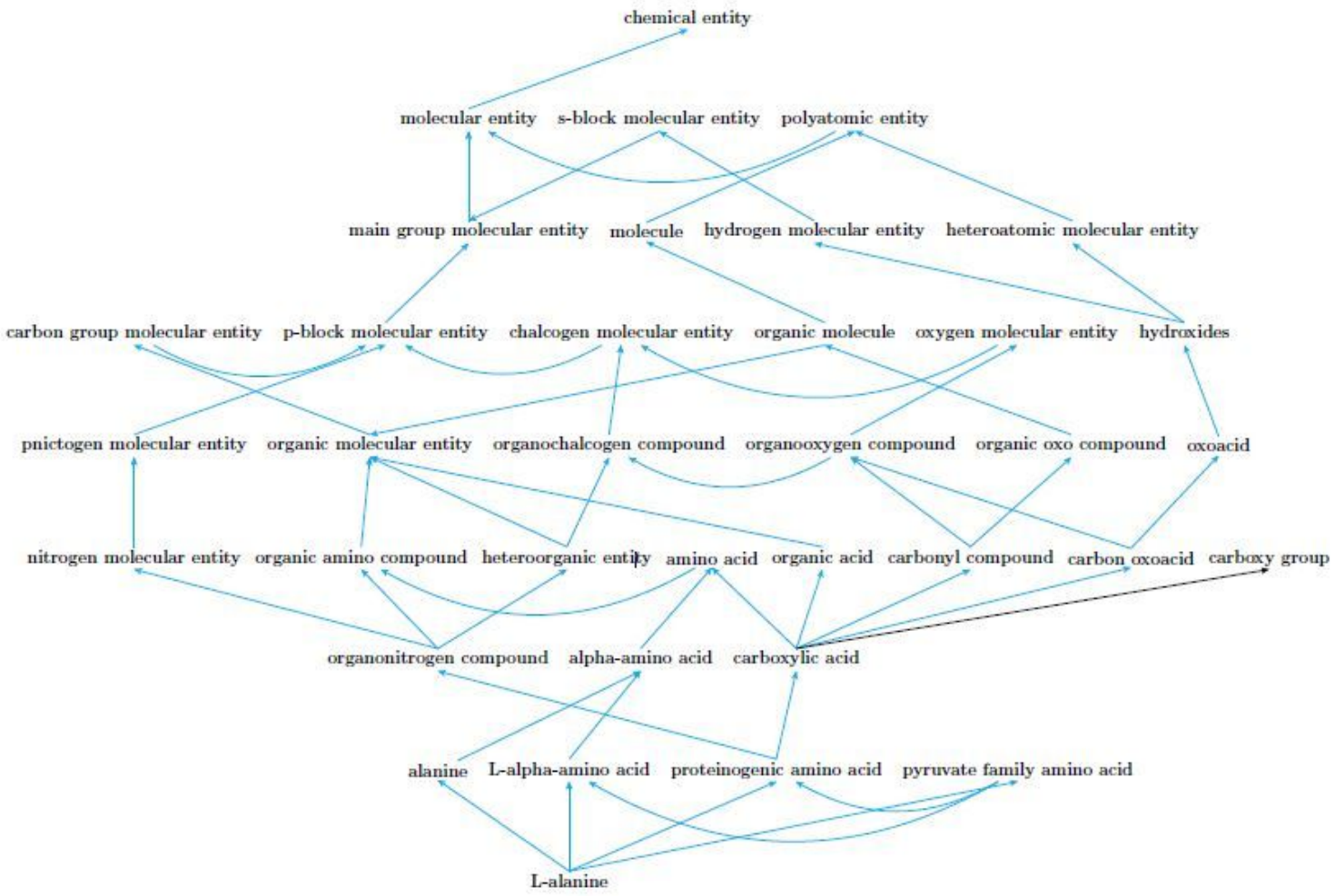

Figure 1

The gure illustrates the many branching and partially overlapping ancestor classes subsuming the molecule L-alanine (CHEBI:16977). Each of the illustrated mid-level classes similarly contains an overlapping range of molecular entity leaf members. For this reason, the ChEBI `chemical entity' ontology can be described as diamond-shaped. Blue arrows indicate subsumption relationships and the black arrow indicates a parthood relationship.

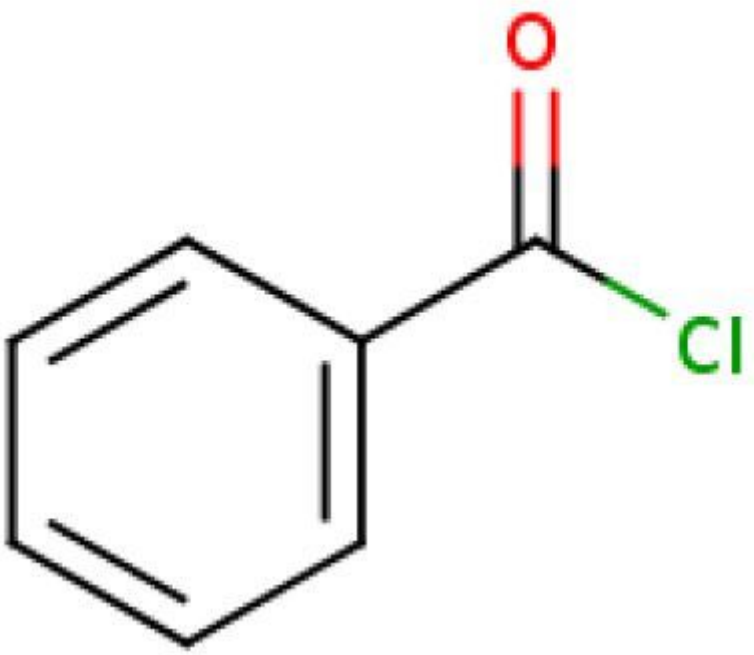

Figure 2

CHEBI:82275 benzoyl chloride, ClC(=O)C1=CC=CC=C1

Tokenisation without atom grouping:

C	l	C	(	=	O	)	C	1	=	C	C	=	C	C	=	C	1

Tokenisation with atom grouping:

Cl	C	(	=	O	)	C	1	=	C	C	=	C	C	=	C	1

Figure 3

Character-wise tokenisation with and without atom groupings

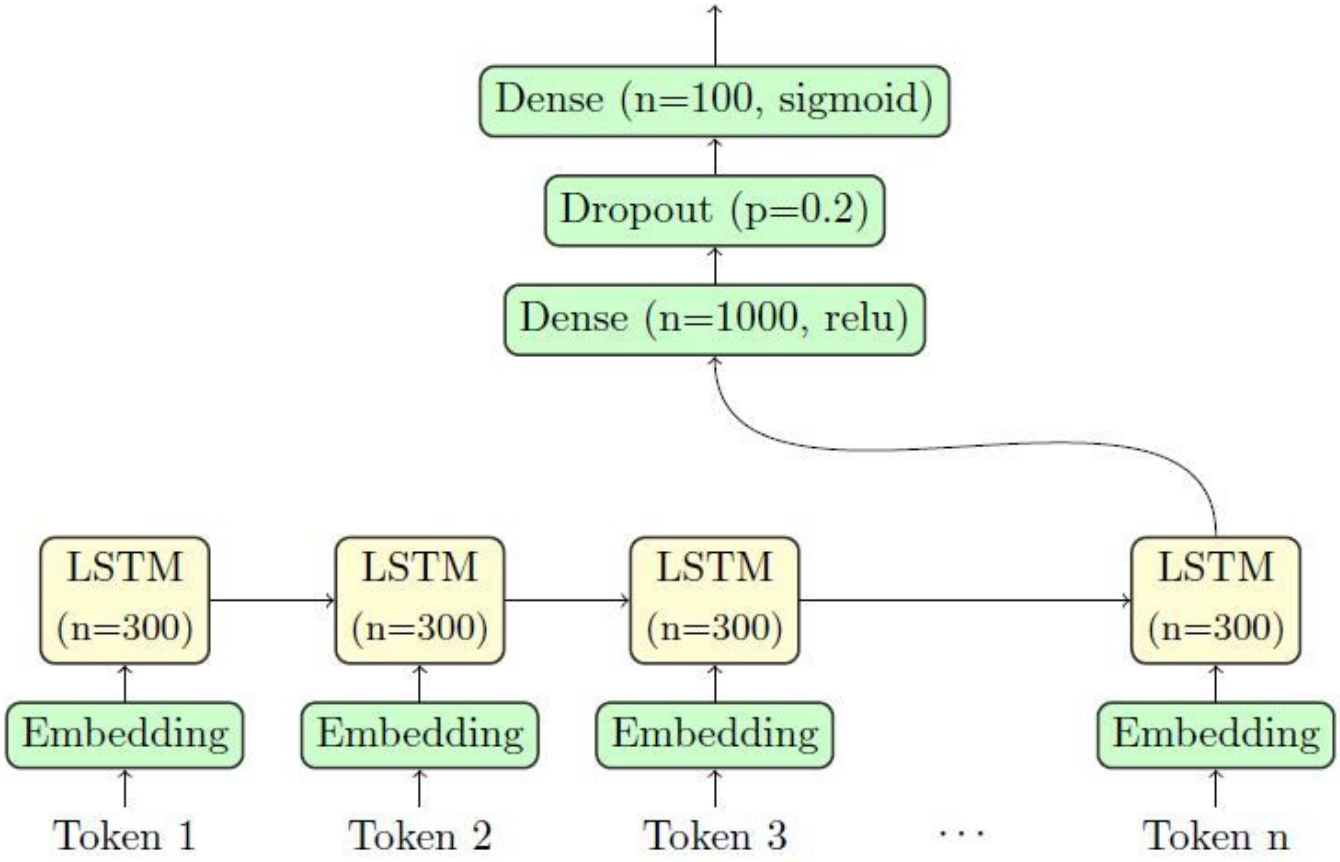

Figure 4

One-directional LSTM

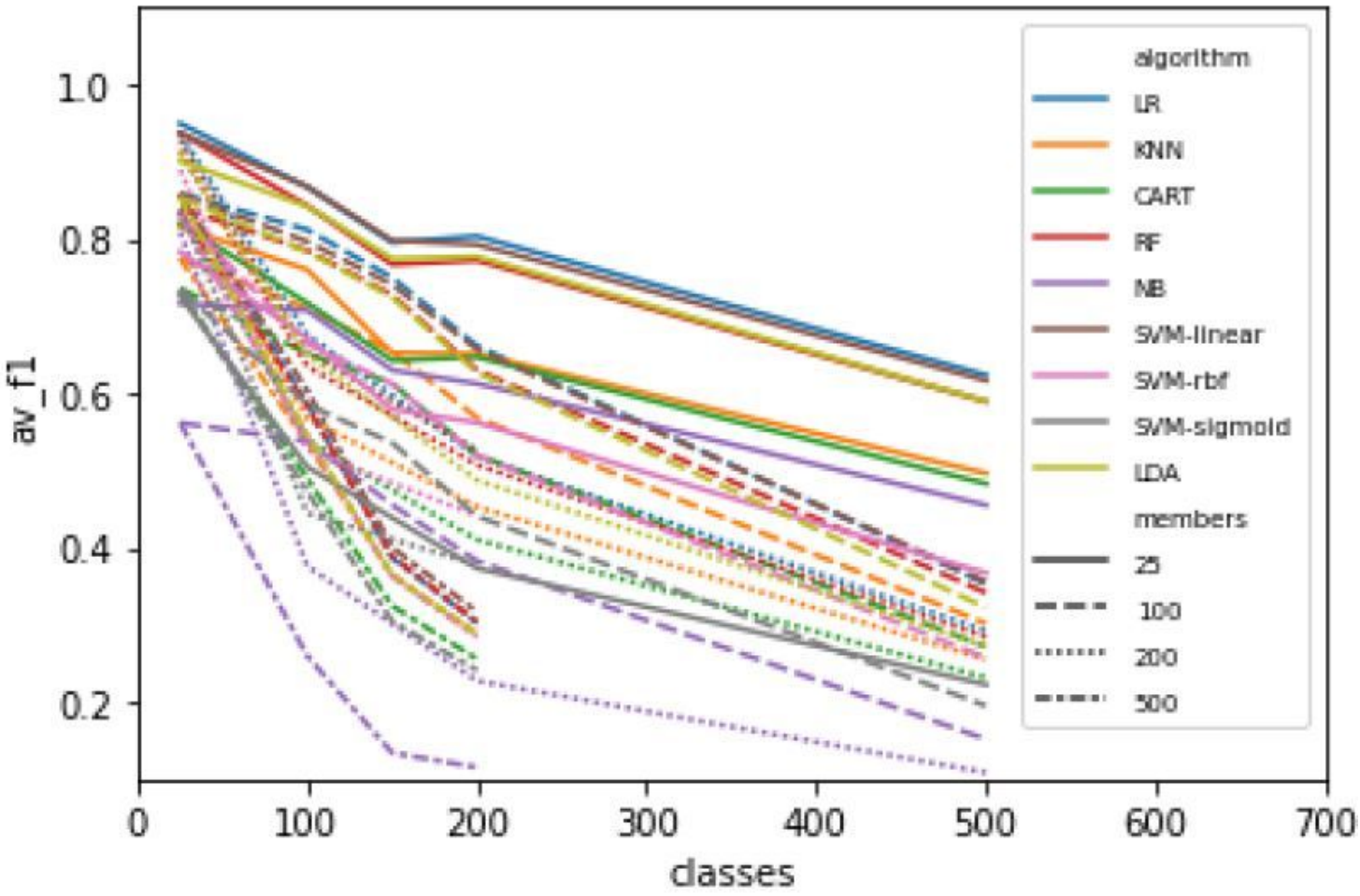

Figure 5

Mean F1 score across all classes for di erent problem sizes. LR=logistic regression;KNN=K-nearest neighbours; CART=decision tree; RF=random forest; NB=naive bayes; SVM=support vector machine; LDA=linear discriminant analysis

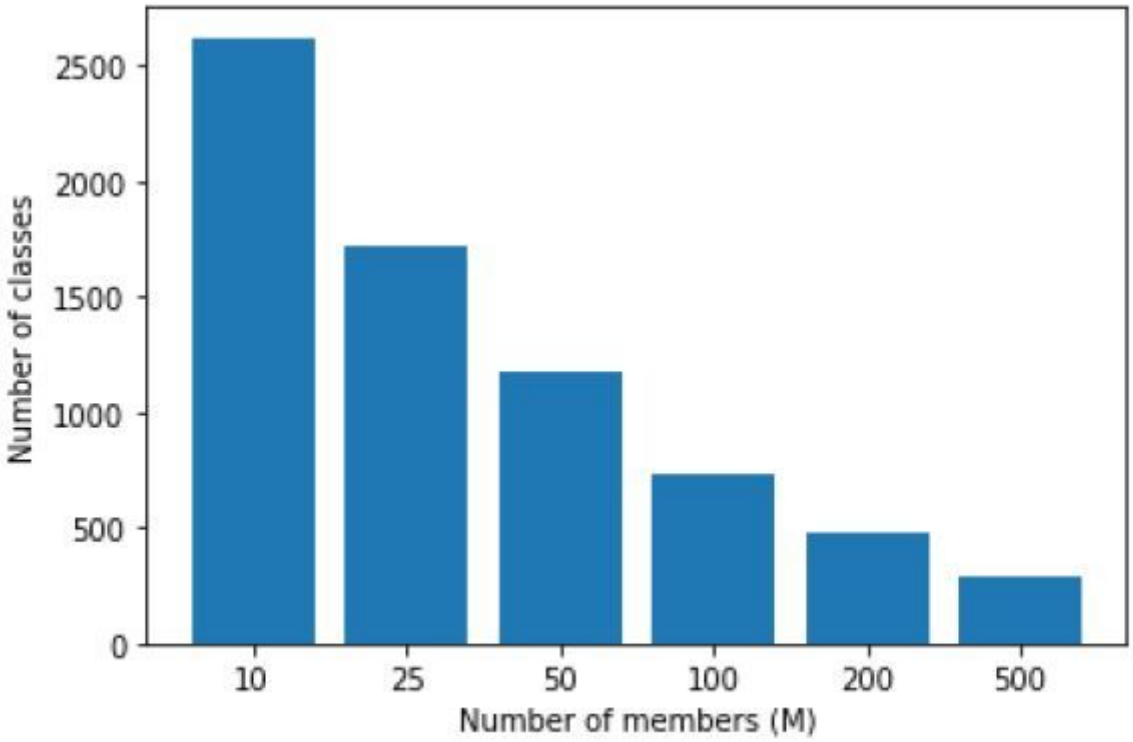

Figure 6

Number of classes with at least M members, for di erent sizes of M, in ChEBI

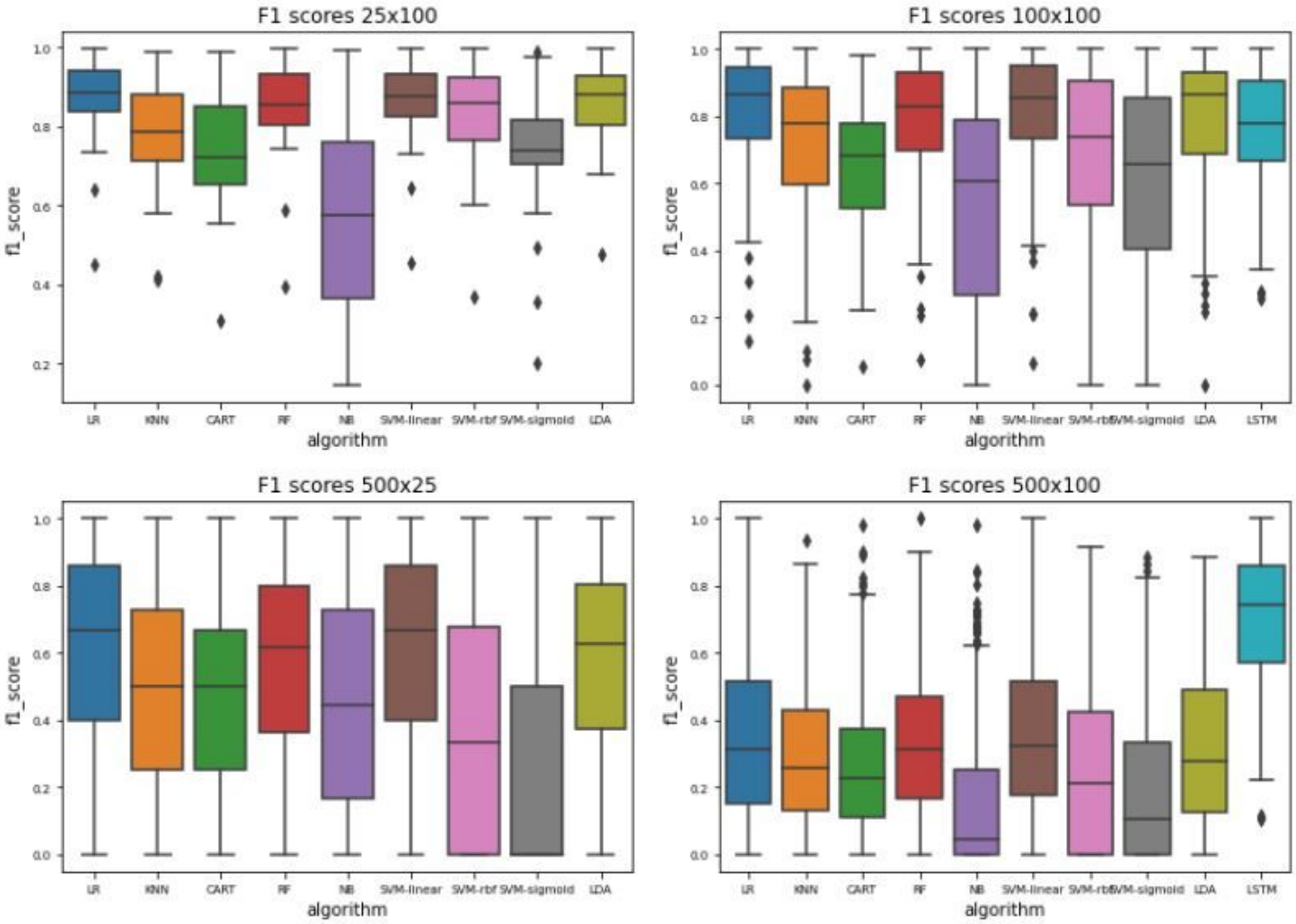

Figure 7

F1 scores per algorithm for the 25x100 problem, 100x100 problem, 500x25 problem and 500x100 problem. LR=Logistic regression;KNN=K-nearest neighbours; CART=Decision tree; RF=Random forest; NB=Naive Bayes; SVM=Support vector machine; LDA=Linear discriminant analysis; LSTM=Long short-term memory network

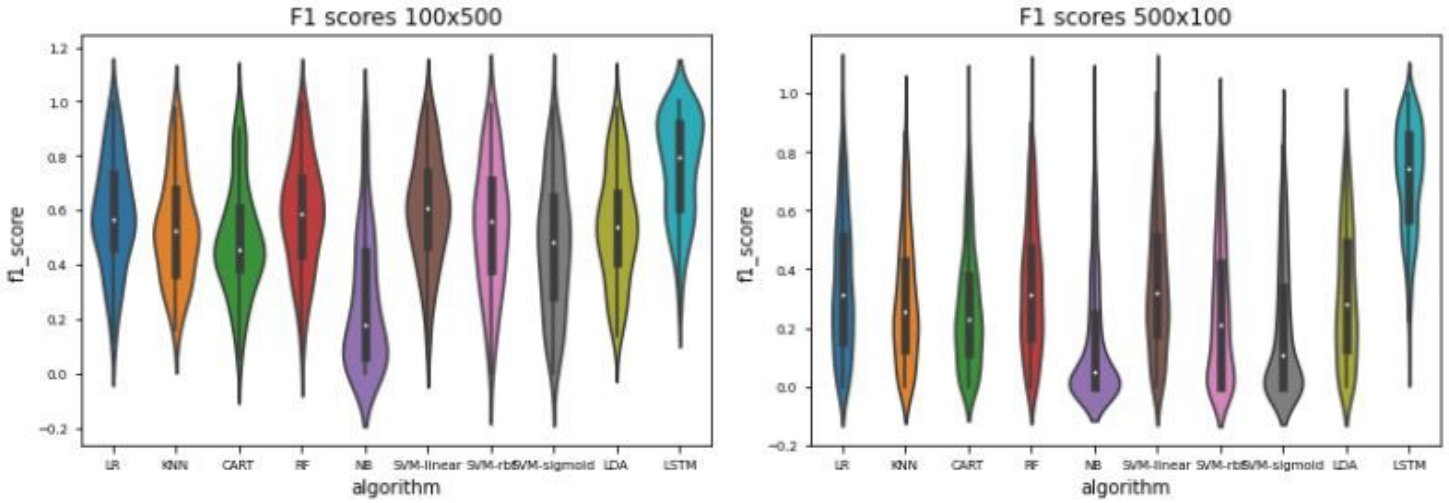

Figure 8

Violin plot of F1 scores per algorithm for the 100x500 problem (left) and the 500x100 problem (right). LR=Logistic regression;KNN=K-nearest neighbours; CART=Decision tree; RF=Random forest; NB=Naive Bayes; SVM=Support vector machine; LDA=Linear discriminant analysis; LSTM-Long short-term memory network

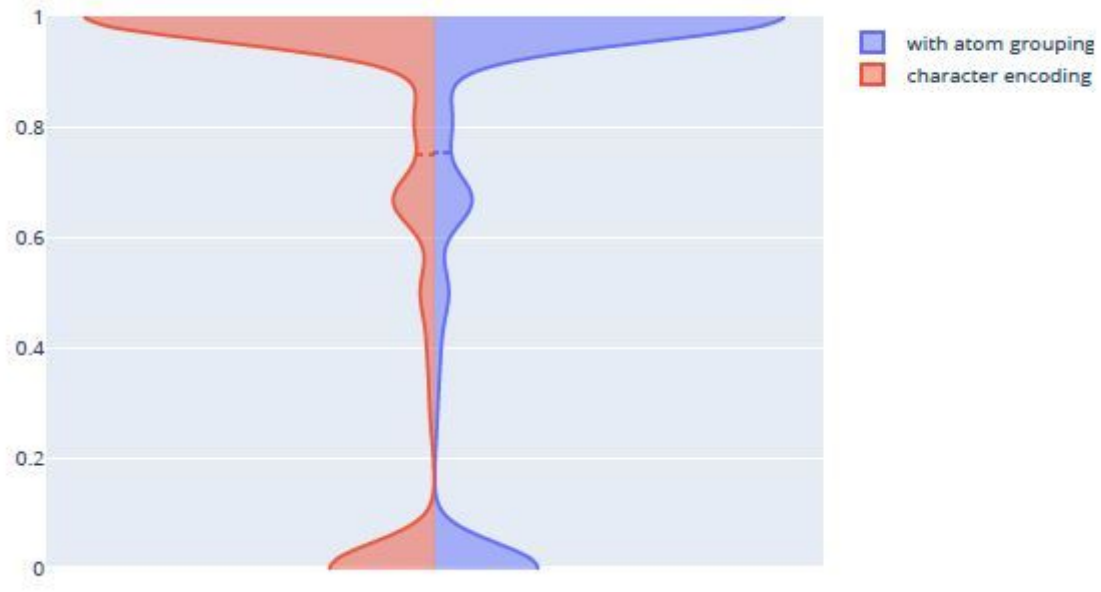

Figure 9

Violin plot of F1 scores on molecules in test set after 100 epochs

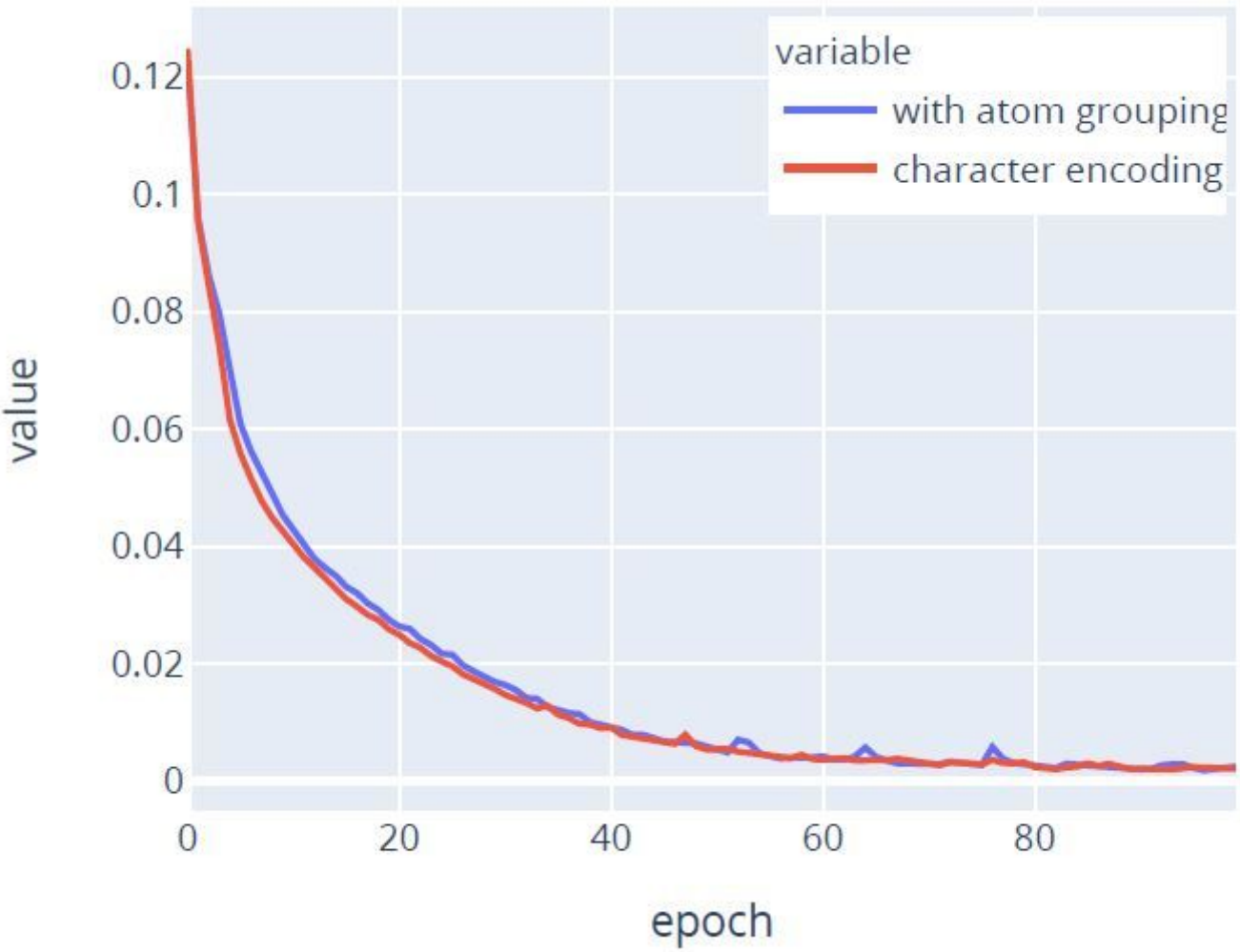

Figure 10

Loss on training data

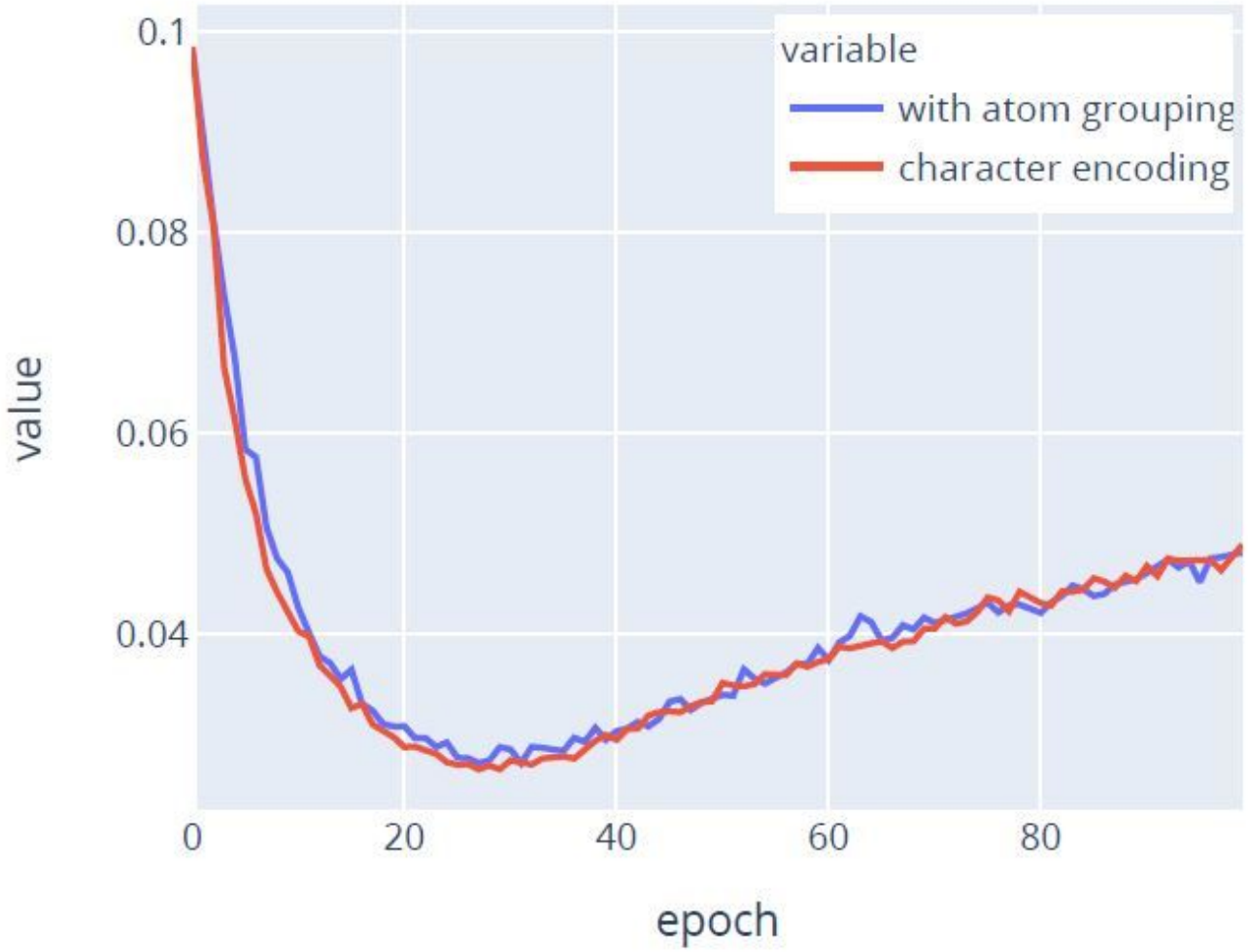

Figure 11

Loss on validation data

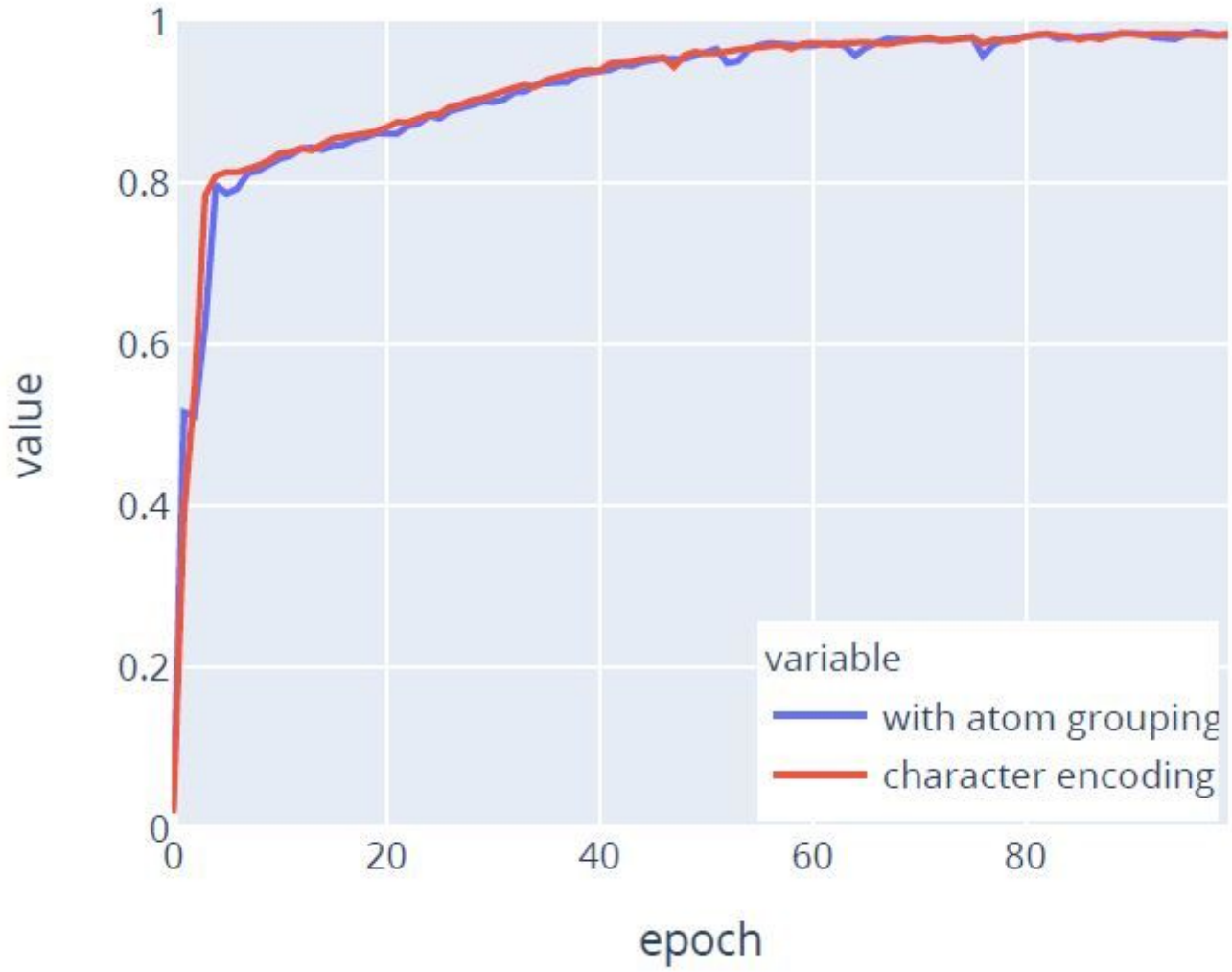

Figure 12

Precision on training data

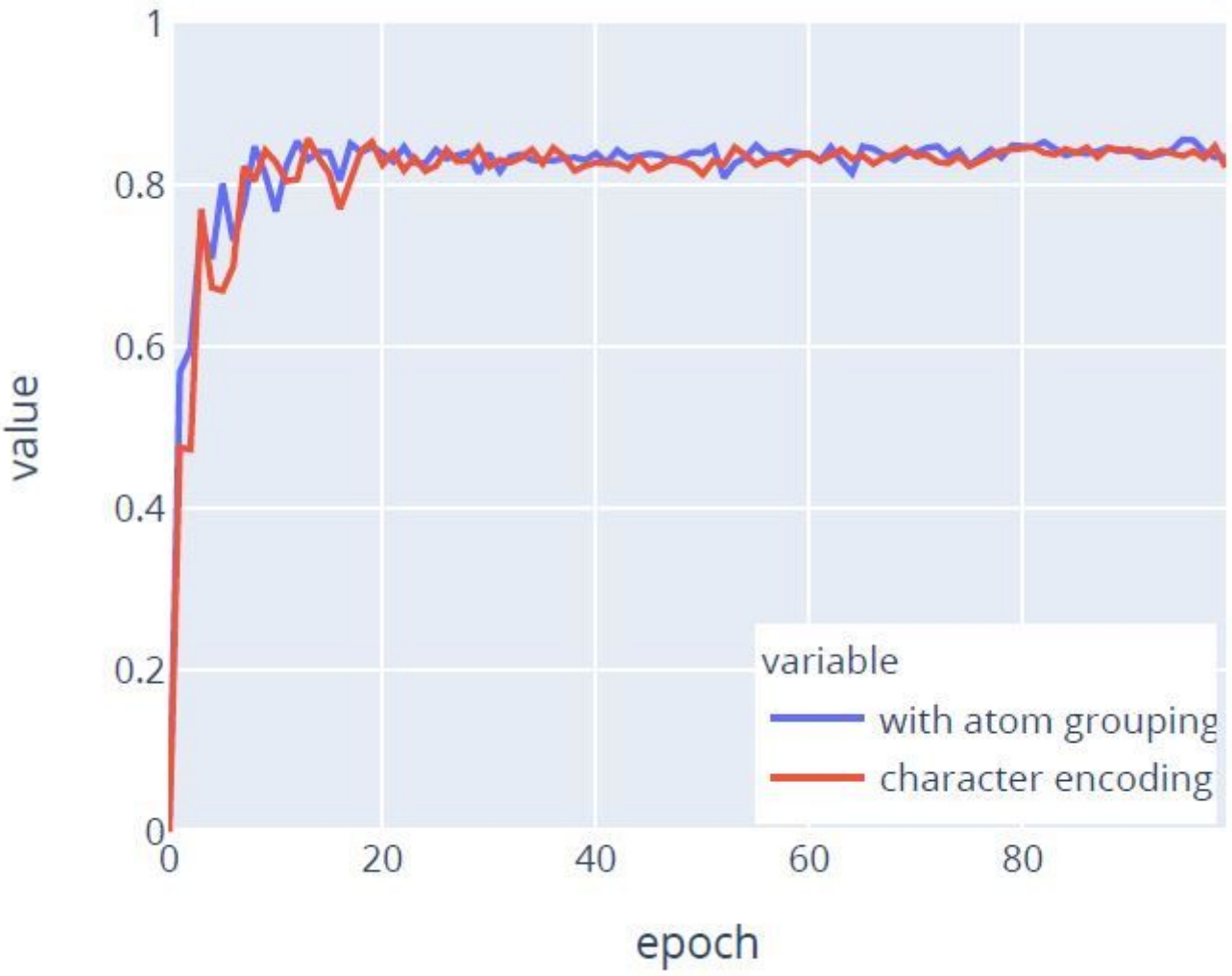

Figure 13

Precision on validation data

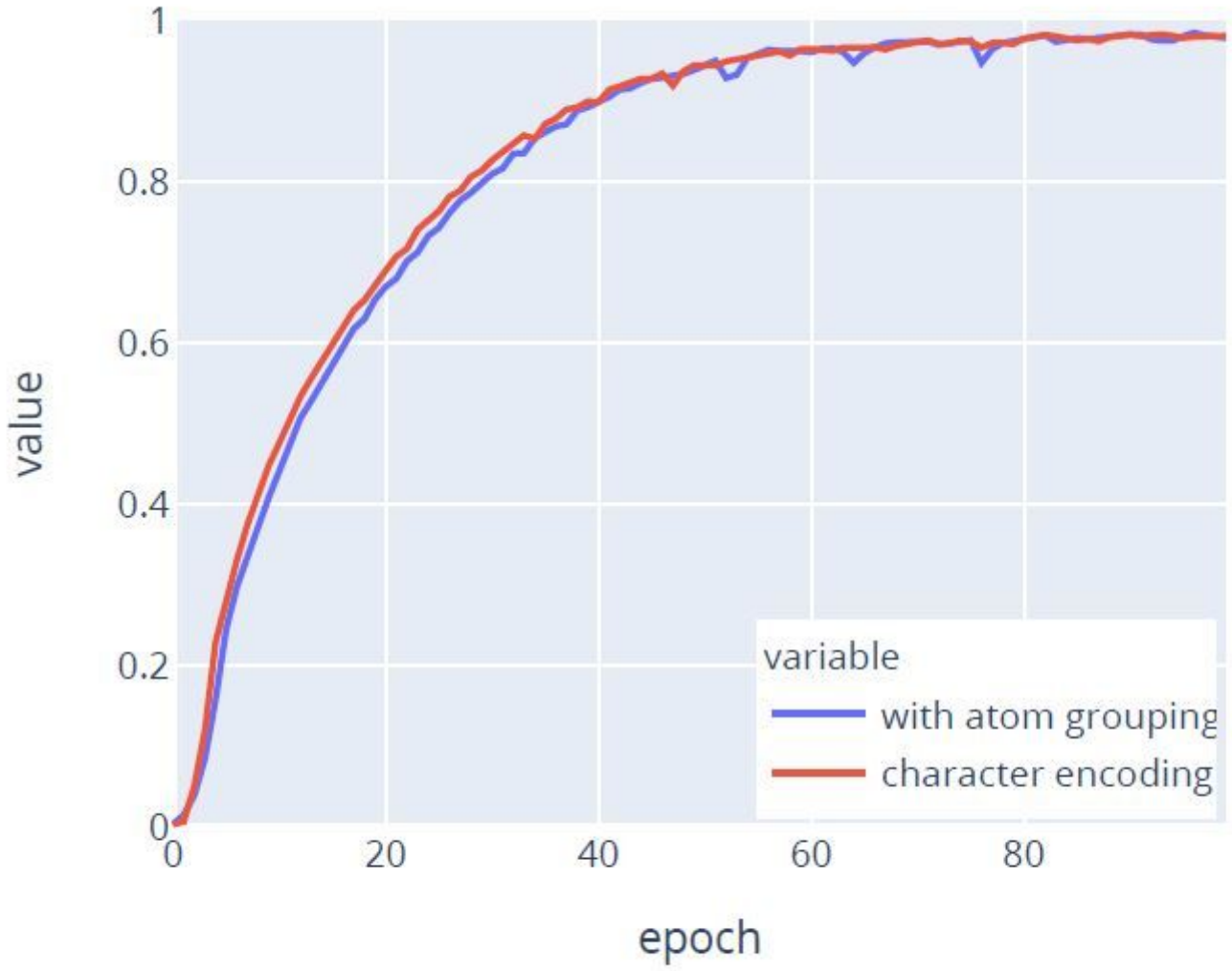

Figure 14

Recall on training data

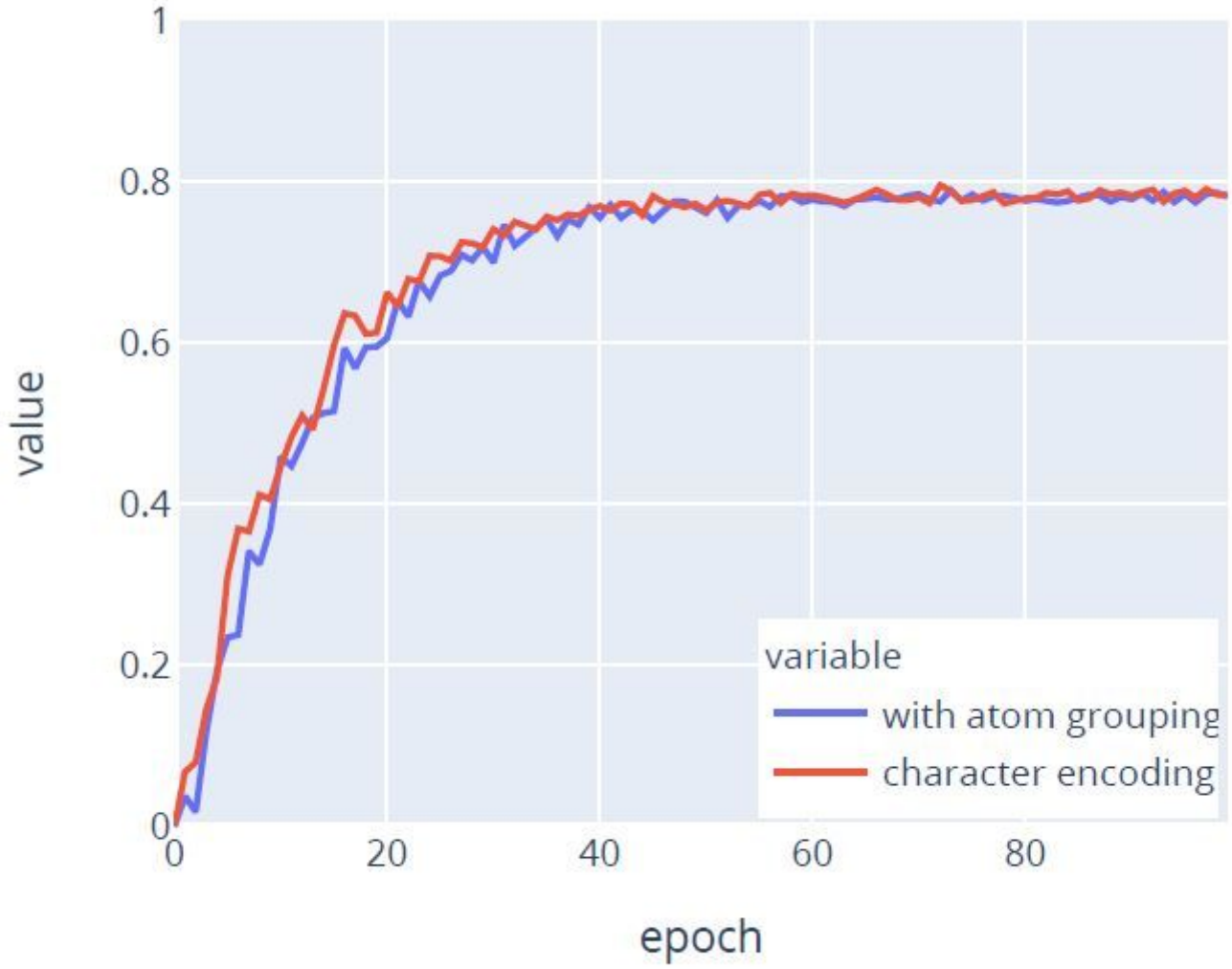

Figure 15

Recall on validation data

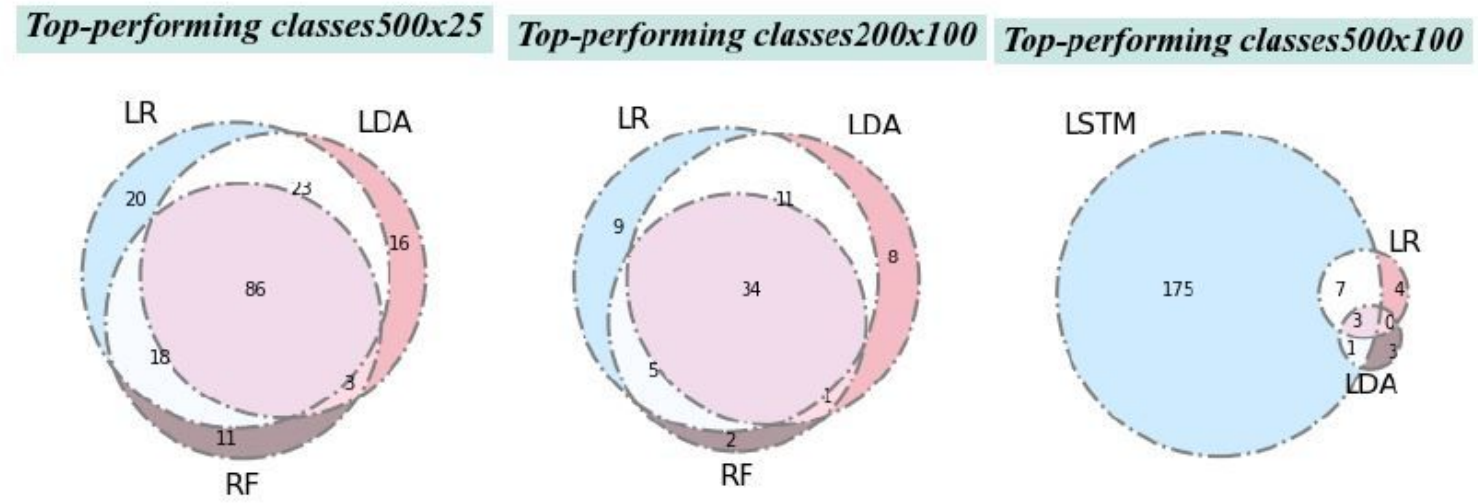

Figure 16

The Venn diagrams show the overlap of classes which scored F1 more than 0.8 (i.e., best-performing classes) for three of the classiers in each of these problem sizes. LR=logistic regression; RF=random forest; LDA=linear discriminant analysis; LSTM=long short-term memory

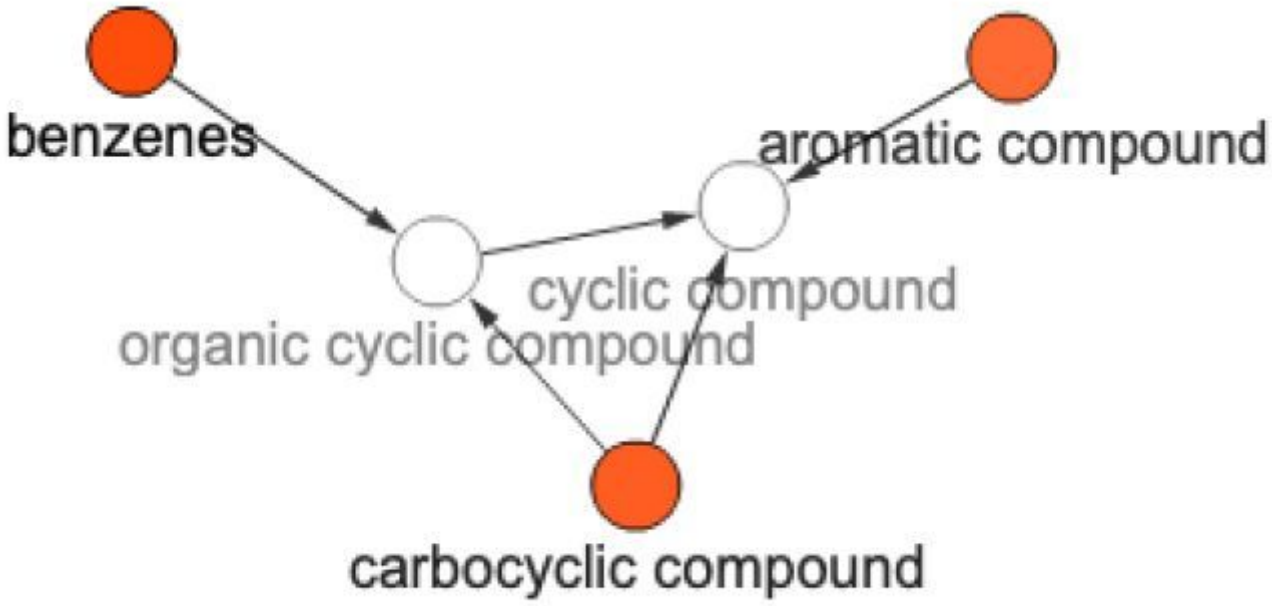

Figure 17

Enrichment analysis result on the ChEBI structural ontology for the 50 worst-performing classes in the LSTM