Curating inorganic chemical datasets to train RNN and transformer ML models to predict IUPAC names from InChI

Thomas Allam - ta1u18@soton.ac.uk

Skills 4 Scientists

Sponsored by: optibrium, dotmatics knowledge solutions

Project Outline

- Improve InChI to IUPAC name prediction for inorganic compounds by training models on larger inorganic datasets.
- Compare effectiveness of recurrent neural network (RNN) and transformer ML models

Background

- Large chemical databases contain discrepancies between inorganic structures and IUPAC names[1]
- Previous work in this area[2] predicted inorganic IUPAC names to 71% accuracy compared to an overall accuracy of 95% (when organic compounds were included)

Methods

- Cleaned and curated datasets totalling 1.2million compounds; containing InChIs (reconnected layers), SMILES and IUPAC names
- Split molecules into inorganic 'types' using SMARTS queries (table 1)
- Datasets used to train RNN models in TensorFlow

Type of inorganic molecule	Final validation accuracy	Epoch
Inorganic Organic Mix	InChI-86% Reconnected-86%	25
Pure Inorganic	InChI-84% Reconnected-84%	50
Organometallic	InChI-83% Reconnected-82%	50

Table 1: Validation accuracy of the models

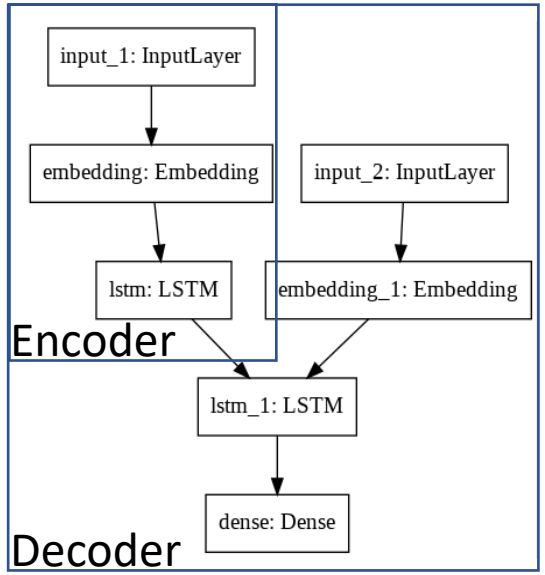

Figure 1: Architecture of the RNN model

Expected name	Most accurate prediction	Training Dataset
bis[(1,2,3,4,5-η)-cyclopentadienyl]iron	bis(triphenylphosphane) chromium	Organometallic Reconnected
Hexaamminecobalt(III) chloride	triammonium hexachlororutheniumdiuide	Inorganic Reconnected
bromo(methyl)magnesium	bromo(ethyl)mercury	Organometallic InChI
butyllithium	pentan-2-yl lithium	InorganicOrganicMix InChI

Table 2: Prediction of IUPAC names through RNN model that were incorrectly predicted in previous work[2]

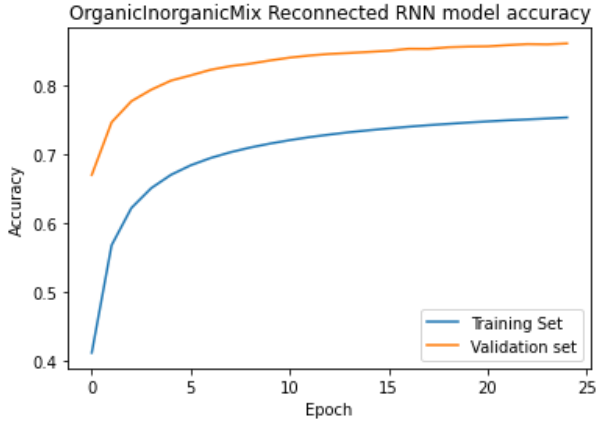

Figure 2: Training accuracy of RNN model over time

Results

- Overall the RNN models show an average validation accuracy of 84.5%
- Models improved on the accuracy of previous work (71%)[2] for inorganics despite limited training.

Further Work

- Establish if using reconnected InChIs will provide improvement when training models by increasing epochs
- Compare RNN with transformer models

References

1) Akhondi, S. A.; Kors, J. A.; Muresan, S. Consistency of Systematic Chemical Identifiers within and between Small-Molecule Databases. Journal of Cheminformatics 2012, 4 (1), 35. https://doi.org/10.1186/1758-2946-4-35.J. 2)

2) Handsel, B. Matthews, N. Knight and S. Coles, Translating the Molecules: Adapting Neural Machine Translation to Predict IUPAC Names from a Chemical Identifier , DOI:10.26434/chemrxiv.14170472.v1.

AI³ Science Network+

PSDS PHYSICAL SCIENCES DATA-SCIENCE SERVICE

RSC INTEREST GROUP CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP