Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Curating a chemical dataset to train recurrent neural network models to predict IUPAC names from InChI's

Project Report

Project Dates: 26/07/2021– 20/09/2021

University of Southampton

Project Student: Thomas Allam, University of Southampton

Supervised by: Professor Simon Coles, University of Southampton

Dr Jennifer Handsel, Data Science Service

Report Date: 20/09/2021

AI3SD-Intern-Series:Report-7_Allam

Curating a chemical dataset to train recurrent neural network models to predict IUPAC names from InChI’s

AI3SD-Intern-Series:Report-7_Allam
Report Date: 20/09/2021
DOI: 10.5258/SOTON/AI3SD0148
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1
Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Table of Contents

1. Project Details

Title	Curating a chemical dataset to train recurrent neural network models to predict IUPAC names from InChI's
Project Reference	AI3SD-FundingCall3_008
Supervisor Institution	University of Southampton
Project Dates	Project Dates: 26/07/2021– 20/09/2021
Website	N/A
Keywords	N/A

2. Project Team

2.1 Project Student

Mr Thomas Allam	
University of Southampton	
ta1u18@soton.ac.uk	

2.2 Project Supervisor

Dr Jennifer Handsel
Physical Sciences Data-science Service

2.3 Researchers & Collaborators

Professor Simon Cole, Professor of Structural Chemistry at the University of Southampton
Dr Jennifer Handsel, Physical Sciences Data-science Service.

3. Lay Summary

In this project, we built machine learning models to predict International Union of Pure and Applied Chemistry (IUPAC) chemical names from International Chemical Identifiers (InChI) which encode molecular information for use in indexing and sorting datasets[1]. The models were trained using newly curated inorganic chemical datasets totalling 1.2 million inorganic molecules split into 3 'types': Pure inorganic (70000 molecules), Inorganic organic mix (900000 molecules) and organometallic molecules (80000 molecules). Using these larger datasets we aim to improve on the 71% accuracy of the machine learning models by Handsel et al[2] for predicting IUPAC names from InChI's specifically of inorganics. The best models built for this project had a validation accuracy of 87.8% showing how the newly curated datasets show great promise for this kind of work in the future.

4. Aims and Objectives

Currently, there are discrepancies between large chemical databases between inorganic structures (contained within MOL files) and chemical identifiers with the lowest observed consistency between MOL files and IUPAC names[3]. To improve on this, we aim to predict IUPAC names from InChI's (rather than MOL files) compounds by training recurrent neural network models on large datasets of inorganic molecules. The main aim of this project was to compile a large inorganic dataset (1.2 million inorganic molecules) from the available SDF files on the PubChem website[4] for training these RNN models.

We aim to verify that simple machine learning models trained on the larger specifically inorganic dataset in this project, can improve on the validation accuracy of the more complex models in the previous work by Handsel et al[2] for inorganic molecules (71%). The other way we can validate if these new models trained on the larger inorganic dataset are an improvement is to see if the new models can correctly predict 4 inorganic compounds that were noted in the work by Handsel et al as being predicted inaccurately. The IUPAC names for these compounds have been difficult for the previous machine learning models to predict. It is hoped that training the models on the larger inorganic datasets will solve this issue.

If the machine learning models meet the targets above it will go a long way to show how these types of machine learning techniques when trained on the correct dataset can be used to solve the issues with discrepancies between large chemical datasets.

5. Methodology

Curating the dataset-

The first step for this project was to curate a large chemical dataset of inorganic molecules to train recurrent neural network (RNN) machine learning models specifically for predicting IUPAC names. This was done by downloading all 160 million Spatial Data Files (SDF) containing the chemical structure metadata for the molecules within the PubChem database[4] [5]. The Cheminformatics library, OpenBabel, for Python[6] was used to convert the SDF files to Simplified Molecular Input Line Entry System (SMILES) to allow the molecules to be processed and obtain datasets contains types of inorganic molecules.

SMILES is one of the simplest ways to input chemical molecules. SMILES are important due to their integrations with large open-source cheminformatics libraries, OpenBabel[7] is a notable mention. They provide a simple way for the user to input a molecule in order for the molecule's information to be used within the library. Having SMILES strings present in datasets allow trained chemists to quickly identify the type of compounds within the dataset without explicitly seeing the molecule[8] [9].

SMILES are particularly important to the curation of the inorganic dataset as they allow the user to use SMARTS queries. SMARTS queries were used in order to specify the substructure of the molecules using OpenBabel. For example, SMILES strings containing carbon-carbon (SMARTS query = [#6]~[#6]) or carbon-hydrogen (SMARTS query = [#6!H0]) were considered organic and outputted to a separate text file. Therefore, molecules that don't pass these criteria are inorganic and are refined by further SMARTS queries until the molecule is categorised as either pure inorganic, inorganic organic mix or organometallic molecules.

As mentioned above InChI's or international chemical identifiers encode molecule information for use in indexing and sorting datasets[10]. Unlike SMILES InChI's are not easy for humans to understand as they follow a more complex set of rules. Both InChI's and reconnected InChI's were calculated using OpenBabel from the SMILES examples shown in Figure 1 below.

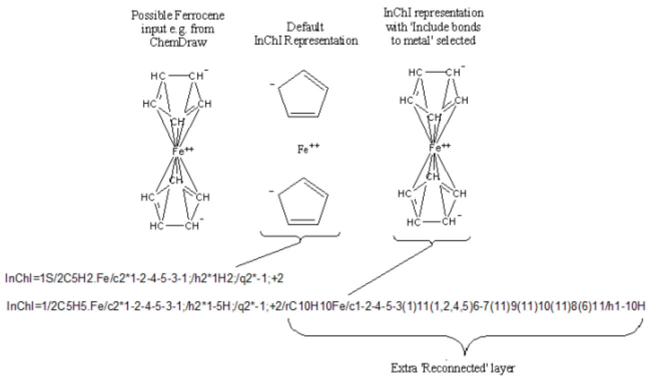

Figure 1: A figure to show the difference between InChI's and reconnected InChI's[11]

As the figure above shows reconnected InChI's provides more information about how organic groups bond to metals as well as transition metal centres within molecules. It is therefore hoped that after enough training, the models using reconnected InChI's as their input will see higher validation accuracies compared to using standard InChI's as the input.

The datasets are made up of SMILES, Standard InChI's, Reconnected InChI's and IUPAC names with an example from the Organic Inorganic Mix dataset is shown below:

IUPAC Name	SMILES	InChI	Reconnected InChI
oxorhenium tetramethanide	O=[Re].[CH3-].[CH3-] .[CH3-].[CH3-]	InChI=1S/4CH3.O.Re/h4*1H3;;/q4*-1;;	InChI=1/4CH3.O.Re/h4*1H3;;/q4*-1;;/r4CH3.ORe/c;;;;1-2/h4*1H3;/q4*-1;

These 3 datasets totalled 1.2 million molecules varying in complexity and size.

Building the models-

Once the datasets had been curated recurrent neural network (RNN) models were built to predict IUPAC names from InChI's. RNN models are a class of machine learning models that are very effective at predicting text[12]. The principles of a simple RNN model is outlined using Figure 2 below:

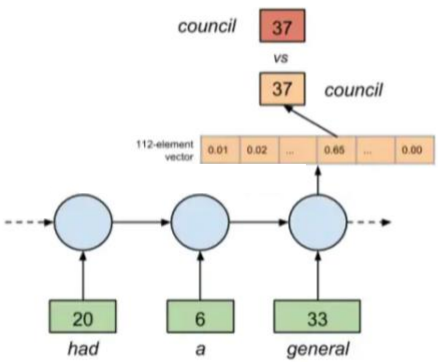

Figure 2: Shows the architecture of a standard RNN mode[13]

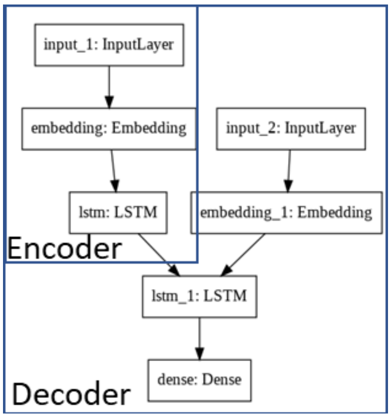

Figure 3: A graphic to show the architecture of the RNN model

In RNN models words (or characters in the case of this project) are the inputs however these need to be encoded into numbers or matrices before they can be understood by computers. For the model in figure 2, each word is assigned a unique number. For the model built in this project (shown in figure 3) each character in the input is one-hot encoded with by each character having a unique matrix the model can interpret.

These encoded inputs are passed to a hidden layer shown in blue in figure 2 where the next word can be predicted. In traditional neural networks, these hidden layers have no memory and will simply predict the next word or character based on the previous word or character alone. The models for this project use RNN models in conjunction with long short term memory (LTSM) networks as these overcome the no memory issue with previous neural networks[14]. Using RNN LSTM models combines the ability of RNN's to make predictions based on the structure of the sentence just before. For example, RNN's could predict the word 'green' given the input 'the grass is'. However, over longer character or sentence strings, this information can be lost. LSTM networks allow the model to retain the important information that would otherwise be lost. For example, if the input string begins with 'I love Italian culture' followed by a long block of text ending with 'My favourite food to eat while in' LSTM networks will retain the previous information about Italian culture and predict 'Italy'[15].

The prediction will be passed to the next layer. This layer is shown by the orange min/max vector in figure 2. The model will assign a probability to each character within the IUPAC name alphabet and will predict the character with the highest probability to be the next character in the sequence. For each character prediction, the model will compare its prediction to the correct character and a loss function will correct the model. The weights of the LSTM layers will be adjusted accordingly to make better predictions on the remaining epochs (the number of times the model is trained on the dataset). This process is repeated for the total number of epochs until the model is sufficiently trained. The final layer decodes the model output, put simply, this reverses the encoding process and outputs the predicted character.

6. Results

The models were then trained either using the standard or reconnected InChI's, on each 'type' of inorganic dataset. The results are in the form of validation accuracies (how well the model predicts a subset of data it is not trained on) and these are shown below:

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

As is shown above despite the lower number of epochs, the validation accuracy for predicting these inorganic compound IUPAC names from InChI's of 84% (average) is higher than that of the work by Handsel et al of 71%. However interestingly the standard and reconnected InChI's have displayed little difference in validation accuracy. This may be due to the small number of epochs. This means that the models do not have enough training to make use of the extra information provided by the reconnected InChIs.

As shown by the organometallic results, the more complex nature of the reconnected InChI's provides a slightly worse validation accuracy. This is despite reconnected InChI's providing more information to this 'type' of inorganic compound. This is due the nature of the metal to organic group bonding in this 'type' of inorganic compound.

These models were then tested by predicting IUPAC names from InChI's for Inorganic compounds inaccurately predicted by the previous models trained on the more generalized dataset in the previous work. The results are shown below:

Expected name	Most accurate prediction	Training dataset that provided the most accurate prediction
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

None of the models could correctly predict the expected IUPAC names though this may be expected due to the models limited training (small number of epochs). They do show the expected naming structures. For example, butyllithium was predicted to be pentan-2-yl lithium. Though the organic fragment is incorrect, the structure of the name shows how the model has learned the correct naming structure for IUPAC names when predicted from the InChI's.

Finally, the RNN model was trained on a dataset containing all of the 'types' of inorganic molecules (Inorganic Organic Mix, Pure Inorganic and Organometallic) used previously. This was to see if given enough training time the broader models could accurately predict the 4 difficult to predict IUPAC names from the previous work. The models were trained until there were 3 consecutive epochs without improvement and the model was deemed to be optimized for this architecture. The table below shows validation accuracy of the models trained on either Reconnected or Standard InChI's:

Type of inorganic molecule	Final validation accuracy	Epoch
All ‘types’ of inorganic molecules	Standard-86.9% Reconnected-87.8%	Standard-77 Reconnected-95

As expected, when the models were trained to their maximum optimization given the dataset, the model trained using Reconnected InChI’s had a higher validation accuracy and was required to train for a larger number of epochs, when compared to the model trained on the standard InChI’s. This was due to the extra information contained within the reconnected InChI’s and their more complex nature. The predictions using these models are shown below:

Expected name	Reconnected InChI prediction	Standard InChI prediction
bis[(1,2,3,4,5-η)-cyclopentadienyl]iron	bis(triphenylphosphane) iron	bis(triphenylphosphane) iron
hexaamminecobalt(III) chloride	trichlorochromium hexahydrate	cobalt tetrahydrate trichloride
bromo(methyl)magnesium	mercury hydrochloride	mercury hydrobromide chloride
butyllithium	but-2-en-1-yl lithium	butan-1-yl lithium

Asides from the bromo(methyl)magnesium where the prediction became more inaccurate. Overall, the model trained on the broader dataset made similar predictions when compared to the best dataset-specific model. This shows how using the dataset containing all ‘types’ of inorganic molecules creates more generalizable models.

7. Conclusions & Future Work

Overall, a large dataset of inorganic molecules has been compiled in this work. Along with this, a similar and much larger organic dataset was procured by the same method. The inorganic dataset has proved a very effective set of data to train machine learning models. It is hoped that these datasets will provide a good grounding to train more complex models involving InChI’s and IUPAC Names.

In future work, building machine learning models with more complex architecture such as transformer models could be used to improve the predictions and validation accuracy of the results.

Overall, as expected, the models trained on the broader purely inorganic datasets have high validation accuracies. This demonstrates, that despite the relativity simple architecture of the model, the datasets curated in this project provide an excellent grounding for training models of this type.

8. Outputs, Data & Software Links

The inorganic dataset as well as the RNN model are available at https://github.com/ta1u18/Curating-a-chemical-dataset-to-train-recurrent-neural-network-models-to-predict-IUPAC-names.

9. References

(1) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *Journal of Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.
(2) Handsel, J.; Matthews, B.; Knight, N.; Coles, S. Translating the Molecules: Adapting Neural Machine Translation to Predict IUPAC Names from a Chemical Identifier. **2021**. https://doi.org/10.26434/chemrxiv.14170472.v1.
(3) 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.
(4) Index of /pubchem/Compound/CURRENT-Full/SDF https://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/ (accessed 2021 -09 -06).
(5) The SDfile Format http://depth-first.com/articles/2020/07/13/the-sdfile-format/ (accessed 2021 -09 -06).
(6) Chun, W. *Core Python Programming*; Prentice Hall Professional, 2001.
(7) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *Journal of Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.
(8) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155. https://doi.org/10.1039/C9CC05122H.
(9) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.
(10) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, *7* (1), 23. https://doi.org/10.1186/s13321-015-0068-4.
(11) Technical FAQ. *InChI Trust*.
(12) Sutskever, I.; Martens, J.; Hinton, G. Generating Text with Recurrent Neural Networks. 8.
(13) Atienza, R. LSTM by Example using Tensorflow https://towardsdatascience.com/lstm-by-example-using-tensorflow-feb0c1968537 (accessed 2021 -09 -08).
(14) Dahl, G. E.; Dong Yu; Li Deng; Acero, A. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20* (1), 30–42. https://doi.org/10.1109/TASL.2011.2134090.
(15) Understanding LSTM Networks -- colah's blog http://colah.github.io/posts/2015-08-Understanding-LSTMs/ (accessed 2021 -09 -08).