Understanding Atmospheric Chemistry using Graph-Theory, Visualisation and Machine Learning.

Dan Ellis

March 2020

Veritatem inquirenti, semel in vita de omnibus, quantum fieri potest, esse dubitandu:
In order to seek truth, it is necessary once in the course of our life, to doubt, as far as possible, of all things.

- Descartes, Rene, *Principles of Philosophy*

Contents

Chapter 1

Computational Learning, Visualisation and Clustering:

Learning species structure using unsupervised machine learning.

"So, in the interests of survival, they trained themselves to be agreeing machines instead of thinking machines. All their minds had to do was to discover what other people were thinking, and then they thought that, too."

- Kurt Vonnegut, *Breakfast of Champions*

1.1 Introduction

Historical significance

The established process of trial and error has always underpinned our survival [Noble, 1957]. Babies are born to rely on a set of sensory reflexes and a framework for physical reasoning [Baillargeon and Carey, 2012], and with these, we develop methods to navigate the influence of change within a physical, and auditory space [Lynch, 2011]. This method of decision making is reflected in our adult lives with ideas and actions being limited in choice by our intuition and experience [Descartes and Lafleur, 1960]. In science, we apply a methodological framework consisting of a continuous assessment of scepticism, educated guessing (hypothesizing) and rigorous practical testing. Specialists accrue years of practical and theoretical knowledge within a narrow field and can identify areas of potential gain and futility. Yet even with all prior knowledge, the discovery of new and untested techniques involve the tortuous traipsing through a sea of uncertainty. Such a methods sometimes prove fruitful, through accidental discoveries of items such as x-rays, penicillin, etc. [Roberts, 1989]; finding novel applications for existing methods such as optical tweezers for chemistry or the abstract field of maths utilised by Einstein [REF], but more often than not end in the constant evolution of a pre-existing project with no clear result.

Theory and Simulation in Science

Until recently much of the experimentation possible was limited by resources, levels of knowledge available technology. With the increase of computation power, we have been able to not only increase our understanding but also run theoretical simulations to guide exploratory efforts with an impact on real-world applications [Oliveira et al., 2006; T. Leube et al., 2018; Morozov, 2016; Yu-ChenLo, 2018]. However, as our ability to record and produce data increases, the need for the scientific method diminishes [Anderson, 2008]. Here the application of ‘big data’ tools and algorithms can provide insights and correlations much more compelling than the predictive capabilities of constantly changing models - “Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration” - Box [1976]. As our level of attainable technology increases, so does the complexity of the data collected. Modern data-sets tend to be large, complex and highly multivariate. Although this greatly improves the quality of science that may be extracted from them, the difficulty lies in trying to represent it in such a way that we may successfully access the reliability of the results. Since simple bar and line graphs are no longer applicable, one solution falls within a class of unsupervised machine learning techniques called dimensionality reduction (DR).

Chapter Aims

In **??** we looked at visual representation as a way of understanding complex systems. **??** showed that the chemical properties could be visually inferred from the node-link graph structure of a mechanism. Similarly, **??** and **??** located the presence of important species and clusters of like properties by applying mathematical algorithms to the graph network. As opposed to attempting to visualise complex data, this chapter looks at learning the structure of a chemical species and simplifying it into two dimensions. Here it is possible to extract key features of like-groups through the use of vector clustering, which unlike the graph clustering in **??** works by determining the density between points on a plane.

The chapter begins with the introduction of the chemical system, and the various methods for representing species structure within it (section 1.2). Next, we define the dimensionality reduction methods which shall be used to simplify the aforementioned inputs (**??**). This is followed by a brief overview of the visualisation methodology (**??**). Finally, all three sections are combined to produce a set of result and conclusions about the use of DR to identify species structure.

1.2 Species of the MCM and ways to represent them.

The master chemical system (as defined in all previous chapters), represents our foremost knowledge of gas phase chemistry within the troposhere. It has been shown that due to its creation protocol (**??**), much of the information about a species structure is encoded within the reaction pathways it can take. This section explores the different methods of representing a species structure, with the aim of providing a machine built algorithm with the greatest amount of information about each species and its functionality. To do this a range of input types will be evaluated against a number of different dimensionality reduction algorithms with the aim of isolating which chemical properties are most 'picked up'.

1.2.1 Input generation

The MCM provides species information in the form of a species 'smiles' (subsubsection 1.2.3.2) and the IUPAC InChi string [Heller et al., 2013]. Within this chapter we use only the smiles string, which is either manually processed using regular expressions or with the aid of pythons RDKIT package [Landrum et al., 2019]. There are seven different methods for representing the chemsitry, each of which are outlined below.

1.2.2 Manual Categorisation

Reactions within the MCM are determined by a set of rules (PROTOCOL SECTION). These are designed to mimic the process a chemist my discover new species, and often rely on the bond availability and functionalisation of a species. Since the present functional groups are the benchmark of whether a DR algorithm has sucessfuly separated species structure, it make sense to run a unit test using the known functional groups of a species as the input.

To genearate the functional groups the regular expressions in Table 1.1 are used[1] on the smiles strings (described in subsubsection 1.2.3.2) for each species. In extracting the functional groups we are able to plot the likelyness a species with a certain group is likely to have another using a chord diagram - Figure 1.1. Since most species are found to contain a multitude of functional groups, the separation of these into 'tidy' clustered groups seems unlikely.

PAN	`C\\(=O\\)OON\\(=O\\)=O$|^\\[O-{0,1}\\]\\[N\\+{0,1}\\]\\(=O\\)OOC|O=N\\(=O\\)OOC\\(=O\\)|C\\(=O\\)OO\\[N\\+{0,1}\\]\\(=O\\)\\[O-{0,1}\\]`
Carb. Acid	`[^O](C\\(=O\\)O$|^OC\\(=O\\))`
Ester	`[\^O](C\(=O\)O\b|OC\(=O\))C`
Ether	`(\([\^O=]+\))*C(\([\^O=]+\))*O(\([\^O=]+\))*C(\([\^O=]+\))*`
Per. Acid	`c\\(=O\\)OO$|^OO\\(=O\\)C`
Nitrate	`O(NO2\b|NOO\b|N\(=O\)=O|\[N\+\](?:\[O-\]|\(=O\)){2})`
Aldehyde	`C=O$|^O=C`
Ketone	`C\(=O\)C`
Alcohol	`CO\\b|(?=^\\b)(?!^\\[)CO.|(?=^\\b)(?!^\\[)OC.|\\(O\\)|C\\)O(\\b|[^O]`
Criegee	`\[O-\]\[O\+\]`
Alkoxy rad	`\[[\/]{0,1}CH{0,1}\]|\b[\^O]\[O\.{0,1}\]`
Peroxyacyl rad	`\\ w\(=O\)O\[O\.{0,1}\]`

Table 1.1: CHECKKKKKKK!!!!!!!!! A set of regular expressions that may be used to determine the number of occurrences of a functional group within a SMILES string.

[1] To see the structure of each functional group type, go to **??**.

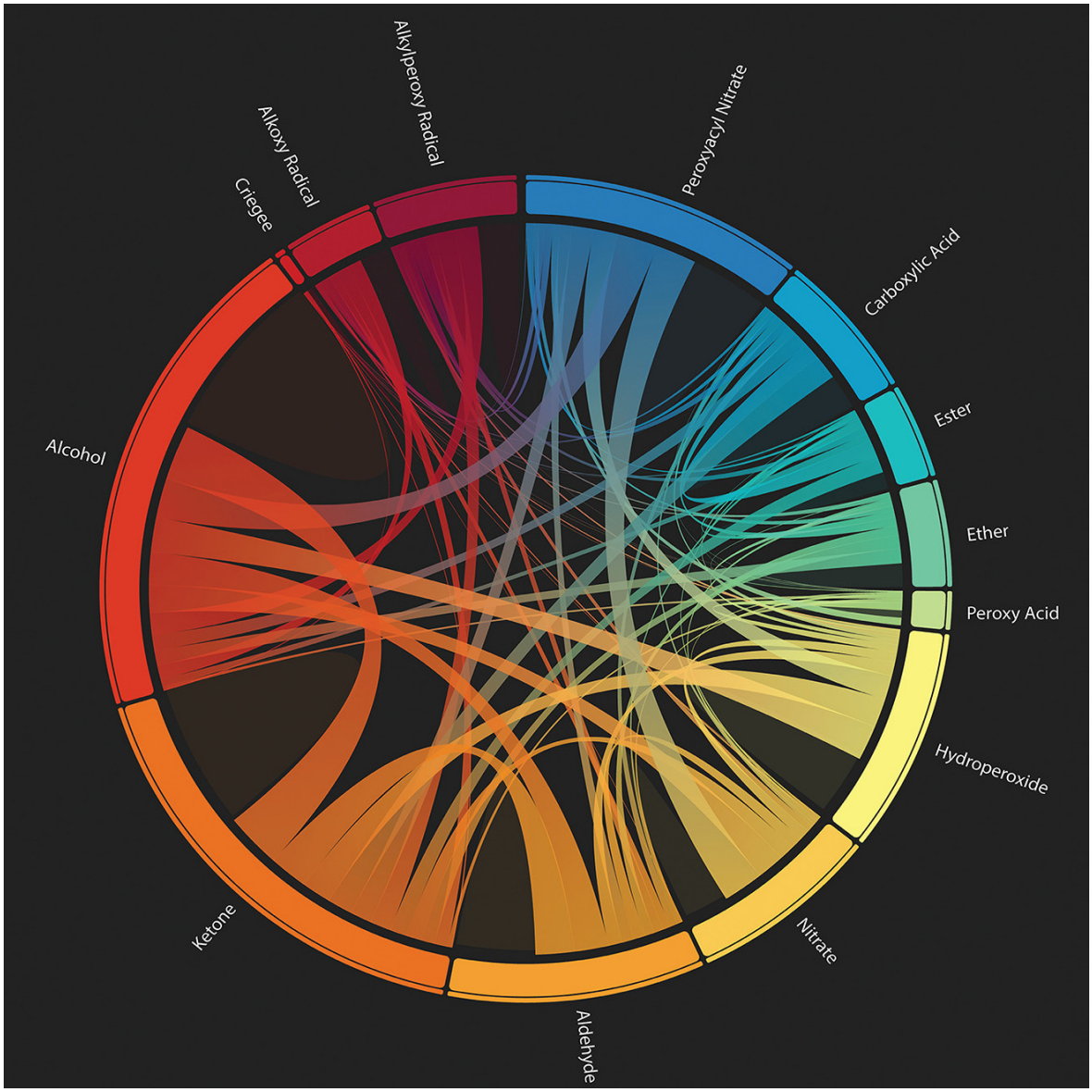


Figure 1.1: **The multifunctionality of the MCM.** A chord diagram showing the functionalisatoin of a species within the MCM. Arc sizes represent what percentage of all functional groups in the MCM mechanism a group contains. Translucent areas of no outwards links represent species with multiples of a certain functional group, of which Alcohols and Ketones have the most. Source: [Ellis, 2019]

1.2.3 Tokenization

As computer algorithms are unable to understand words, or their meaning, we have to first categorise the data into groups. Tokenisation is the conversion of a string into characters and representing them with a numerical equivalent. In doing so a string of characters can be converted into a numerical vector, allowing for its representation in a latent vector space. Within our input selection, we have two sets of inputs we can convert. These are the species names, and their smiles string representation.

1.2.3.1 Species Names

In **??** it was shown that the dedicated species names for species in the CRI mechanism were often representative of their structural properties. This addage also applies for the MCM, where an intitive naminc convention has been selected. This is is often derrived as part of the construction protocol, where a species names reflects its own, or its precursors structure (which it will have atleast in-part inherited).

Although this is not the most robust method of defining structure, it allows for an easy test of the algorithms, for which the user can quickly compare the human readable output.

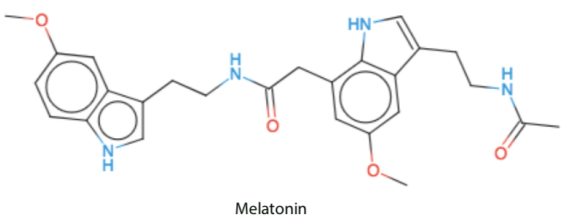
1.2.3.2 SMILES strings

Smiles ('Simplified Molecular-Input Line-Entry System') provide a human-readable representation of molecular structure, [Weininger, 1988]. They provide a linear human-readable representation of the chemical structure within a molecule. This makes it easy for us to visually check the structure of a species without any additional work. In addition their role in generating the molecular fingerprints in subsubsection 1.2.3.4 makes it a useful comparison to make when evaluating methods of structure representation.

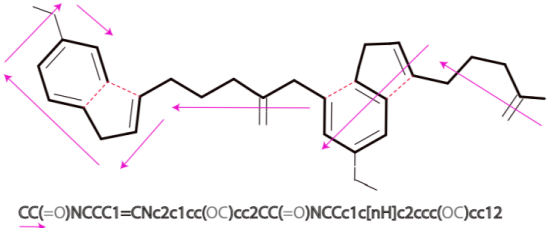
Construction Methodoly of SMILES strings

Smiles strings are constructed in three parts. We begin with a species backbone, then add break cycles and branches producing a smiles string. A visual description of this procedure is given below.

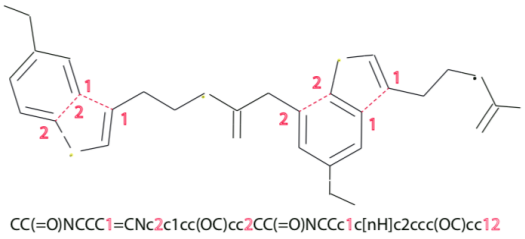
1. The smiles string is built by creating the longest possible chain to form a molecule backbone. Figure 1.2b
2. This may within itself contain aromatic rings denoted by the lowercase carbons and a number corresponding to the location of each break cycle. Figure 1.2c
3. Finally all the functional groups and branches attached to the main backbone are added. These are nested within parenthesis to show that they are not part of the skeletal backbone. Figure 1.2d



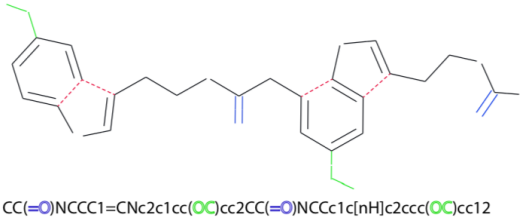
(a) Structure of Melatonin



(b) Step 1 : Building the C chain backbone.



(c) Step 2 : Aromatic Rings



(d) Step 3 : Functional Groups

Figure 1.2: **Construction process of a smiles string.** The example compound is Melatonin. Although this does not exist within the atmosphere, it provides a clear example of the smiles string methodology. Figure 1.2a is made using smiles drawer: [Probst and Reymond, 2018]

1.2.3.3 Graph Inspired

?? - **??** have shown the role of graphs in revealing network properties and structure. Graphs in themselves are able to simplify relational data into two/three dimensions for visualisation and algorithmic clustering. Continuing this trend we can represent a species strucuture in the form of a graph (section 1.2.3.3), as well as converting the structure of a mechanism for dimensionality reduction (Figure 1.2.3.3)

The species graph

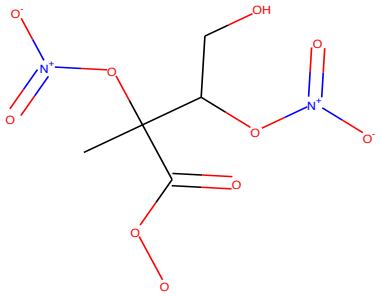
The structure of a species has long represented using a graph-like layout, **??**. It therefore follows that other methods for representing the graph structure would also apply. One such method is the use of an adjacency (or relational) matrix to describe the relationships between atoms and bonds in a species. Such a methodolgy is already used in the construction of bond and z-matrixes, [**?**Parsons et al., 2005].

The construction of a structure matrix/graph begins with a chemical species. Here the relationships between atoms (Figure 1.3b) is converted into an adjacency matrix (Figure 1.3c). However since species have different numbers of each atom, a template allowing us to compare different graphs is required. To do this a maximum occurance table (Figure 1.3a) is created. Here for example BCARY $C_{15}H_{24}$, a sesqueterpine contains the most carbon atoms of any species within the MCM. This universal matrix is now able to contain any possible combination of atoms in a species.

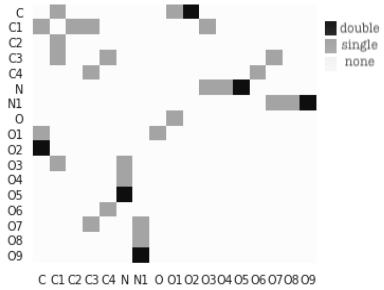
As machine learning algorithms only vectors as an input, it is possible to decompose the 37^2 element adjacency matrix into rows, which can then be joined together, Using this method we create a flat array (vector) of 259 elements (518 bytes) which can be used to represent our species.

Atom	Max
C	15
Cl	4
O	12
N	3
S	1
BR	2

(a)



(b)



(c)

Figure 1.3: **Constructing a graph from species structure.** (a) shows the maximum number of times an atom occurs for any single species in the MCM. (b) depicts the graph-like chemical structure of INB_1NBCO_3. This is a highly processed species stemming from Isoprene, and this makes for a good example of the bond matrix. Finally a matrix representing the bonds in INB_1NBCO_3 is created from the maximum possible occurance matrix from (a). For simplicity, empty row/column pairs have been removed to produce (c). This matrix will always be symmetrical as the bonds do not have a direction.

MCM graph: Node Embeddings

As shown in chapter YYY, representing the reactions within a mechanism in the form of a graph exposes patterns presented by the generation protocol, and thus the species chemical structure. We saw this through the categorisation of graph structures into aromatics, terpenes and an alkane/alkenes REF, and through van-krevelin ratios showing the progressive oxidation of species, to the production of CO_2 (CO in the mcm) and water, Figure 1.4.

We can make use of the fact that chemical structure is encoded within the mechanism through its construction process. To extract this we may use node2vec, [Grover and Leskovec, 2019], a program used to covert the structure of a graph into a numerical vector for use in machine learning. This is discussed in detail in **??**.

1.2.3.4 Molecular Fingerprints

Molecular fingerprints (structural keys) are a way of encoding molecule structure into a queryable series of binary digits. These are predominately used in the filed of chemical-informatics as means of exploring chemical space (a type of property space constructed using pre-determined properties and boundary conditions). Properties are often split into structural and psyico-chemical groups, allowing for the use of data mining techniques in search of molecule similarity (with uses such as the discovery of natural analogues to circumvent side effects and in-tolerances [Spahn et al., 2017]). Unlike line

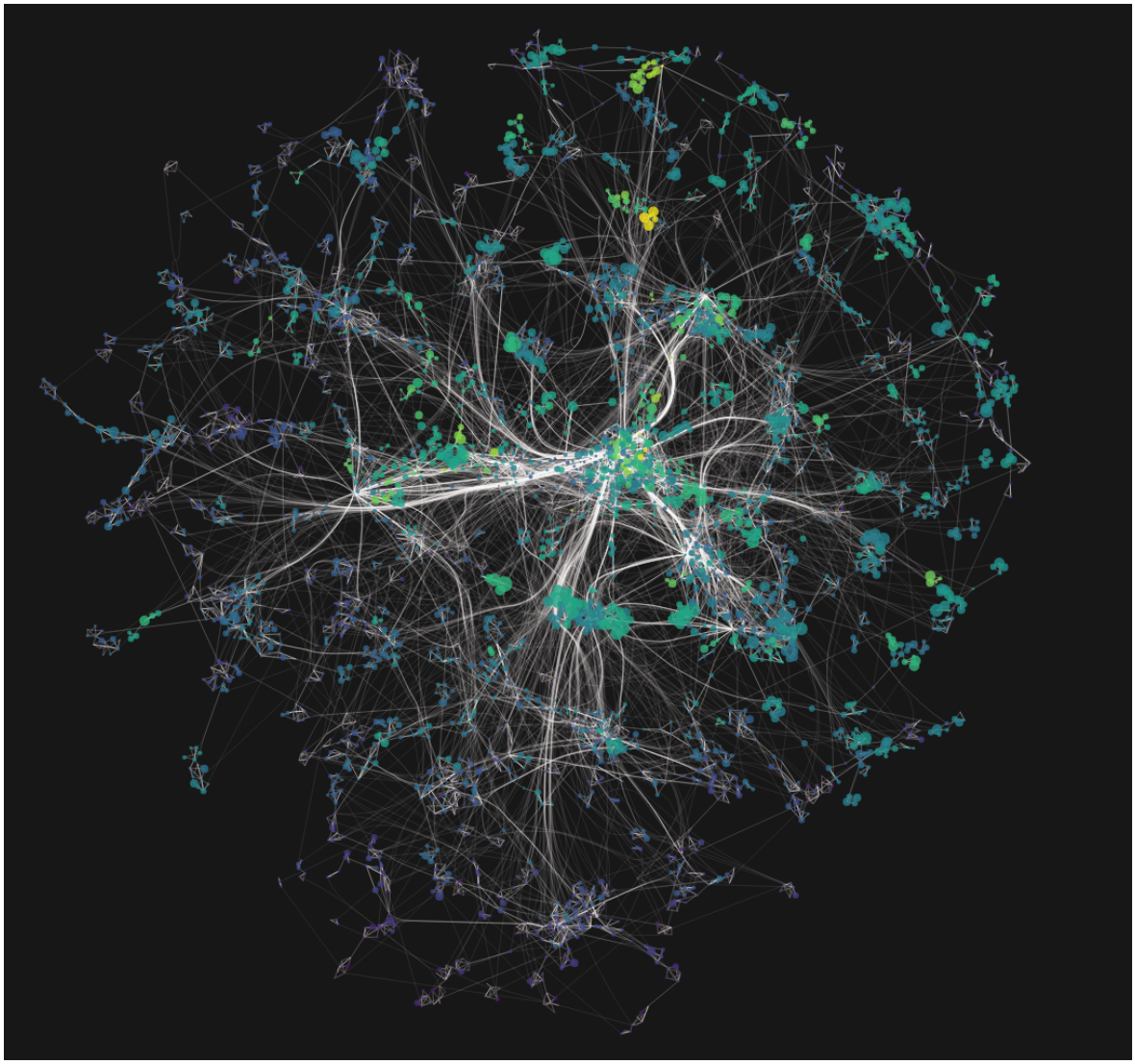


Figure 1.4: A graph representing the MCM mechanism derived from measured emissions in Beijing. This shows an increase of the O-C ratio as species are oxidised towards CO (center).

notations, such as smiles and InChi, molecular fingerprints provide a multi-dimensional classification for chemical species which makes them ideal for machine learning inputs.

Molecular Quantum Numbers (MQN)

In chemistry the shape, phase and electron occupancy of an atom may be described through the use of four quantum numbers[2]. The rationalisation of elements based on their structure, and by consequence reactivity, has led to the most iconic tool of the modern-day chemist - the periodic table[3] [Wang and Schwarz, 2009]. In representing a molecule as a set of 42 quantum numbers, MQN fingerprints produce a multi-dimensional mapping of atom, bond, polarity and topology count [Nguyen et al., 2009]. Its

[2]These are n principle quantum number, I angular momentum quantum number, M_i magnetic quantum number and M_s spin quantum number.

[3]Increasing atomic numbers follow the principal quantum number.

binary nature not only... application..

[ref fae others...]

Molecular ACCess System (MACCS)

MACCS keys are a 164[4] bit structural keys formulated through answering a series of structure-related questions. Developed by MDL Information Systems [, MDL], their main purpose lies in being a SMILES Arbitrary Target Specification (SMARTS) system for substructure searching. However their distinct structure key format

makes them highly suitable for similarity detection. In many cases, the optimised version of MACCS keys is cited ([Durant et al., 2002]), although most use cases exploit a variation of the undocumented 166bit keys. We use the implementation presented by [Landrum et al., 2019; rdkit, 2019] for all molecular fingerprints in this section.

1.3 Results

1.3.1 CLuster distribution

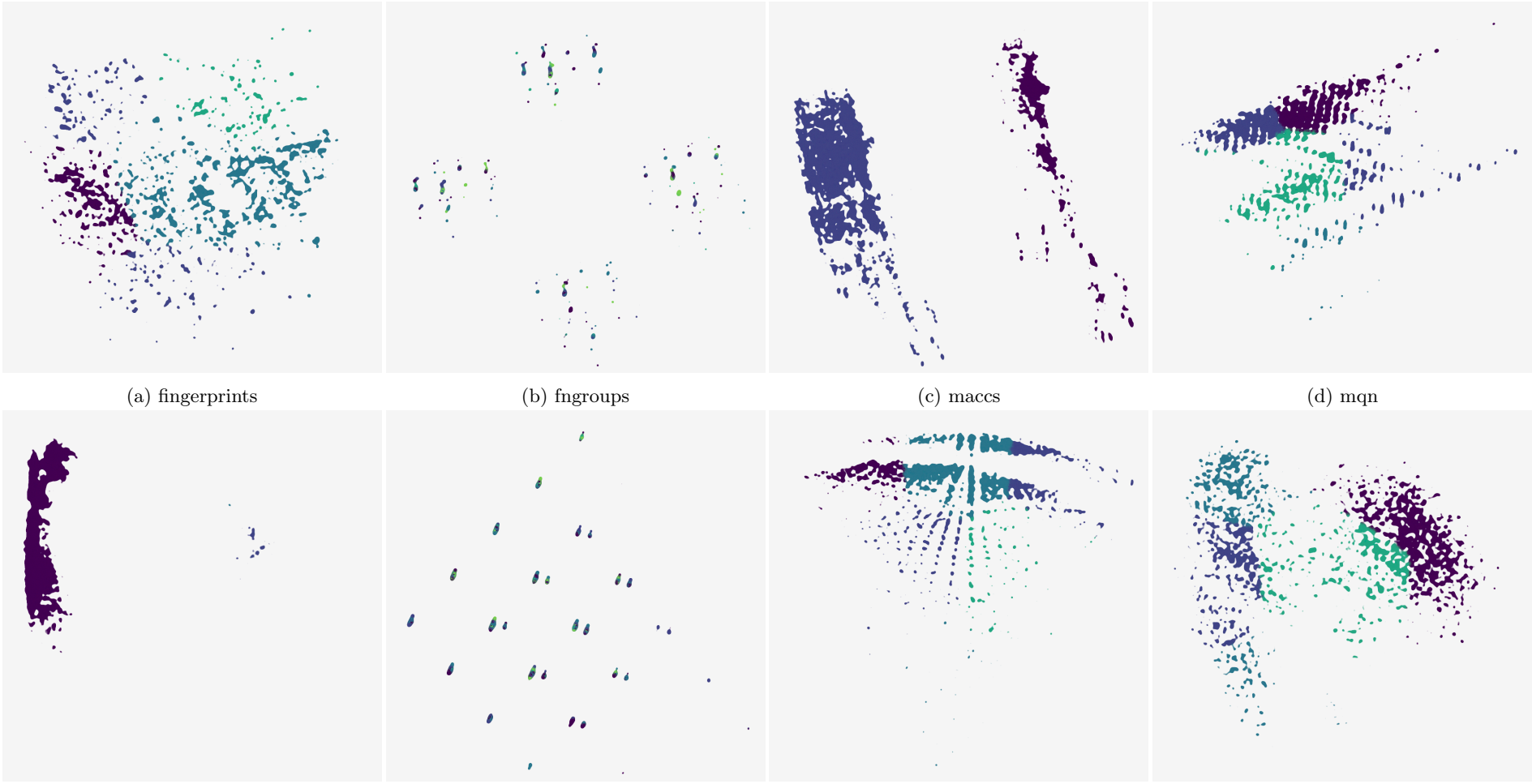
Start with the visual comparison and compare it with the silhouette values.

Principle Component Analysis

DR	input	silhouette	groups
PCA	fngroups	0.9122	141
PCA	protocol	0.8761	149
PCA	node2vec	0.8569	3
PCA	maccs	0.6563	2
PCA	mqn	0.4041	8
PCA	smiles	0.3648	6
PCA	fingerprints	0.3529	6
PCA	spec	0.3364	6

Table 1.2: The inputs to the PCA dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.

[4]Although they are 166-bit keys, there is no real agreement to what the 44th keys' purpose is, and therefore it is often omitted. Within RDKIT this is denoted by a ? [rdkit, 2019].



(a) fingerprints (b) fngroups (c) maccs (d) mqn

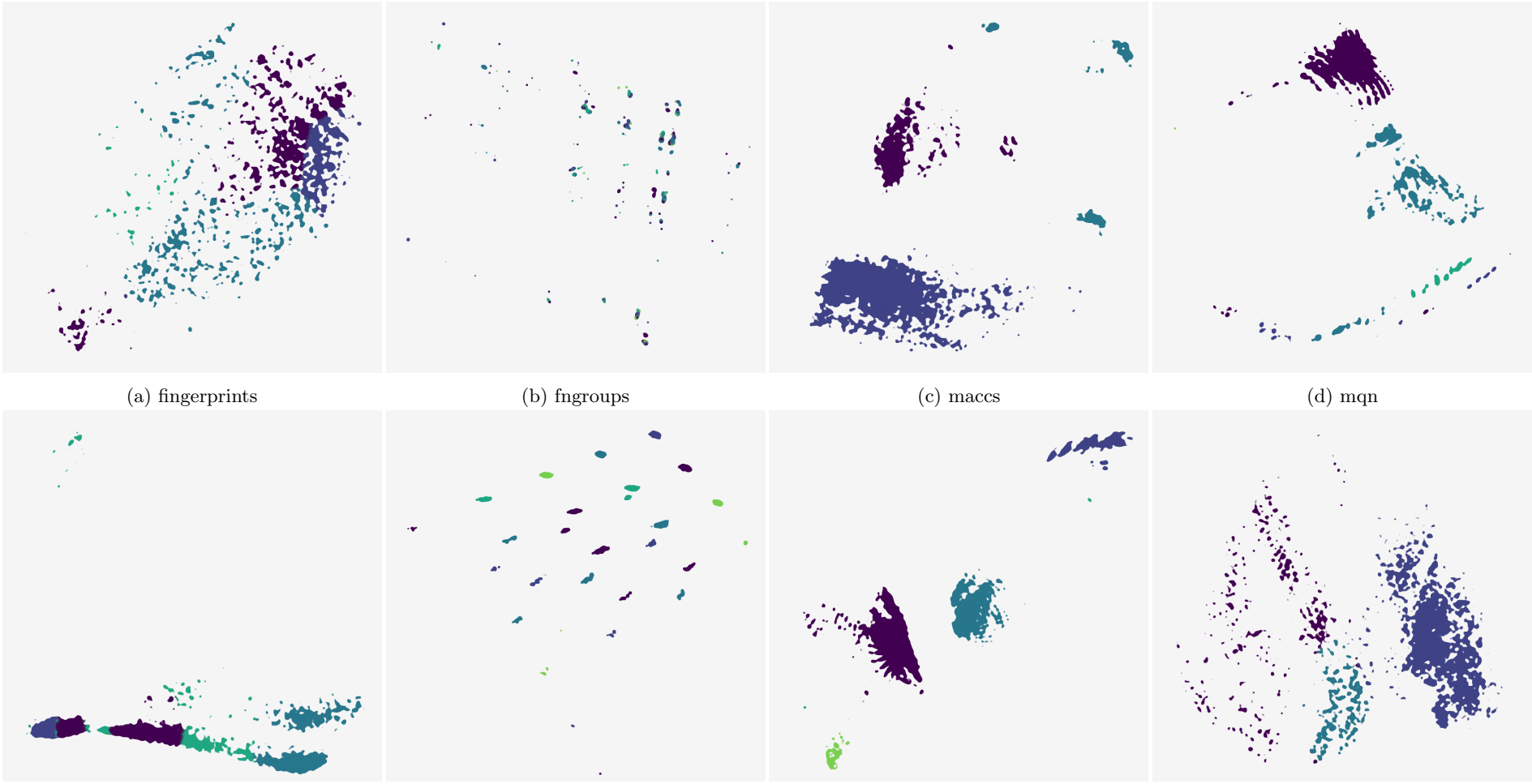
(e) node2vec (f) protocol (g) smiles (h) spec

Figure 1.5: **Comparing clusters for all inputs after a reduction to 2 dimsnsions using Principle Component analysis.** Each graphs has indergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient have been chosen. Colours follow the greedy 4 colour theorem and are there only to indicate contrast between cluster boundaries.

Auto Encoder Encoding

DR	input	silhouette	groups
AE	fngroups	0.9249	140
AE	protocol	0.8992	27
AE	smiles	0.6897	5
AE	mqn	0.6572	12
AE	maccs	0.6241	3
AE	node2vec	0.5476	5
AE	spec	0.4238	3
AE	fingerprints	0.3189	8

Table 1.3: The inputs to the AutoEncoder dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.



(a) fingerprints (b) fngroups (c) maccs (d) mqn

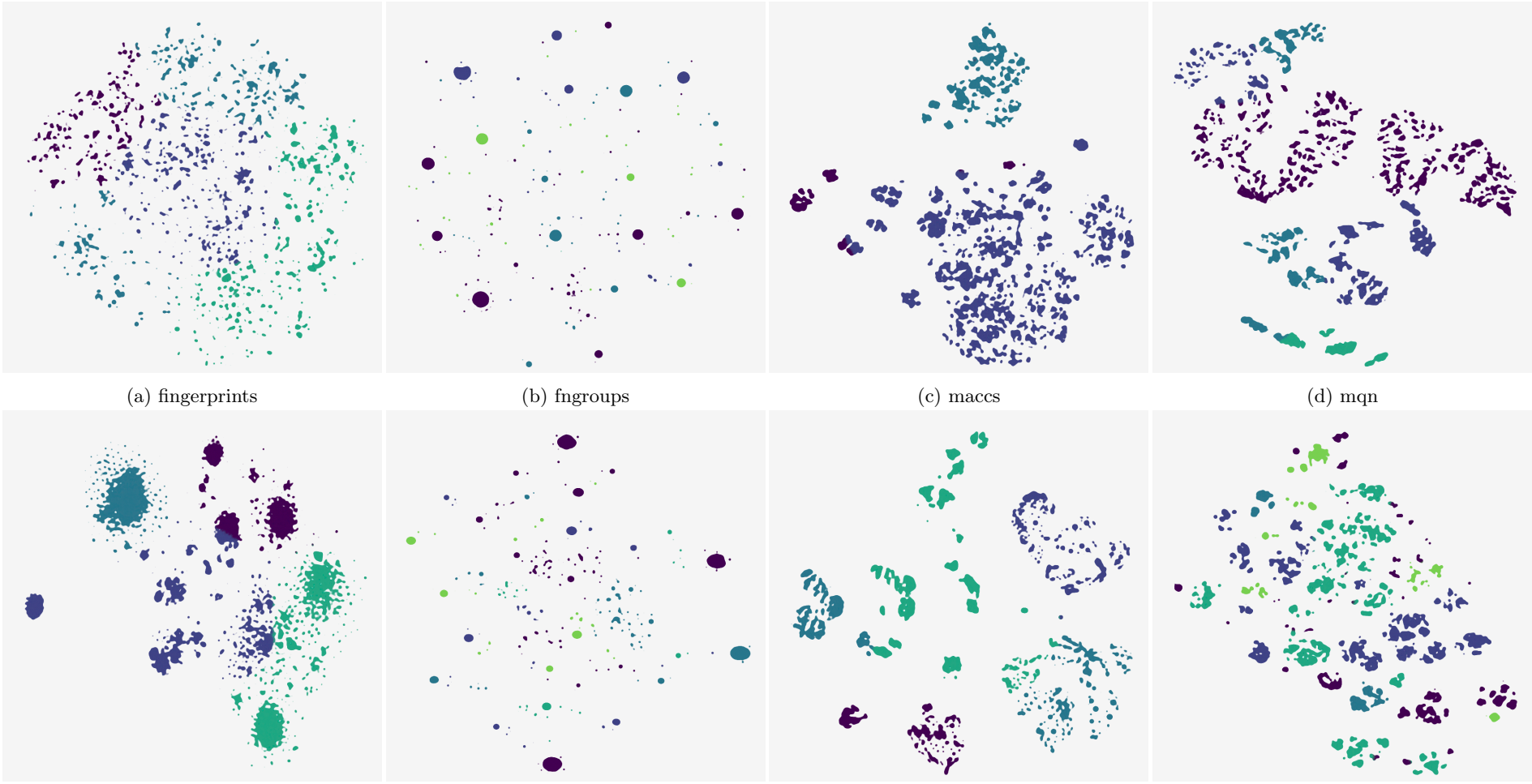
(e) node2vec (f) protocol (g) smiles (h) spec

Figure 1.6: **Comparing clusters for all inputs after a reduction to 2 dimsnsions using an AutoEncoder.** Each graphs has indergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient have been chosen. Colours follow the greedy 4 colour theorem and are there only to indicate contrast between cluster boundaries.

t-Distrubuted Stochastic Neighbor Embedding

DR	input	silhouette	groups
t-SNE	fngroups	0.7458	106
t-SNE	protocol	0.5688	51
t-SNE	smiles	0.4808	6
t-SNE	node2vec	0.4359	6
t-SNE	maccs	0.4295	3
t-SNE	spec	0.3781	35
t-SNE	mqn	0.3684	8
t-SNE	fingerprints	0.3539	6

Table 1.4: The inputs to the t-SNE dimensionality reduction algorithm sorted by the best obtained silhoette coefficient.



(a) fingerprints (b) fngroups (c) maccs (d) mqn

(e) node2vec (f) protocol (g) smiles (h) spec

Figure 1.7: **Comparing clusters for all inputs after a reduction to 2 dimsnsions using t-SNE.** Each graphs has indergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient have been chosen. Colours follow the greedy 4 colour theorem and are there only to indicate contrast between cluster boundaries.

Bibliography

Anderson, C. (2008). The End Of Theory: The Data Deluge Makes The Scientific Method Obsolete. *online*.

Baillargeon, R. and Carey, S. (2012). Core cognition and beyond: The acquisition of physical and numerical knowledge. *Early childhood development and later outcome*.

Box, G. E. P. (1976). Science And Statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Descartes, R. and Lafleur, L. J. (1960). *Meditations On First Philosophy*. Bobbs-Merrill New York.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization Of Mdl Keys For Use In Drug Discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

Ellis, D. (2019). Chemical Kinetic Interactions Cover Image. `https://s100.copyright.com/AppDispatchServlet?startPage=i&publisherName=Wiley&publication=kin&contentID=10.1002%2Fkin.21180&endPage=i&title=Cover+Image%2C+Volume+50%2C+Issue+6`. Accessed: 2019-6-6.

Grover, A. and Leskovec, J. (2019). Node2vec: Scalable feature learning for networks. Accessed: 2019-10-21.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). Inchi - The Worldwide Chemical Structure Identifier Standard. *Journal of cheminformatics*, 5(1):7.

Landrum, G., Tosco, P., Kelley, B., sriniker, gedeck, NadineSchneider, Vianello, R., Dalke, A., Cole, B., AlexanderSavelyev, Turk, S., Ric, Swain, M., Vaucher, A., N, D., Wójcikowski, M., Pahl, A., JP, strets123, JLVarjo, O'Boyle, N., Berenger, F., Fuller, P., Jensen, J. H., Sforna, G., DoliathGavid, Cosgrove, D., Nowotka, M., Leswing, K., and van Santen, J. (2019). Rdkit 2019-03-2 (q1 2019) release.

Lynch, H. (2011). *Infant Places, Spaces And Objects: Exploring The Physical In Learning Environments For Infants Under Two*. PhD thesis.

(MDL), M. I. S. (1984). Maccs-ii.

Morozov, A. (2016). Modelling biological evolution: Linking mathematical theories with empirical realities. *Journal of Theoretical Biology*, 405:1 – 4. Advances in Modelling Biological Evolution: Linking Mathematical Theories with Empirical Realities.

Nguyen, K. T., Blum, L. C., van Deursen, R., and Reymond, J.-L. (2009). Classification Of Organic Molecules By Molecular Quantum Numbers. *ChemMedChem*, 4(11):1803–1805.

Noble, C. E. (1957). Human Trial-And-Error Learning. *Psychological reports*, 3(2):377–398.

Oliveira, B., Pereira, F., de AraÃºjo, R., and Ramos, M. (2006). The hydrogen bond strength: New proposals to evaluate the intermolecular interaction using dft calculations and the aim theory. *Chemical Physics Letters*, 427(1):181 – 184.

Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., and Strauss, C. E. M. (2005). Practical Conversion From Torsion Space To Cartesian Space For In Silico Protein Synthesis. *Journal of computational chemistry*, 26(10):1063–1068.

Probst, D. and Reymond, J.-L. (2018). Smilesdrawer: Parsing And Drawing Smiles-Encoded Molecular Structures Using Client-Side Javascript. *Journal of chemical information and modeling*, 58(1):1–7.

rdkit (2019). Rdkit.

Roberts, R. (1989). *Serendipity: Accidental Discoveries In Science*. Wiley Science Editions. Wiley.

Spahn, V., Del Vecchio, G., Labuz, D., Rodriguez-Gaztelumendi, A., Massaly, N., Temp, J., Durmaz, V., Sabri, P., Reidelbach, M., Machelska, H., Weber, M., and Stein, C. (2017). A Nontoxic Pain Killer Designed By Modeling Of Pathological Receptor Conformations. *Science*, 355(6328):966–969.

T. Leube, B., Inglis, K., J. Carrington, E., Sharp, P., Shin, F., R. Neale, A., Manning, T., Pitcher, M., J. Hardwick, L., Dyer, M., Blanc, F., Claridge, J., and J. Rosseinsky, M. (2018). Lithium transport in li 4.4 m 0.4 m Â 0.6 s 4 (m = al 3+ , ga 3+ and m Â= ge 4+ , sn 4+): Combined crystallographic, conductivity, solid state nmr and computational studies. *Chemistry of Materials*, 30.

Wang, S.-G. and Schwarz, W. H. E. (2009). Icon Of Chemistry: The Periodic System Of Chemical Elements In The New Century. *Angewandte Chemie*, 48(19):3404–3415.

Weininger, D. (1988). Smiles, A Chemical Language And Information System. 1. Introduction To Methodology And Encoding Rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Yu-ChenLo (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538 – 1546.