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

Introduction

"In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move"

- Douglas Adams, *The Restaurant at the End of the Universe*

1.1 Background

1.1.1 A Preface on Humanity and the Climate

The development of humanity is not unlike the chirography of an Aristotelian tragedy. It starts with a simple/primitive species cradling a noble cause - to improve their chances of survival. Here the protagonist (humankind) develops a fatal flaw: an insecurity and latent distruction of their home due to a sudden rise to power. Having acknowleged this flaw, we now strive to imporve our understanding of the universe, correct past mistakes and stem the tide of inevitable change.

With tragedy being an imitation not of humanity, but of action and life, happiness and misery, it is only expected that such a comparison to our current affairs should stir feelings of catharsis when exploring our need for research and scientific advancement. It is with that I begin this thesis with the begining of the planet, its atmosphere and consequently the start of humankind.

1.1.2 Formation of the Atmosphere

4.5 billion years ago the Earth began as a disk of dust and gas orbiting our sun. The movement of such gasses produces a resonant drag instability, which causes them to clump together [Hopkins and Squire, 2018; Woo, 2018]. As these 'clumps' become denser, other forces come in to play and further increase their size. These eventually produced the hot mix of gas and solid which was to become Earth. As the Earth cooled, the vollotile componenets of the primodial gas cloud surrounding it begin to form an atmosphere. At this point in time oxygen was not only absent in the atmopshere, but also had many sinks within the Earths anoxidised crust. It was not until oxygenic photosynthesis ([Peretó, 2011]) that the concentrations of oxygen in the atmosphere started to increase. Eventually the development of multicellular cyanobacteria[1] resulted in biologically induced oxygen accumelating in the atmosphere, [University of Zurich, 2013]. This led to the most significant climate event in the planets history: the Great Oxigenation Event (2.5 billion years ago), [Planavsky et al., 2014]. This increase of oxygen allowed oragnisms to become larger and more active, eventually resulting in the human race.

1.1.3 Rise of the Homo Spiens ('Wise Man')

2-6 million years ago there were many varieties of the 'homo' genus (**??**,**?**). 70,000 years ago homo sapiens came into existance and started the cognitive revolution. Here an increase in brain size

[1]The phylum of phtosynthetic prokaryotic (cells not containing a distinct nucleus) bacteria - e.g. blue-green algae

resulted in an increase of communicaiton, tool development and analysis capabilites. However the evolutuionary brain enlargenment required an increase in net energy intake [Navarrete et al., 2011] (the brain makes up for 2-3% of human body mass but consumes 25% of the bodys energy at rest [Harari, 2015]).

This energy imbalance was soon adressed by change of diet [Aiello and Wheeler, 1995], provisioning and sharing (cooperative breeding) and tool assisted processing such as cooking [Wrangham, 2009] - the first known case of anthropogenic indoor air pollution. The increase of cerebral power eventually led to the agricultural revolution[2] (12,000 years ago) and the scientific revolution[3] (500 years ago), [Harari, 2015].

As part of this air pollution and climate have always been a concern for the human race. Concerns about lead in the air can be documented back as far as 6000 years ago [se ref,], in ancient Rome [1145] and in 1285 where after a visit from Queen of England to a coal burning to Nottingham, the first air pollution act was deployed [1147]. air pollution = animals air qulaity pollicy kingxx With the this increased capability, a language capable of communicating information, allowing for the ability to not only hunt larger prey but also. Ability to metaphorical, allowed fruther knwoplege transfer , cvave paings and metaphoirical for people over 150 .... REFERENCES TO OTHER CHAPTERS... - vis - accounting via metaphors - and an interest in science, and atmosphere

Preseent organismas are incresingly shaped by intelegent design rather than natural selection

1.2 Motivation (How the atmosphere affects us)

The atmosphere makes up an integral part of the earth system. It is responsible for shielding the earth from harmful radiation, allowing the transport of energy (weather and climate forcing) and interacting with the biosphere. This section explores the many roles of the atmosphere, and consequently the interests and motivation of climate and atmospheric science. We start with the composition of the atmosphere and air quality (Subsection 1.2.1), and then relate this to the different roles of Ozone (**??**), concluding on changing climate and radiative forcing, for with OH plays a vital role (Subsection 1.2.3).

1.2.1 Air Quality - it is the air we breathe

The atmosphere consists mainly of Nitrogen and Oxygen (forming 99% of its total mass), as well as a vast range of other species [Pryor et al., 2015]. Human beings rely on oxygen to convert sugars and

[2]Domestication of plants and animals.

[3]Humankind admit ignorance and gain unpresidented power

fatty acids into energy. The procurement of this lies through the breathing of the air surrounding us - the composition of which can have dire effects on our respiration system. Pollutants such as particulate matter (PM) to ozone (O_3), nitrogen (NO_2) and sulpher (SO_2) dioxides can cause respiotory problems, heart disease, strokes, cancer and chronic obstructive pulmonary disease Organization [2018]. Over 80% of people who live in urbain environmets[4] are exposed to poor air quality levels exceeding the reccomended limits by World Health Organisation, air quality poses a significant risk to human life - It is estimated that 4.2 million premature deaths globally are linked to ambient air pollution[5] (Figure 1.1).

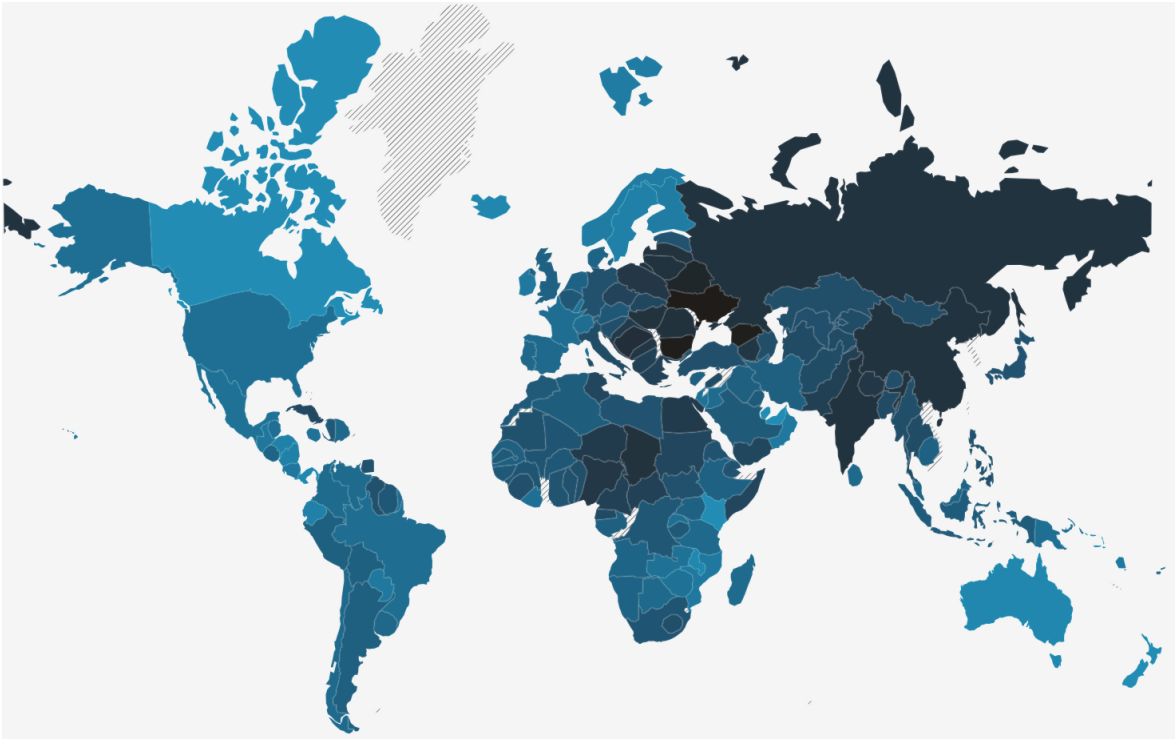


Figure 1.1: **Reported eaths attributed to air pollution by country (2016)**A cartogram cloropleth showing the number of premature deaths attributed to ambient air pollution per 100,000. The colour bar range is from 9 (light blue) to 170 (navy) people. Data Source:[**?**]

1.2.2 Protective barrier - the ozone hole

Ozone plays a vital role in the stratosphere. This was seen in the 1980s where the use of Cloro Floro Carbon (CFC) aerosols resulted in the thinning of the atmospheric ozone [Farman et al., 1985][6]. This resulted in an increase in UV-B radiation, and in consequence skin cancerss, immune supression and disorders of the eye [Bais et al., 2018]. However since their ban in the Montreal Protocol, the atmo-

[4]Which measure the levels of air pollution.

[5]A similar number can also be attributed to indoor air pollution - which also falls under the umbrella term of Air-Quality.

[6]Here the chlorine attacks the double bond and 'steals' an oxygen atom from the O_3 molecule.

spheric hole in ozone has recently recovered to levels similar to its discovery 35 years ago [Ellen Gray, 2019].

1.2.3 Changing Climate

Over the last 30 years a large body of scientists have established that humans have a warming effect on the planet [Houghton et al., 1996b,a; IPCC, 2007, 2013; IPBES, 2019]. Here it has been shown that changes in temperature can lead to the melting of glaciers, rise of sea levels, extreme weather events and the extinction of many species.

[Höhne et al., 2020]

1.3 Tropospheric Chemistry

The lowest part of the atmosphere (<18km)[7] is called the troposphere. This contains 75% of the atmospheres mass, and comes from the greek $\tau\rho o\pi o\varsigma$ which means 'way' or 'turn towards change'. This describes the turbulent mixing that happens due to friction in the lower 2km of the atmosphere (the boundary layer). As the troposphere is the closest part the ground, this where most of the complex chemistry which affects us at the surface happens. This section describes the basic chemical processes which exist in the atmosphere.

1.3.1 Ozone Production/Loss

In the troposphere, the mixing ratio of ozone is controlled by the photostationary state relationship (Equation 1.1-1.3). Since the concentrations of ozone (20-60 pp**b**v)[8] are often much higher than that of the nitrogen oxides, NO (1-60 pp**t**v)[9] and NO_2 (5-70 pp**t**v), the rapid rate of reaction between Equation 1.1 does lead to a net change in O_3 concentration [10] [Jacobson, 2005].

$$NO + O_3 \xrightarrow{k_1} NO_2 + O_2 \tag{1.1}$$

$$NO_2 + hv \xrightarrow{J} NO + O \quad (\lambda < 420nm) \tag{1.2}$$

$$O + O_2 + M \xrightarrow{k_2} O_3 + M \tag{1.3}$$

[7] 18km at the tropics, 17km in the mid latitudes and 6km at the poles.

[8] ppbv: parts per billion volume

[9] pptv: parts per trillion volume

[10] In urban areas NO concentrations may rise to be be greater than those of O_3 during the night. This leads to a decrease in from Equation 1.1

Using Equation 1.1 and Equation 1.2 it is possible to describe the change in NO_2 as:

$$\frac{d[NO_2]}{dt} = k1[NO][O_3] - J[NO_2] \tag{1.4}$$

If the relative change of NO_2 is small, it can be though of as being in steady state. This means that Equation 1.4 can be simplified to producde a relationship between O_3, NO and NO_2 in steady state (Equation 1.5). Here if any two concentrations are known , the third can be calculated.

$$[O_3] = \frac{J[NO_2]}{k1[NO]} \tag{1.5}$$

As ozone is a secondary pollutant (made not emitted), and its main reaction produces a null cycle, the production of ozone in the atmosphere requires an increase in nitrogen dioxide concentraions.

1.3.2 The NOx cycle

Ozone production/loss in the troposphere is directly dependant on the concentration of available Nitrogen Oxides (NOx) (Subsection 1.3.1). These are predominantly emitted by motor verhicles and power stations and can are known to cause respiotory problems in children and asmatics as well as disrupting terrestirial and aquatic ecosystems [EEA, 2018]. Although NOx may be released naturally, the anthropogenic influence on their emissions was hilighed in early 2020 where the COV-19 corona virus disrupted travel across mainland china, causing a significant drop in anthropogenic emissions - Figure 1.2.

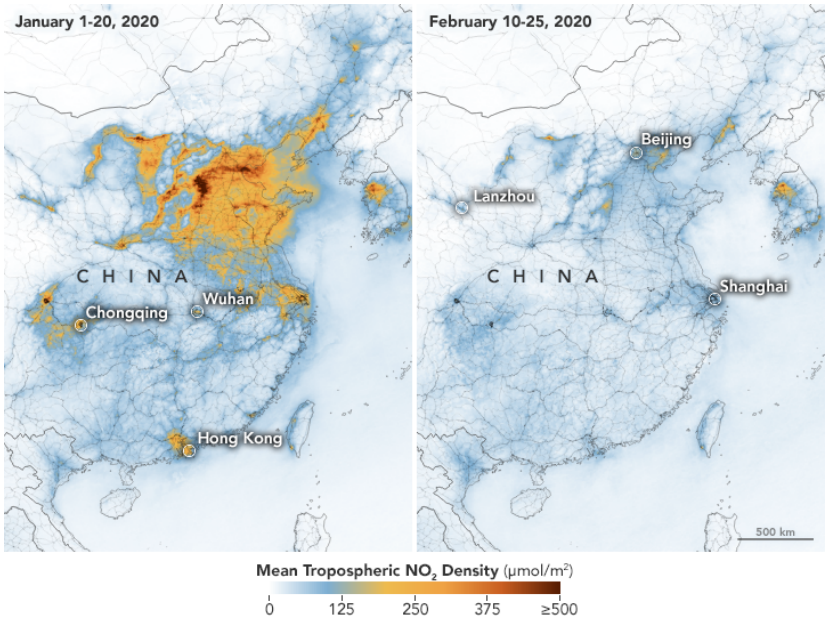


Figure 1.2: **Changes in NOx concentrations due to anthropogenic emissions.** A reduction in activity and trasport resutls in a large decrease of Nitrogen dioxide concentrations in the troposphere. Source: [Stevens, 2020]

During the day nitrate (NO_3) radicals can be formed through the reaction with O_3: Equation 1.4 and Equation 1.6, however this is quickly destroyed through rapid photolysis (Equation 1.7) [Ng et al., 2017]. Photolysis reactions such as Equation 1.7 and Equation 1.2 are no longer possible and th ozone production process shuts down.

$$NO_2 + O_3 \xrightarrow{k_3} NO_3 + O_2 \tag{1.6}$$

$$NO_3 + hv \xrightarrow{J_2} NO_2 + O(3\,P) \tag{1.7}$$

The increased amount of NO_3 can now react with NO_2 to produce dinitric pentoxide (N_2O_5) and an aqueous nitric acid (HNO_3) - Equation 1.8 and Equation 1.9. Equation 1.8 is a three body forwards pressure dependant reaction, and a reverse temperature dependant reaction. During the day at the lower troposphere it is warm and this can occur within seconds, however at night or at high altitudes it can take anywhere from hours to months [Jacobson, 2005].

$$NO_3 + NO_2 \underset{}{\overset{M}{\rightleftharpoons}} N_2O_5 + O \tag{1.8}$$

$$N_2O_5 + H_2O \xrightarrow{k_4} 2\,HNO_3 \tag{1.9}$$

1.3.3 HOx Cycle

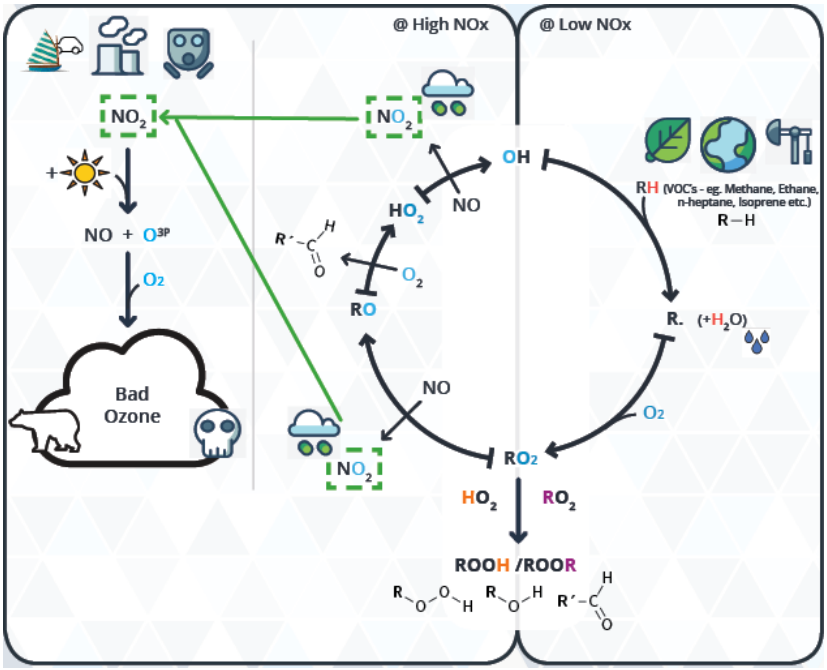


Figure 1.3: **The HOx cycle.**

1.4 Modelling the Earth

In the previous section the air quality and its detrimental effects on human health was seen to influence polcicy for cities and industry. Koyoto, Islands suing powerstations. For a policy to be passed there needs to not only evidence of the problem, but a strong suggestion that any proposed changes will have the desired effect. As it is not possible to perform experiments on complex, and often unknown, chemistry at every location on the planet, we are forced to rely on the numerical simulation of the Earth System, and the constituent parts within it.

1.4.1 Earth System Models (ESM)

ESMs are models capable of predict past or future interactions of the planetary system. They represnt our foremost understanding of the complex interplay between land-surface (geosphere), ocean (hydrosphere), ice (cryosphere) and the air (atmosphere), and act as a surrogate to manual experimentation - which is just not possible on the global scale. ESMs can be split into their individial parts. One example of this is the Chemistry section of the Goddard Earth Observing System (an integrated ESM and data assimulation model hosted by NASAs Goddard space flight centre [**?**]) - GEOS Chem. GEOS-Chem is a global 3D model of atmospheric chemistry which is driven by the meteorology pro-

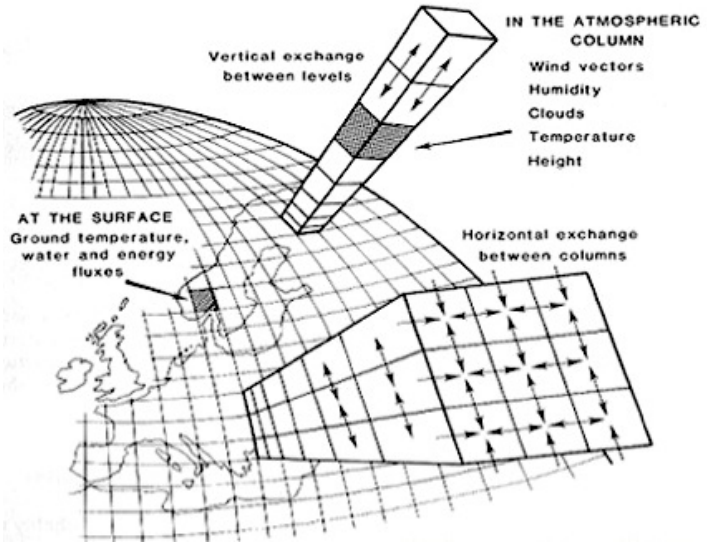


Figure 1.4: **A diagram showing the longitudal, lateral and vertical decomposition of a 3D global model.** Source: [Henderson-Sellers, 2015]

vided by NASA [GEOS-Chem, 2020]. Here the earth is split up into cubic sphere cells longitudally and latitudally, as well as vertically (Figure 1.4)[11]. Each one of these cells performs several purtubations of the chemistry within them, before any long-lived species are transported, and the process is repeated. If extracted separately a single one of these cells may be used to explore the sensetivity of different species for a range of input conditions. This is the bases of the atmospheric box model.

1.4.2 The box model.

In exploring the sensetivities of individual species within a simulation, it is possible to use a zero dimensional box model. This is in essence a single cell within the global structure, constrained in location and height (pressure). mechanism, integrator, etc.

It is then possible to take the many species, their rates or reaction and loss to produce a chemical mechanism detailing their properties in real life.

1.4.2.1 Chemical Mechanisms

Mechanisms are at the heart of every chemistry simulation. They are a mathematical representation of the possible reactions (and the rates at which these may occur) for every s

[11] This image is not from GEOS-Chem.

A note on model type, lifetime and mechanism size

Species such as OH and ... are very short lived and Ozone for example is long lived and can be transported

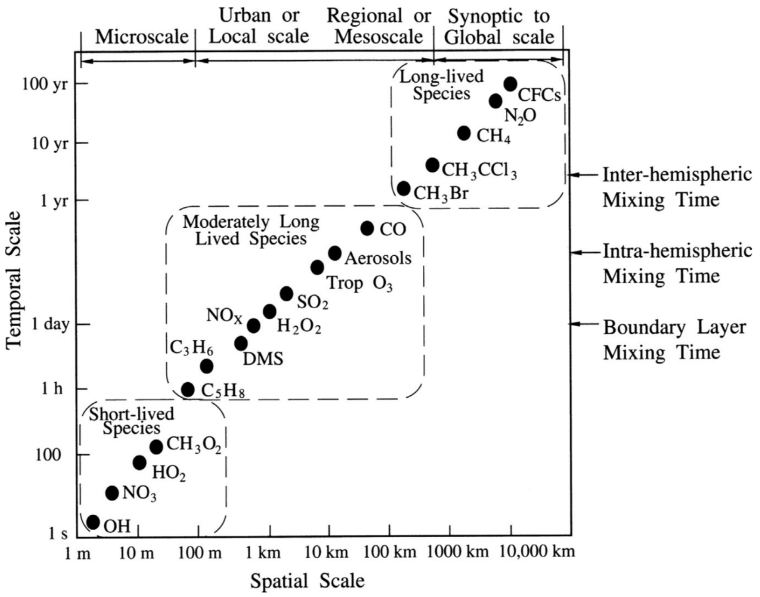


Figure 1.5: **Spatial and temporal scales of variability of atmospheric species.** Source: [Seinfeld and Pandis, 2016]

The atmosphere consists of thousands of species, with tens of thousands of reactions between them. These models represent real wodlc reactions In modelling these we can describe their rate of production and loss with respect to the species they react with.

1.4.3 Numerical integration

For example, it is possible to figure out how quickly each species in a reaction is changing if the reaction mechanism (the exact way it happe ns) and some simple data are known. This representation of how quickly the concentrations are changing is the same as a slope, or derivative. Integration allows us to find the actual change over time and not just how quickly the change is happening. Fo r example, given the following reaction, In a mechanism we are concerned with calculating how quickly a species changes within the chemical system. Taking the reaction of N_2O_5 (Equation 1.10) we can write the rate of change for each species over time (Equation 1.11)[12]. In integrating this equation, we are able to calculate the actual change in concentration (Equation 1.12) - this is the foundation of atmospheric models.

$$N_2O_5 \longrightarrow NO_2 + NO_3 \tag{1.10}$$

[12]This is also known as the flux.

$$\mathrm{d}[N_2O_5]/\mathrm{dt} \longrightarrow \mathrm{d}[NO_2]/\mathrm{dt} + \mathrm{d}[NO_3]/\mathrm{dt} \tag{1.11}$$

$$\int \mathrm{d}[N_2O_5]/\mathrm{dt} \longrightarrow \int \mathrm{d}[NO_2]/\mathrm{dt} + \int \mathrm{d}[NO_3]/\mathrm{dt} \tag{1.12}$$

1.4.3.1 Non-Stiff Equations

Computational systems cannot integrate numbers analytically we rely on a series of computational algorithms. Since integration is the calculation of the area under a curve, the simplest of these

1.4.3.2 Numerically stiff equations (atmospheric chemistry)

Figure 1.5 shows the lifetimes of species can range between x orders of magnitude, similarly the componenets for each reaction (differential equation) evolve on significantly different timescales. This makes the atmospheric chemical mecahnism

1.4.4 The model development cycle

Scientific understanding is the product of many cycles of trial and error, Figure 1.6. In atmospheric chemistry we start with a hypothesis or a question, e.g. will changing X have a negative response on Y. We then construct a theoretical model to represent the chemsistry within. This chemistry is updated to reflect the rates and reactions that have been recorded in laboratory/chamber experiments. This cycle is then repeated until the model and real-world observations produce a comparable result.

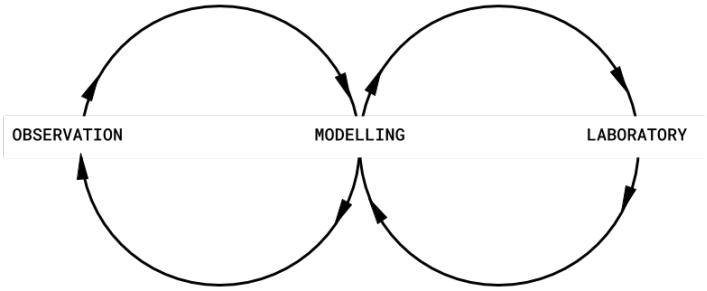


Figure 1.6: **The scientific development cycle.**This shows the iterative nature between modelling, observation and laboratory experimentation

ESM A series of box models.

1.4.5 The Dynamically Simple Model of Atmospheric Chemical Complexity

1.5 Thesis Layout

This thesis will explore a series of methods for describing and understaning the complex chemistry which may exist as part of an atmospheric chemistry mechanism. The mechanism used is a near-explicit representation of our foremost understanding of how gas phase chemistry in the troposphere reacts - the Master Chemical Mechanism, [**?**]. We begin by exploring the use of visualisation to convey complex scientific data (**??**). Next we apply this to the representation of species in a mechanism, and the relationships between them. To do this it is found that the node-link style graph format is the most beneficial, the use of which is then explored further (**??**). However in doing so, large complex network are shown to reach the limits of human cognition and visual representation. To overcome this a series of mathematic metrics are used to leverage our understanding of the species in a chemical network using graph theory (**??**). The use of computation to aid in graph analysis is further extended when graph clustering methods are applied as a method to group similar species within a chemical network (**??**). Finally in a bid towards the use of graph neural networks (see future work, **??**), a range of different chemical representations for machine learning are explored using a number of dimensionality reduction algorithms (**??**).

Bibliography

Aiello, L. C. and Wheeler, P. (1995). The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Current anthropology*, 36(2):199–221.

Bais, A. F., Lucas, R. M., Bornman, J. F., Williamson, C. E., Sulzberger, B., Austin, A. T., Wilson, S. R., Andrady, A. L., Bernhard, G., and McKenzie (2018). Environmental Effects Of Ozone Depletion, Uv Radiation And Interactions With Climate Change: Unep Environmental Effects Assessment Panel, Update 2017. *Photochemical & photobiological sciences: Official journal of the European Photochemistry Association and the European Society for Photobiology*, 17(2):127–179. http://dx.doi.org/10.1039/c7pp90043k.

EEA (2018). Air Quality In Europe - 2018. https://www.eea.europa.eu/publications/air-quality-in-europe-2018.

Ellen Gray, Theo Stein, S. B. (2019). 2019 Ozone Hole Is The Smallest On Record Since Its Discovery. http://www.nasa.gov/feature/goddard/2019/2019-ozone-hole-is-the-smallest-on-record-since-its-discovery.

Farman, J. C., Gardiner, B. G., and Shanklin, J. D. (1985). Large Losses Of Total Ozone In Antarctica Reveal Seasonal Clox/Nox Interaction. *Nature*, 315(6016):207–210. https://doi.org/10.1038/315207a0.

GEOS-Chem (2020). Geos-Chem Publications. http://acmg.seas.harvard.edu/geos/geos_pub.html.

Harari, Y. (2015). *Sapiens: A Brief History Of Humankind*. Harper. https://books.google.co.uk/books?id=FmyBAwAAQBAJ.

Henderson-Sellers (2015). Climate Data Services | Nasa Center For Climate Simulation. https://www.nccs.nasa.gov/services/climate-data-services.

Höhne, N., den Elzen, M., Rogelj, J., Metz, B., Fransen, T., Kuramochi, T., Olhoff, A., Alcamo, J., Winkler, H., Fu, S., Schaeffer, M., Schaeffer, R., Peters, G. P., Maxwell, S., and Dubash, N. K. (2020). Emissions: World Has Four Times The Work Or One-Third Of The Time. *Nature*, 579(7797):25–28. http://dx.doi.org/10.1038/d41586-020-00571-x.

Hopkins, P. F. and Squire, J. (2018). The Resonant Drag Instability (Rdi): Acoustic Modes. *Monthly notices of the Royal Astronomical Society*, 480(2):2813–2838. https://academic.oup.com/mnras/article-pdf/480/2/2813/25498305/sty1982.pdf.

Houghton, J., Filho, L. M., Callander, B., Harris, N., Kattenberg, A., and Maskell, K. (1996a). *Climate Change 1995 The Science Of Climate Change*. The Intergovernmental Panel on Climate Change.

Houghton, J., Jenkins, G., and Ephraums, J. (1996b). *Climate Change 1990 The Science Of Climate Change*. The Intergovernmental Panel on Climate Change.

IPBES (2019). Global Assessment Report On Biodiversity And Ecosystem Services | The Intergovernmental Science-Policy Platform On Biodiversity And Ecosystem Services. https://ipbes.net/global-assessment.

IPCC (2007). *Fourth Assessment Report: Climate Change 2007: The Ar4 Synthesis Report*. Geneva: IPCC. http://www.ipcc.ch/ipccreports/ar4-wg1.htm.

IPCC (2013). *Climate Change 2013: The Physical Science Basis. Contribution Of Working Group I To The Fifth Assessment Report Of The Intergovernmental Panel On Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. www.climatechange2013.org.

Jacobson, M. (2005). *Fundamentals Of Atmospheric Modelling*. Cambridge University Press. https://www.cambridge.org/core/books/fundamentals-of-atmospheric-modeling/A6B866737D682B17EE46F8449F76FB2C.

Navarrete, A., van Schaik, C. P., and Isler, K. (2011). Energetics And The Evolution Of Human Brain Size. *Nature*, 480(7375):91–93. http://dx.doi.org/10.1038/nature10629.

Ng, N. L., Brown, S. S., Archibald, A. T., Atlas, E., Cohen, R. C., Crowley, J. N., Day, D. A., Donahue, N. M., Fry, J. L., Fuchs, H., Griffin, R. J., Guzman, M. I., Herrmann, H., Hodzic, A., Iinuma, Y., Jimenez, J. L., Kiendler-Scharr, A., Lee, B. H., Luecken, D. J., Mao, J., McLaren, R., Mutzel, A., Osthoff, H. D., Ouyang, B., Picquet-Varrault, B., Platt, U., Pye, H. O. T., Rudich, Y., Schwantes, R. H., Shiraiwa, M., Stutz, J., Thornton, J. A., Tilgner, A., Williams, B. J., and Zaveri, R. A. (2017). Nitrate radicals and biogenic volatile organic compounds: Oxidation, mechanisms, and organic aerosol. *Atmospheric Chemistry and Physics*, 17(3):2103–2162. https://www.atmos-chem-phys.net/17/2103/2017/.

Organization, W. H. (2018). Who | Ambient Air Pollution: Health Impacts. https://www.who.int/airpollution/ambient/health-impacts/en/.

Peretó, J. (2011). *Oxygenic Photosynthesis*, pages 1209–1209. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-11274-4_1721.

Planavsky, N. J., Asael, D., Hofmann, A., Reinhard, C. T., Lalonde, S. V., Knudsen, A., Wang, X., Ossa Ossa, F., Pecoits, E., Smith, A. J. B., Beukes, N. J., Bekker, A., Johnson, T. M., Konhauser,

K. O., Lyons, T. W., and Rouxel, O. J. (2014). Evidence For Oxygenic Photosynthesis Half A Billion Years Before The Great Oxidation Event. *Nature geoscience*, 7(4):283–286. https://doi.org/10.1038/ngeo2122.

Pryor, S., Crippa, P., and Sullivan, R. (2015). Atmospheric chemistry. In *Reference Module in Earth Systems and Environmental Sciences*. Elsevier. http://www.sciencedirect.com/science/article/pii/B9780124095489091776.

Seinfeld, J. and Pandis, S. (2016). *Atmospheric Chemistry And Physics: From Air Pollution To Climate Change*. Wiley. https://books.google.co.uk/books?id=n_RmCgAAQBAJ.

Stevens, J. (2020). Airborne Nitrogen Dioxide Plummets Over China. https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china?fbclid=IwAR1z9jXZfY8xNZsCCRRo8Eor2hCjbNDIV70wXGOlzmNyFPkFBesURDCAwB4.

University of Zurich (2013). Great Oxidation Event: More Oxygen Through Multicellularity. *Science Daily*. https://www.sciencedaily.com/releases/2013/01/130117084856.htm.

Woo, M. (2018). Planet Formation? It'S A Drag. *Scientific American*. https://www.scientificamerican.com/article/planet-formation-its-a-drag/.

Wrangham, R. (2009). *Catching Fire: How Cooking Made Us Human*. Basic Books.

Chapter 2

Computational Learning of Species Structure using Visualisation and Vector Clustering

"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*

2.1 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 (hypothesising) 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 experience, 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... [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 apparent 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. New datasets tend to be large, complex and highly multivariate. Although this dramatically improves the quality of science, 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 inferred (visually) from the node-link graph structure of a mechanism. Similarly, **??** and **??** located the presence of important species and clusters of similar 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 2.2). Next, we define the dimensionality reduction methods, which are to be used to simplify the inputs above (**??**). 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.

2.2 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 troposphere. **??** shows that information about a species structure is encoded within its reactions, much of which can be attributed to the well-defined construction protocols.

This section explores the different methods of representing a species structure, intending to provide a machine built algorithm with the highest amount of information about each species and its functionality. A range of input types will be evaluated against several dimensionality reduction algorithms to isolate which chemical properties are most 'picked up'.

2.2.1 Input Generation

The MCM provides species information in the form of a species 'smiles' (Subsubsection 2.2.3.2) 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 chemistry; these are outlined below.

2.2.2 Manual Categorisation

Reactions within the MCM are determined by a set of rules (PROTOCOL SECTION). These 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 successfully separated species structure, it makes sense to run a unit test using the known functional groups of a species as the input.

To generate the functional groups the regular expressions in Table 2.1 are used[1] on the smiles strings (described in Subsubsection 2.2.3.2) for each species. In extracting the functional groups, we can plot the likeliness a species with a certain group is likely to have another using a chord diagram - Figure 2.1. Since most species 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 2.1: 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 Section C.2.

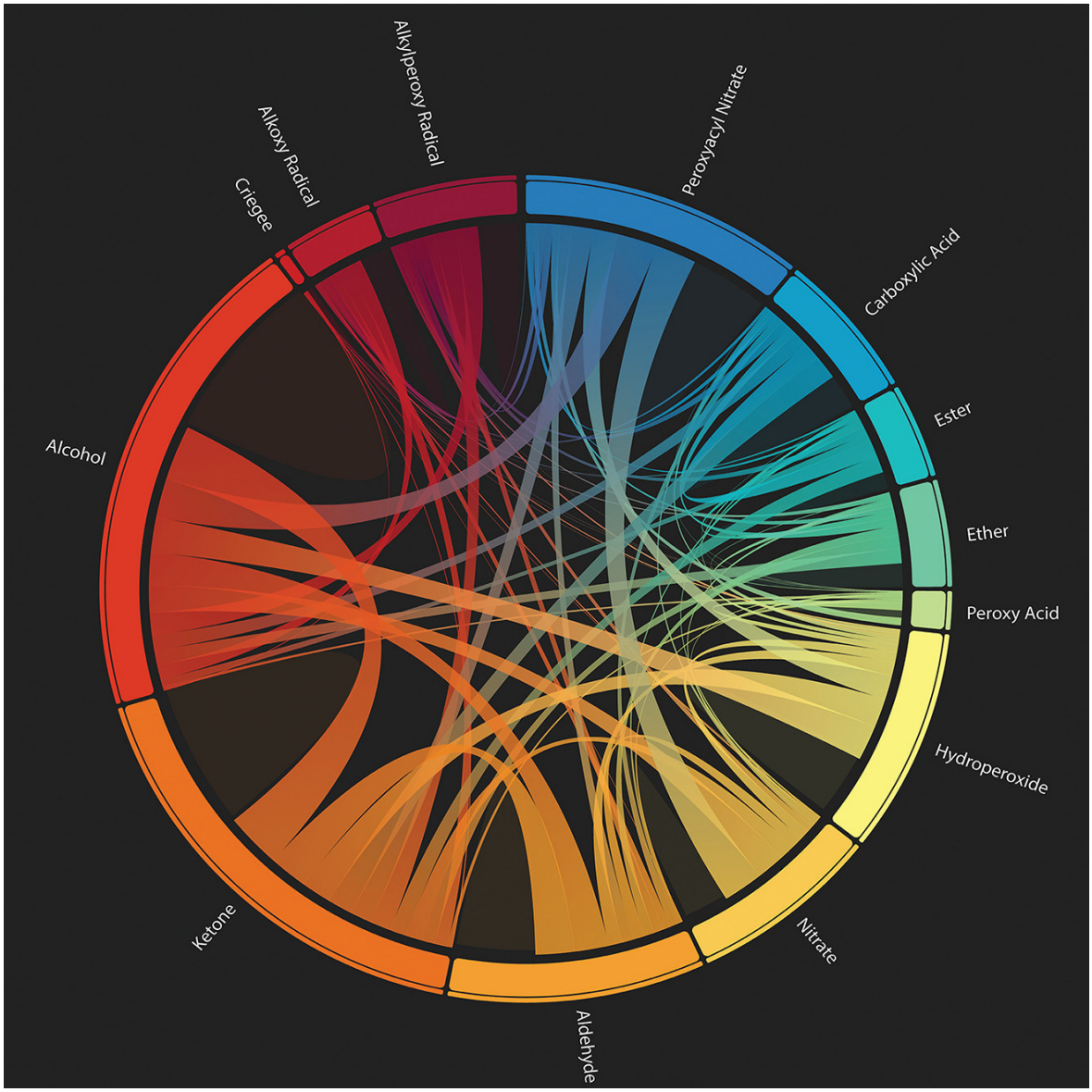


Figure 2.1: **The multifunctionality of the MCM.** A chord diagram showing the functionalisation of a species within the MCM. Arc sizes represent what percentage of all functional groups in the MCM mechanism a group contains. Translucent areas of no outwards links represent species with multiples of a certain functional group, of which Alcohols and Ketones have the most. Source: [Ellis, 2019]

2.2.3 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.

2.2.3.1 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 adage also applies for the MCM, where an intuitive naming convention is used. This is often derived as part of the construction protocol, where a species names reflect its own, or its precursor's structure (which it will have at least in-part inherited).

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

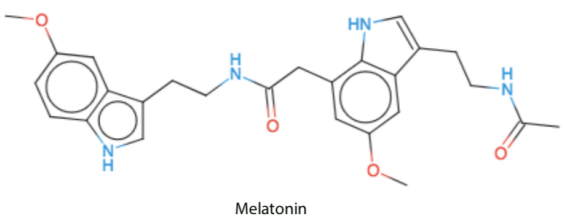
2.2.3.2 Smiles Strings

Smiles ('Simplified Molecular-Input Line-Entry System') provide a human-readable representation of the molecular structure, [Weininger, 1988]. They offer a linear human-readable description of the chemical composition within a molecule - making it easy to visually check the construction of a species without any additional work. Besides, their role in generating the molecular fingerprints in Subsection 2.2.5 makes it a useful comparison to make when evaluating methods of structure representation.

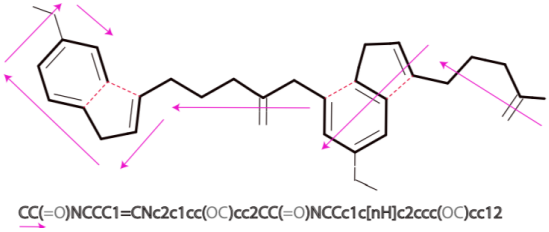
Construction Methodology of SMILES strings

The construction of a SMILES string happens in three parts:

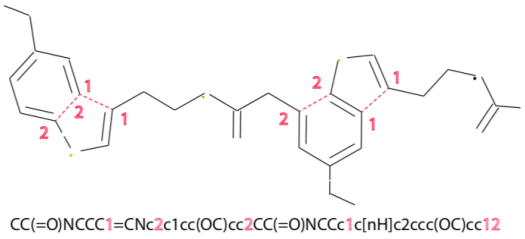
1. The smiles string is built by creating the longest possible chain to form a molecule backbone. Figure 2.2b
2. This may within itself contain aromatic rings denoted by the lowercase carbons and a number corresponding to the location of each break cycle. Figure 2.2c
3. Finally all the functional groups and branches attached to the main backbone are added. These are nested within the parenthesis to show that they are not part of the skeletal backbone. Figure 2.2d



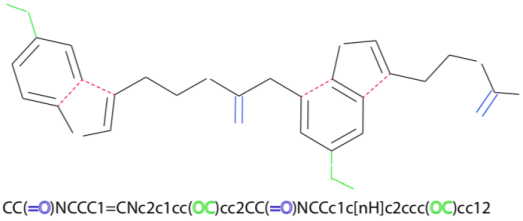
(a) Structure of Melatonin



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



(c) Step 2 : Aromatic Rings



(d) Step 3 : Functional Groups

Figure 2.2: **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 2.2a is made using smiles drawer: [Probst and Reymond, 2018]

2.2.4 Graph Inspired

?? - **??** have shown the role of graphs in revealing network properties and structure. Graphs in themselves can simplify relational data into two/three dimensions for visualisation and algorithmic clustering. Continuing this trend, we can represent a species structure in the form of a graph (Subsubsection 2.2.4.1), as well as converting the structure of a mechanism for dimensionality reduction (Subsubsection 2.2.4.2)

2.2.4.1 The Species Graph (Fingerprint)

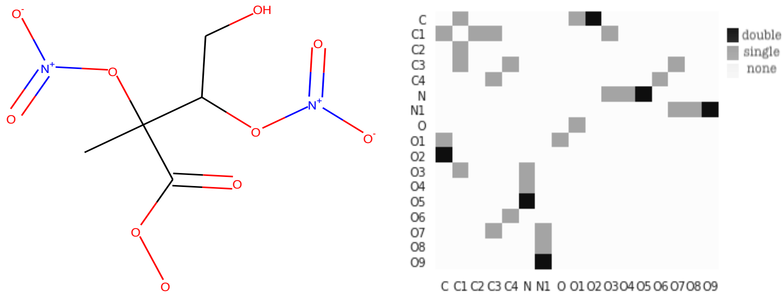
The structure of a species has long represented using a graph-like layout, **??**. It, therefore, follows that other methods for representing the graph structure would also apply. One such way is the use of an adjacency (or relational) matrix to describe the relationships between atoms and bonds in a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

The construction of a structure matrix/graph begins with a chemical species. Here the relationships between atoms (Figure 2.3b) is converted into an adjacency matrix (Figure 2.3c). However, since species have different numbers of each atom, a template allowing us to compare different graphs is required. To do this a maximum occurrence table (Figure 2.3a) is created. Here, for example, BCARY $C_{15}H_{24}$, a sesquiterpene 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 one-dimensional array (vector) of 259 elements (518 bytes) to represent our species.

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

(a)



(b) (c)

Figure 2.3: **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 occurrence 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.

2.2.4.2 Node Embeddings (Node2Vec)

?? and **??** showed that the underlying structure of a chemistry mechanism graph contains information about the species and reactions within it. In Figure 2.4 colour represents the ratio of potential oxidation of a species. Here as emitted species become progressively more processed, the number of bonds which may be oxidised diminishes (lighter colours near the centre) until they eventually form carbon dioxide and water.

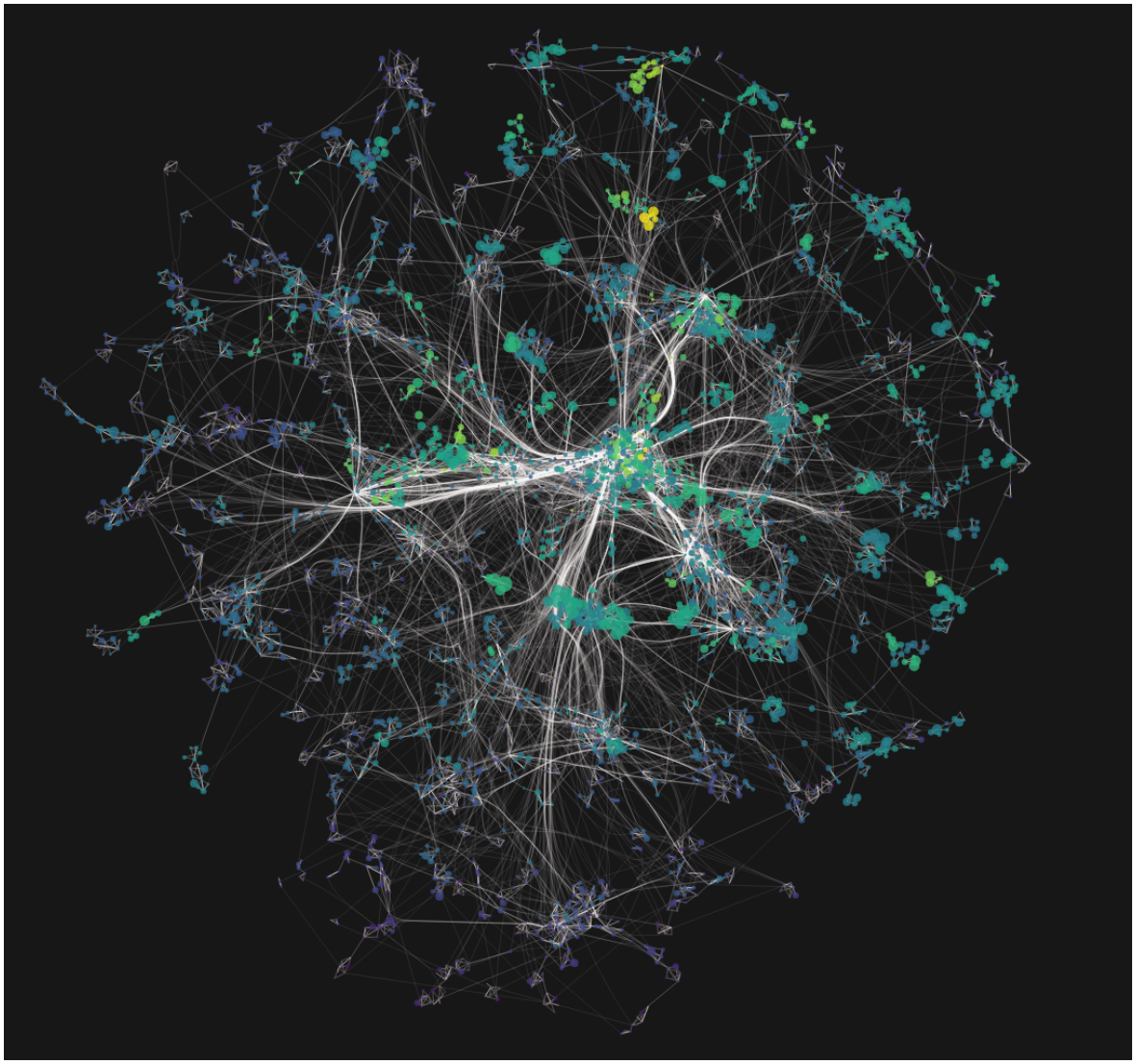


Figure 2.4: **The graph of an MCM subset representing the chemistry within Beijing.** Here colours show the increase of O−C ratio as species are oxidised (lighter). All emitted species ultimately tend towards carbon monoxide which is at the centre of the graph.

This type of structural information can be extracted through the use of a natural language processing package capable of transforming a graph into a vector - node2vec [Grover and Leskovec, 2019]. Since this may also be used for dimensionality reduction, it is described within the next section (Subsection 2.3.6).

2.2.5 Molecular Fingerprints

In the field of chemical informatics, molecular fingerprints (or structural keys) are used to encode and query structural properties of species. Their binary representation makes them suitable for dimensionality reduction and the exploration of fo chemical space (a type of property space constructed using pre-determined features and boundary conditions).

Here species properties are often split into structural and psyico-chemical groups - which has used such as the discovery of natural analogues (which circumvent problems such as intolerances in medicine [Spahn et al., 2017]). Although there exist many different types of molecular fingerprints, the two main ones that will be explored are molecular quantum numbers (MQN) and the molecular access system (MACCS).

2.2.5.1 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: the n principle quantum number, I angular momentum quantum number, M_i magnetic quantum number and M_s spin quantum number. 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, where increasing atomic numbers follow the principal quantum number [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].

2.2.5.2 Molecular Access System (Maccs)

MACCS keys are a 164^2 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.

2.3 Dimensionality Reduction Methods

In the last section, we described several methods in which the chemical structure of a species could be encoded for direct comparison. However, since each input consists of a multitude of elements, it is still not a simple task to determine the differences and similarity between all species in mechanisms. Dimensionality reduction is the process of reducing the number of random variables and only presented a set of principal values, by mapping a high-dimensional space into a low-dimensional one [Roweis

[2]They are 166-bit keys, although 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].

and Saul, 2000]. This allows us to flatten a multivariate input into the two dimensions required for a simple scatter plot.

In this section, we begin by explaining the data preparation required for dimensionality reduction (**??**) before describing the different possible methods of reducing the dimensions of a dataset.

2.3.1 Preperation Of The Data

Real-world data is rarely preformatted in such a way that it can be used directly within a computational model. Often values need to be cleaned and corrected to be fit for purpose. In the interest of completeness, the two main methods of data adjustment for machine learning are outlined below. These are normalisation and standardisation.

Normalisation

In the data is without (dimensionless) or of a single unit, it is possible to rescale the data between a range - most commonly 0,1. In doing so it is possible to interpret the importance of value in contrast to the largest recorded value. This gives us a percentage scale spanning the range of the data. Such a range is useful in the definition of colourmaps and describing the importance of value relative to the dataset. To rescale a dataset we shift the minimum value to zero, then divide by the new maximum of the dataset (Note this is equivalent to the range of the unshifted dataset.)

$$n(x_i) = \frac{x_i - \min_x}{\max_x - \min_x} \tag{2.1}$$

Standardisation

If the components we wish to compare are of different units or are expressed with a different scale, normalising them would not produce meaningful data. Instead, it is possible to standardise the data by looking at each points deviation from the mean. Here the variation of the mean for a dataset is divided by the standard deviation to produce a value between {-1,1}, Equation 2.2. In statistics this is known as the 'z-score'[3]

$$z(x_i) = \frac{x_i - \mu_x}{S} \tag{2.2}$$

[3]Possibly because of the American spelling of standardi**Z**ation?

2.3.2 Principle Component Analysis

One of the most well-known dimensionality reduction methods is the determination of the principal components through the use of Principal Component Analysis (PCA). PCA increases the readability of a dataset by creating a set of new uncorrelated variables which maximise the variance [Jolliffe and Cadima, 2016].

PCA works on the assumption that components within a dataset are linear combinations of each other. By simplifying these linear combinations, it is possible to identify the elements which explain the most variability in a dataset - these are the principal components.

A more straightforward interpretation of this would be to adjust the direction of each axis of the data, such that its projection has the most prominent variability. In doing so, it is possible to determine which components contribute the most to changes in the dataset [F.R.S., 1901; Hotelling, 1933]. An example of this is seen in Figure 2.5, where the second component of the original data can be removed with little effect on the overall result of the data. Such methods have applications in compression and signal filtering [Hernandez and Mendez, 2018; Hamadache and Lee, 2017].

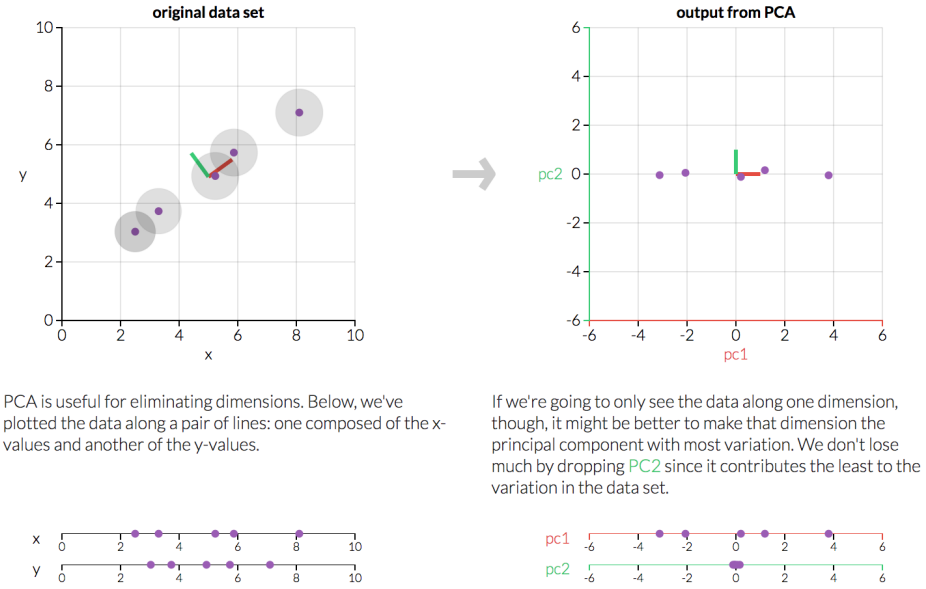


Figure 2.5: **Determining the Principal Compnent of a sample dataset.** It can be seen that in a change in axis to follow the first principal component (right), it is possible to explain most of the variation in the samle dataset (left). Source: [Powell, 2020]

2.3.2.1 Mathematical Explanation Of Pca

Note: *The basic statistics/mathematics required to understand this section is shown in* **??**. *Please read this if you are not familiar with any of the terms below.*

The mathematics behind PCA consists of first calculating the covariance matrix - an $n \times n$ matrix

outlining how strongly each variable changes with every other. Using this we can calculate both the eigenvalues and eigenvectors of the matrix [4]. This can be done using a computational package such as numpy or scipy [Oliphant, 2006; Jones et al., 01].

We can now sort the eigenvector columns by influence using their eigenvalues—this way a feature dataset can be produced by removing vectors of low importance. The final feature dataset can now be transposed and multiplied by the transpose of the original dataset. This results in an output dataset containing each principal component of the desired dimension.

2.3.3 T-Distributed Stochastic Neighbor Embedding (T-Sne)

t-SNE is an algorithm designed with visualisation in mind [Maaten and Hinton, 2008]. Rather than representing the data through a series of linear transformations, t-SNE uses local relationships to create a low-dimensional mapping, much in the same way as a fully connected force graph, Figure 2.6. This allows the ability to capture non-linear structures in the data which cannot be accomplished through linear mapping methods (e.g. PCA).

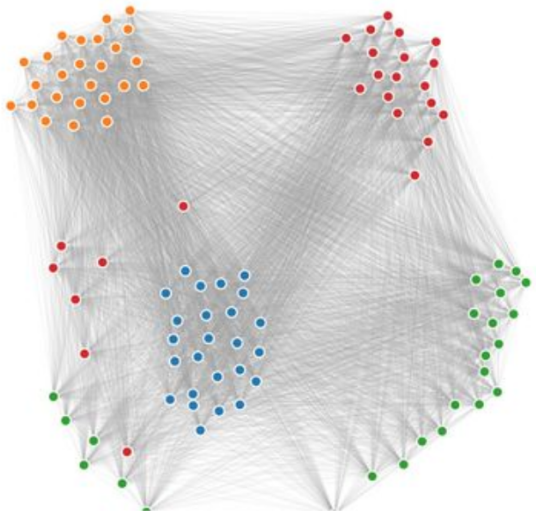


Figure 2.6: **Representing the t-SNE algorithm as a fully connected force graph.** Here each node is attached to every other node. Nodes with a strong relationship are pulled closer together than those with a weaker one.

The algorithm itself can be simplified into two parts,

1. Create a probability distribution which dictates relationships between neighbouring points
2. Recreate a lower-dimensional space following the probability distribution established in 1.

[4]These need to be unit vectors, although most packages already do this out of the box.

and is described in Subsubsection 2.3.3.1.The main reason t-SNE produces good results is that it can handle the **'crowding problem'** very well. The crowding problem is a product of the 'curse of dimensionality'. In a high dimensional space, the surface of a sphere will grow much quicker than one in a lower dimension space. This means that the higher dimension spaces will have more points at a medium distance from a certain point, Figure 2.7. When we map our data into a lower dimension, data will try to gather at its medium distance, resulting in a more 'squashed', and thus crowded, output.

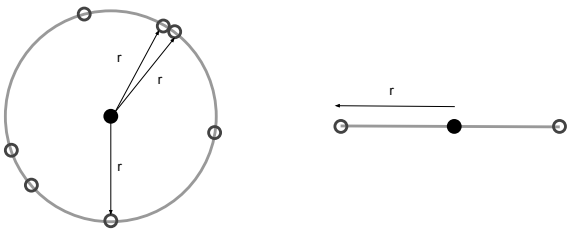


Figure 2.7: **An example of how the curse of dimensionality affects the mapping of points a certian distance from eachother.**

2.3.3.1 Mathematical Explanation Of T-Sne

In the original paper [Maaten and Hinton, 2008], the algorithm is described using the etymologic dissection of its name.

Step 1

First we begin with Stochastic Neigbour Embedding (SNE) - the distribution across neignbouring datapoints in our high dimension space. This is done by converting the high dimensional Euclidian distances between points into conditional probabilities representing their similarity:

$$p_{ij} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2)}{\sum_{k \neq l} \exp(-\left\|x_k - x_l\right\|^2 / 2\sigma_i^2)} \tag{2.3}$$

Here $p_{i|j}$ is the conditional probability that x_i may pick x_j as a neigbour. This is proportional to the probability density of a Gaussian σ_i centered at x_i.

Perplexity Since we want the number of neighbours of each point to be similar in number and prevent a single point from having a disproportionate influence on the entire system we introduce a hyperparameter named *perplexity*. Perplexity works by ensuring that σ_i is small for points in densely populated areas and large for spare ones and can be thought of as a scale of the number of neighbours considered for any one point in the system. Generally, values between 5 and 50 are considered to give

good results, with larger perplexities taking global features into account, and by consequence smaller ones, local features.

Step 2

Now a probability distribution describing the relationship between points has been formulated, we wish to express this as a low dimensional mapping of our inputs X in terms of our output dimensions Y. Naturally, we would want to make the low dimensional mapping represent a similar (Gaussian) distribution as in Step 1. However, it often causes issues presented by the 'overcrowding problem', Subsection 2.3.3, as the gaussian has a 'short tail', and thus nearby points are likely to be pushed together. A solution to this is the student t-distribution which has a longer tail [5]:

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{2.4}$$

Note: *The definition and explination of the Student t-distribution is given in* **??**.

The optimisation of this equation is achieved through the use of *gradient decent*[6] on the Kullback-Leibler divergence Subsection A.2.2 between distributions p and q. Here the gradient is used to apply an attractive and repulsive force on the items[7].

2.3.4 Pca Vs T-Sne, A Quick Comparison.

PCA has been around for much longer than t-SNE, and its uses are well established within the scientific community - an example of this would be the use of sensitivity analysis within mechanism reduction [Turanyi and Tomlin, 2015]. It is fast, simple and easy to use and very intuitive. The PCA algorithm works by creating a lower-dimensional embedding which best preserves the overall variance of the dataset. Clusters created from the algorithm are grouped in ways, such that they retain the highest variance of the data.

The main drawback of PCA is that it is a linear projection. If our data happened to be in a 'swiss roll' (spiral) pattern, we would not be able to 'unroll' it. The reason for this is that the PCA algorithm works by viewing the data from different perspectives, much like casting a shadow from various directions. With such an example, there is no one way we can do this that unfurls the spiral.

[5] The distribution employed is a t-distribution with only one degree of freedom and is identical to the Cauchy distribution

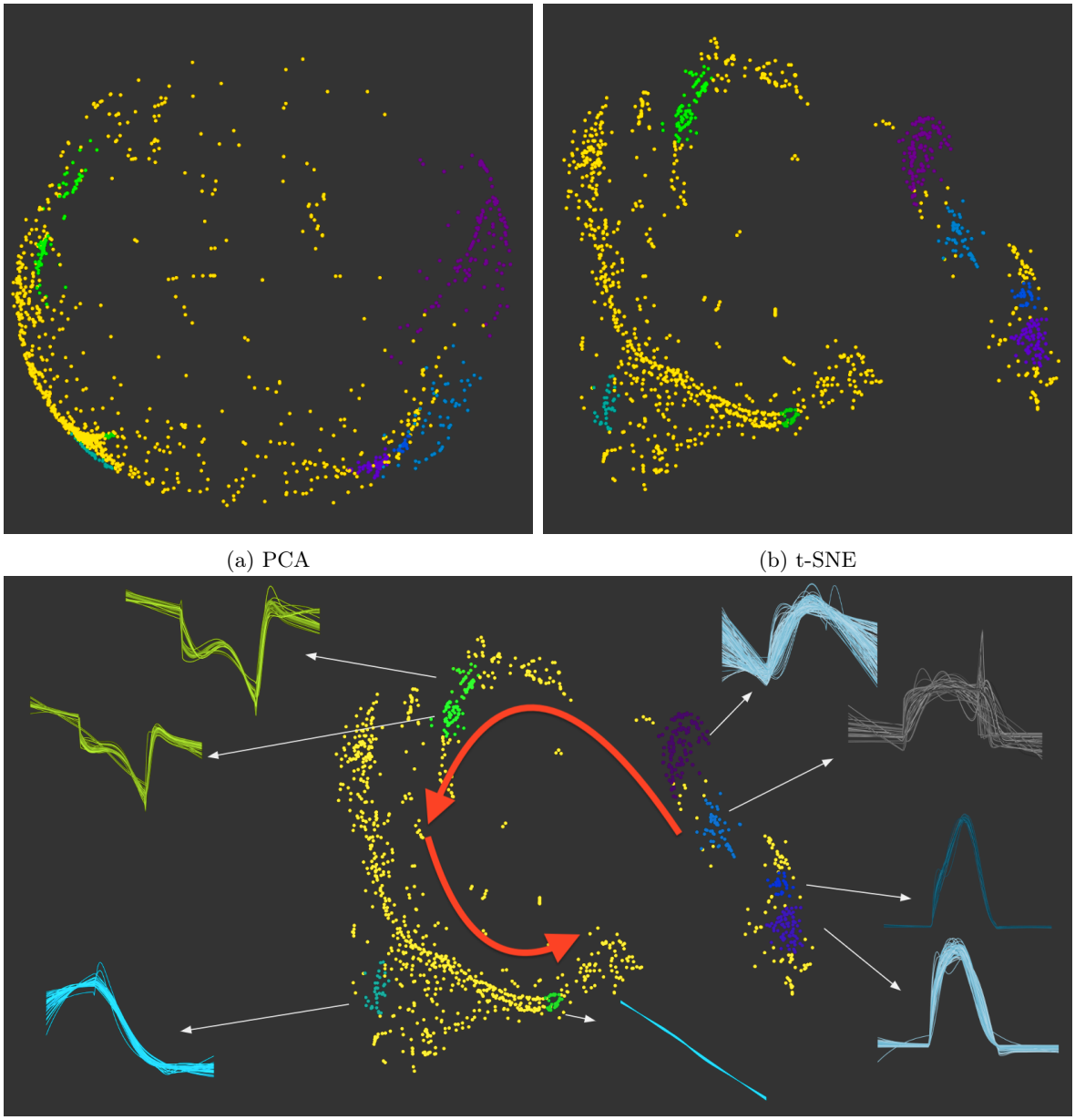
[6] **Gradient Decent** - an optimisation algorithm used to minimise a function by iteratively moving in the direction of the steepest descent. Gradient descent is used to find local minima and is defined by the negative of the gradient of the system. Its primary usage in machine learning is the updating of parameters (coefficients in linear regression and weight in neural networks).

[7] A positive gradient signifies attraction, while a negative one corresponds to repulsion.

t-SNE, on the other hand, is a relatively new method [Maaten and Hinton, 2008]. Its greatest asset is that linear projections do not limit it. Although more computationally intensive for large datasets, t-SNE produces visibly cleaner results. Unlike in PCA, t-SNE cannot be trained on additional data at a later point; however, the output clusters are more visually distinct (they have less of an overlap). Much like in a force graph, the output from t-SNE is scale-invariant. This means that while the location of clusters in a PCA reduced representation has an attributable quality, those produced by t-SNE will not necessarily contain the same information.

A box model run representative of the chemistry within Beijing was used to compare the differences between PCA and t-SNE. The aim is to classify the diurnal profiles of each species concentration (much like the cosine similarity in **??**). Diurnal profiles were extracted on the third day of a spun up model initialised with initial conditions representative of the chemistry within the Beijing environment (**??**). These were then standardised and converted into temporal vectors for use in the algorithms.

Figure 2.8 shows the output of both dimensionality reduction algorithms on the dataset. Different colours represent the location of clusters of similar diurnal profiles. A higher dispersion between clusters and species overlap is seen within the PCA output, Figure 2.8a. This makes it harder to distinguish species from each other or other groups around them. Since the distance between clusters within t-SNE does not hold the same mathematical meaning as PCA, the algorithm can provide a better distribution of points, creating better-defined clusters, Figure 2.8b. The concentration profile shapes for each coloured group is shown in Figure 2.8c.



(a) PCA

(b) t-SNE

(c) t-SNE with cluster outlines.

Figure 2.8: **Showing the difference between PCA and t-SNE clustering.** These figures show the clustering of a set of standardized concentration profiles (c) across two styles of dimensionality reduction: PCA (a) and t-SNE (b).

2.3.5 The Auto-Encoder (Ae)

Auto-encoders are a subclass of neural networks with primary use in compressing data (dimensionality reduction). Rather than predicting a numerical output, AutoEncoders focus on the construction and deconstruction of data through the use of an encoder and decoder pair. The encoder takes an n-dimensional input and applies a compression, reducing it to the number of dimensions in the bottleneck layer. The reduced dataset is then reconstructed within the decoder. Such a process not only allows for an easy understanding of the error of the reduced data but can also be used in the filtration of

noisy or pixelated data [Leite et al., 2018; Dataman, 2019] and as an input to more complex machine learning models.

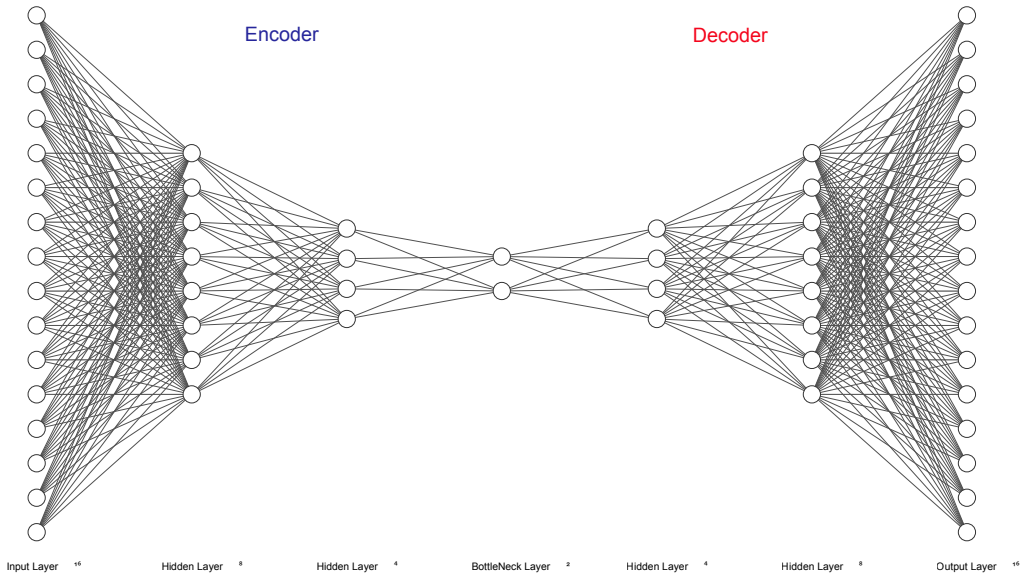


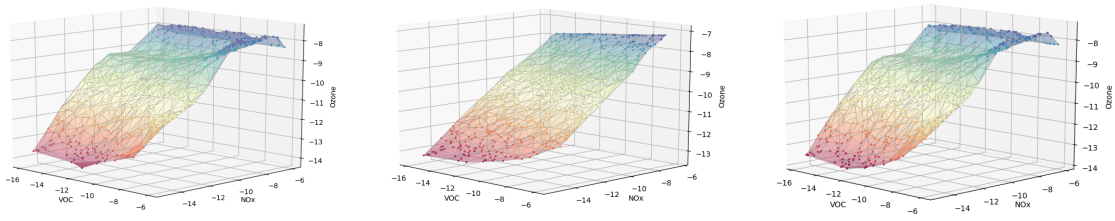
Figure 2.9: An example autoencoder structure which reduces a 16 dimentional input to 2. Draw with the aid of [Krizhevsky et al., 2012]

There are two features of an autoencoder that make it powerful. The first is the ability to sample your latent space using the decoder. The implications of this are that we can establish features that correspond to gaps between our data points - which can have its application if the data used is sparse or incomplete. Next comes the inherent non-linearity of the model. As an autoencoder is just a neural network, the amount of information passed through each link between layers is governed by an activation function. Should this activation function be linear, the reduced dimension will be much akin to a PCA decomposition. Where PCA reduces the dimensions of a dataset by discarding those with a little effect on the variance, an autoencoder opts to combine it- here the entirety of the dataset remains encoded within the links of the AE network. To decide how data flows along the edges of the network, a series of threshold (activation) functions are used for each layer. These are described in B.

2.3.5.1 Demonstration Of Non-Linear Activation Functions

To demonstrate the effect of these we take a sample isopleth of Methane and Ozone, reduce it to two dimensions. This is then reconstructed back into three dimensions using the DR algorithms. Figure 2.10 shows the difference between the original dataset (Figure **??**) and that of the PCA (Figure **??**) and AutoEncoder (Figure **??**) reconstructions. Here we see a loss in the non-linearity of the original data for the PCA reconstruction. However, the use of a non-linear (tanh) activation function

within AutoEncoder produces a result much closer to the original. Use of a linear activation function, however, produces a similar result to the PCA algorithm.



(a) Original (b) PCA (c) AE (Tanh)

Figure 2.10: **Comparing the result of the 2D encoding and decoding of an Ozone-NOx-Methane isopleth.** The original data (a) is reduced to two dimensions and then reconstructed back into 3D. This is done with Principal Component Analysis (b) and an AutoEncoder (c). The original isopleth is created using 300 simulations of different intial conditions: NOx (variable), Methane (variable) and Ozone (constant). These were designed using a latin hypercube and converted into a surface plot using Delaunay triangulation.

2.3.6 Node2Vec

Finally, Node2Vec is an embedding algorithm designed to generate vector representations of the nodes in a *undirected* and *unweighted* network. Although it can be used to reduce a complex network into a 2D vector (dimensionality reduction), for this experiment we shall only use it to generate a fingerprint for a species' position within a mechanism network graph - and then apply this as an input to the DR methods above. This method of input creation has been found more computationally efficient, by circumventing the need for expensive composition, in producing better predictions on network-related tasks compared to more classical methods such as PCA [Grover and Leskovec, 2019].

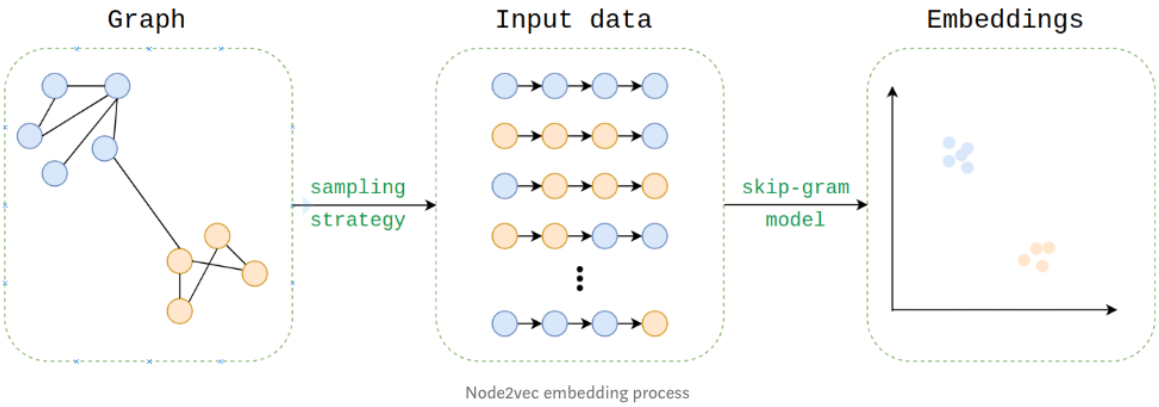


Figure 2.11: The process of converting a graph into a vector using Node2Vec. Source:[Cohen, 2018]

The process of converting the graph structure (Figure 2.11) into a numerical vector node embedding starts by taking a series of 2^{nd} order random walks. These describe the neighbourhood of a node in the

form of a set of random walk paths, much in the same way words are dependant on their neighbours within a sentence: Equation 2.5.

$$ISOPRENE \rightarrow OH \rightarrow TISOPA \rightarrow ISOPBO_2 \rightarrow TISOPA \rightarrow ... \tag{2.5}$$

This methodology allowed for the use of word2vec algorithm, converting the walk into a vector (Subsubsection 2.3.6.2)

2.3.6.1 Sentence Construction By Sampling Of A Network

The probability and path depend both on a set of arguments and a random seed provided to the model. The return and input parameters (p & q) determine how fast we explore the network and our probability to leave the neighbourhood, Figure 2.12. In a system, where the previous path is from t to v, we may calculate the probability of returning to t as $1/p$, going to a mutual node connected between t and v as 1, and viewing a new node as $1/q$. If $q > 1$ we have a high probability to end up at nodes close to t, and with $q < 1$ we are likely to explore other nodes. Additionally if we chose $p > \max q, 1$ we are less likely to return to an already visited node ($p < \min q, 1$ is likely to generate a backwards step). Since we wish to generate a 'local' view, but do not wish to return to t we select $q \geq 1$ and $p > q$ our parameters as $p = 2.0, q = 1.1$. In the case of a weighted graph (something that we are *not* exploring within this chapter) the resultant *alpha* value calculated is further multiplied by the edge weight.

To run the simulation, we use the python2 code provided by the original paper [Grover and Leskovec, 2019] with a set of 50000 random walks, each of length 9. The reasoning behind this is that we have a large graph, with a power-law like structure (where species are often heavily connected, **??**).

NOTE: This process takes over a week to compute (in serial), and then the binary file containing all walks in character form approaches 10 GB, for the complete MCM.

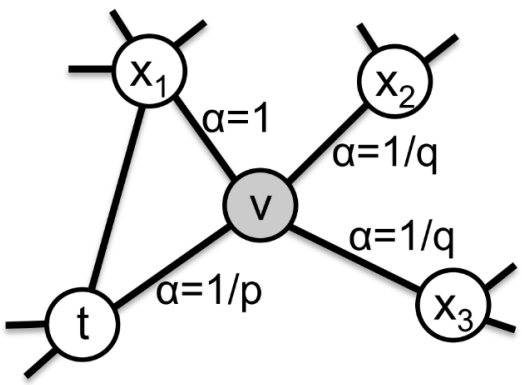


Figure 2.12: Calculation of the random walk path. Source:[Grover and Leskovec, 2019]

2.3.6.2 Word2Vec

Once we have constructed our random path 'sentences' (e.g. Equation 2.5), we can make use of Googles word2vec algorithm [Mikolov et al., 2013]. This is similar to an auto-encoder in many regards; however, the algorithm looks at neighbouring words (or species) in the corpus rather than learning word embeddings using reconstruction. This form of representation has found many uses beyond the realm of natural language processing. Some of these are objects, people, code, tiles,genes and graphs [Lynch, 2011; People2Vec, 2019; Alon et al., 2019; Jean et al., 2018; Du et al., 2019; **?**].

2.3.7 Summary

There exist several methods of reducing a complex dataset into a smaller one. PCA is the simplest method to understand but is constrained to linear decompositions. AutoEncoders can have both a linear and non-linear response, based on the activation functions that they use, and t-SNE applies a non-linear grouping which mimics a complete force-directed graph.

Having defined each method, we next explain how they will be evaluated (Subsection 2.6.1), before applying them to the MCM in Equation 2.5.

2.4 Visualisation Of Clustering

In assessing the validity of clustered space, we require a level of exploratory data analysis. To reveal features of interest, we plot the reduced 2D dataset and apply interactivity coupled with a selection of visualisation techniques described below. This section outlines the different visualisation methods which are used.

2.4.1 Viewing The 2D Species Embeddings

Since the different DR algorithms return data on various scales, comparison between the outputs is not straightforward. To overcome this outputs in x and y are normalised (scaled between $\{0,1\}$), before being plotted as a scatterplot.

2.4.2 Exposing Overlapping Data

If the nodes within a tight-knit cluster overlap, this can cause obfuscate the results and limit the user's ability to select them. As an initial test, node sizes can be reduced. However, this may often result in points too small to pick. The other solution which was used is to create a force-directed graph where

each point is strongly attracted to their initial position. Here we can apply collision detection, while still preserving the overall grouping of nodes within a cluster - a technique that was seen in **??**.

2.4.3 Gooey Effect (Gaussian Blur)

Taking a quote from Reinhardt [1975]: *"The more stuff in it, the busier the work of art, the worse it is. More is less. Less is more. "* and combining it with the work from **??**, we realise that showing each species, when observing overall clusters just add unnecessary clutter to the images. Instead, since we are only interested in the clusters as a unit, a 'gooey effect' filter can be applied. This works by merging nearby points into a single water-like blob using a gaussian blur[8]. Here since each point is allocated a colour, if a colour gradient exists, then there are multiple clusters occupying the same place. The aim of this is to reduce the cognitive load on the end-user by reducing the number of distinct objects that they need to take in.

2.4.4 Four Colours Theorem

When plotted, the number of clusters detected often exceeds the number of categorical colours available. In cartography, it has been noted that the colouring of neighbouring polygons should at most take four colours. This is the origin of the four colours theorem Appel and Haken [1976], of which a greedy implementation is applied.

The aim of this is to show item boundaries (for instance countries, or in our case clusters) while reducing ambiguity (if, say, two neighbours have the same colour). The algorithm I adapted uses the Delaunay tesselation scripts contained within DataDrivenDocuments.js (d3js) Bostock [2012]. This partitions our plane into polygon-regions, each of which includes boundaries at the furthest distance from each point (Voronoi cells) Watson [1981]. First, we chose a random cell and assign it a colour. Next, all its neighbours are recursively iterated, giving them the lowest possible colour in a list, which does not match any of their neighbours. Although such a greedy approach does not produce an optimum result, it allows for the colouring of data with ≤ 5 distinct colours, as is shown in Figure 2.13.

[8]Here a gaussian blur of standard deviation 3.7 and a colour matrix [1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 37 -5] is used.

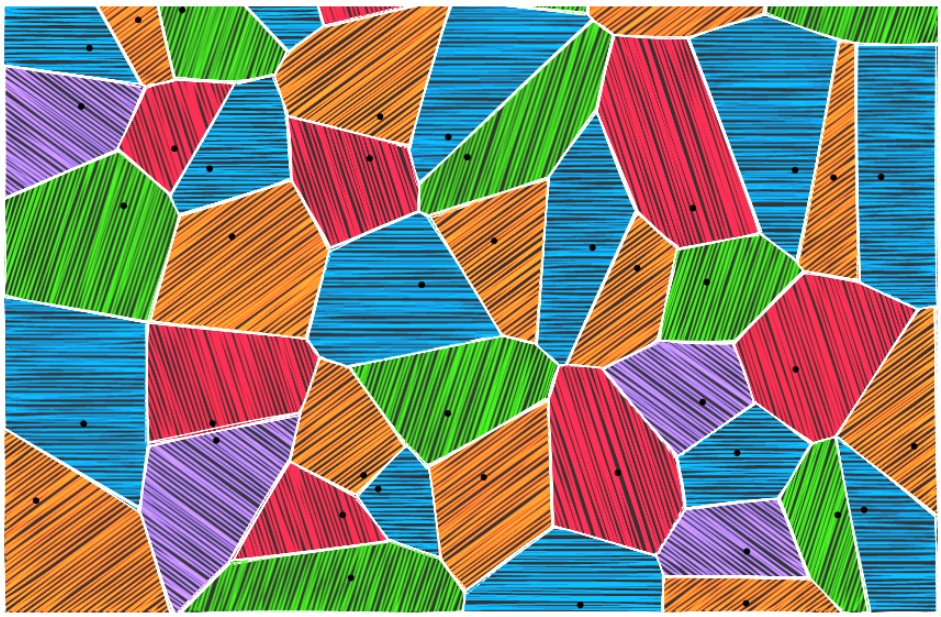


Figure 2.13: **An example 4 colour matching** This uses the first implementation of the algorithm mentioned in Subsection 2.4.4. The greedy approach does not often find the optimum solution, which may result in 5 colours instead. Observable Notebook : Daniel Ellis [2019]

Having defined all the visualisation techniques we move on to explain the clustering algorithms which are used, and how 'goodness of fit' may be measured in the clustering context.

2.5 Cluster Evaluation

The previous section discussed methods of visualising the reduced data for use with interactive exploratory data analysis. In this section we look at the use of vector clustering algorithms[9] (Subsection 2.5.1) to highlight groups in a 2D dataset, as well an automated method of assessing the quality of the clusters selected (Subsubsection 2.5.1.1) and feature extraction (Subsection 2.5.2).

2.5.1 Automated Selection Of Clusters

When it comes to clustering data points in a dataset, there exist a range of methods which may accomplish a task, Figure 2.14. Most often, the k-means [MacQueen, 1967], is used as it is fast and straightforward to understand. However, its linear method of partitioning cannot capture the splits between non-linear relationships of real data. The other problem is that an estimate for the number of expected clusters is required - something that is often unknown without prior understanding of the data. When this is the case, often it is easier to select the nodes with interactivity manually.

[9]Vector clustering is the grouping of data based on their proximity or density to other nearby points

In contrast, density-based clustering techniques such as GMM ([Pedregosa et al., 2011a]) or DBSCAN ([Ester et al., 1996]) tend to be better at locating non-linear trends in the data. The DBSCAN algorithm asses the distribution of data across a specific location. This allows clusters with a high density of datapoints to be located without the need for a predefined number as an input. Another method: OPTICS (Ordering Points To Identify the Clustering Structure) [Ankerst et al., 1999], shall be used[10]. This is an adaptation of the DBSCAN algorithm which does not require the specification of a minimum distance between points (for the density estimate)- instead, we specify a gradient for the distribution and the minimum number of points for a cluster to be classified.

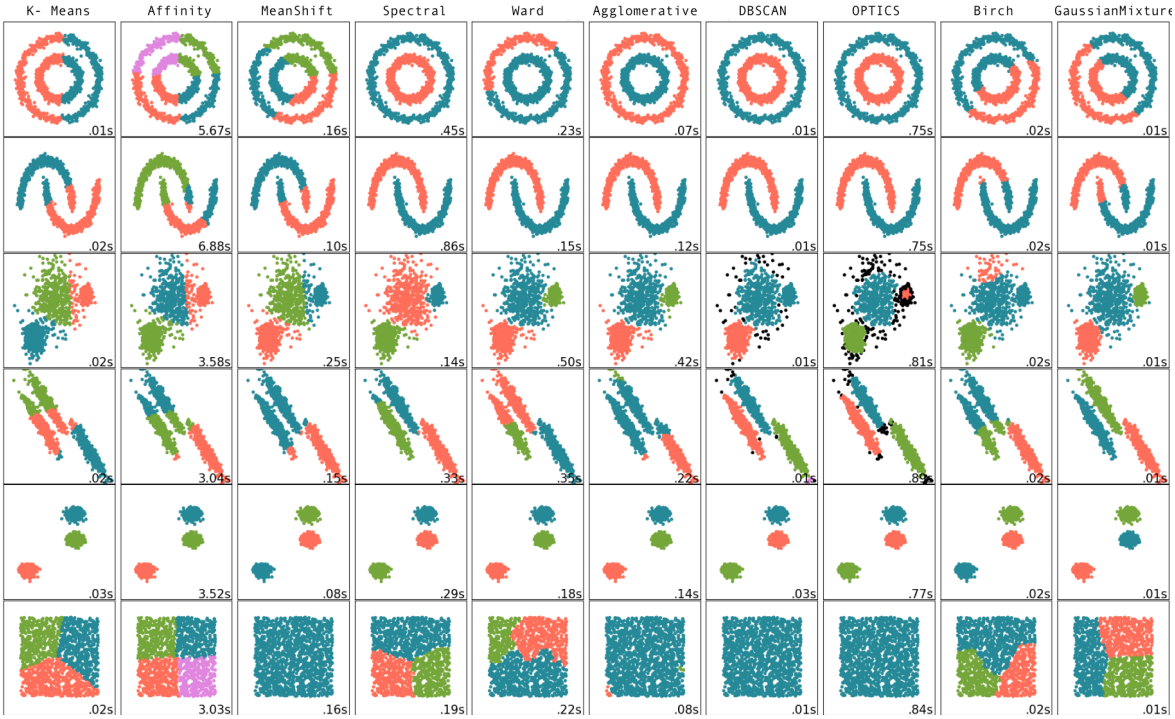


Figure 2.14: **A comparison of different clustering methods on a toy dataset.** The plot shows the performance of several vector clustering algorithms in Scikit-Learn. Cluster algorithms are represented across the horizontal axis and several types of datasets are across the vertical. Clustered groups are coloured. Source: sklearn [2019]

When deciding which algorithms to use, each algorithms' ability to partition non-linear data is considered. The first two rows of Figure 2.14 show data which cannot be partitioned linearly, here spectral, DBSCAN and optics are the only clustering algorithms to identify both correctly. It is for this reason that we shall look at these for the remainder of the chapter.

In selecting a value for the results section, several clustering algorithms, with a wide range of input parameters, are run. From these, the simulation with the best silhouette coefficient (Subsubsection 2.5.1.1) is taken.

[10]If using Python 2, the library for this needs to be extracted from the sci-kit-learn library for python3 package and altered to run with the previous version. (See copy in attached code.)

2.5.1.1 Clustering (Silhouette) Coefficient

The silhouette measure is a tool used for acessing the validity of a set of clusters. Here each cluster is represented as a silhouette, based on the comparison of its tightness and separation. To calculate the silhoette coefficient we look at the intra-cluster a and the mean inter-cluster[11] distance b. The silhouette cluster can then be described using **??**:

$$s(i) = \frac{b(i) - a(i)}{\max a(i), b(i)} \tag{2.6}$$

This gives a value $-1 \leq s(i) \leq 1$. Values near zero suggest overlapping clusters, 1 - dense, well-separated clusters and negative values indicate that a sample may have been incorrectly classified. In using this method, we can get an overview of how well individual objects lie within their assigned cluster.

2.5.2 Feature Extraction

Upon establishing a set of DR datasets, and their groups (the clusters of species they contain), it is important to evaluate what input features they represent. Rather than doing this manually we make use of Random Forests - described below.

2.5.2.1 Random Forrests

Random forests [Breiman, 2001], are a subset of ML algorithms called ensemble learning. This means that they train a large number of decision trees, each on a random subset of the original features. A decision tree is a tree formed from a series of conditionals[12], much like a perceptron network (**??**) with binary activation functions. Random forests introduce a level of additional randomness by selecting only a subset on which to create each decision tree. This may introduce a higher bias, but lowers the overall model variance, which creates a better (more robust) model. Such methods have been applied to replacing the computationally expensive process of chemistry integration of GEOS-Chem (a global 3D model of tropospheric chemistry) [Keller and Evans, 2019] and the prediction of global sea-surface iodine based on observations coupled with sea-surface temperature, depth, and salinity [Sherwen et al., 2019].

[11] Inside and between different clusters.

[12] Questions with a True/False answer

2.5.2.2 Calculating Importance Using Random Forrests

Since random forests are in essence a collection of decision trees, it is possible to generate a 'decision tree aggregate' to visualise the ensemble structure of the random forest [Ellis and Sherwen, 2019] (Figure 2.15). Alternatively, if all that is required is the relative importance of each feature, the RandomForrestClassifier from Pedregosa et al. [2011b] provides a quick and easy way of understanding which features matter, [Géron, 2017]. This works by aggregating the weighted nodes which use a certain feature using the number of samples and then scales the result to 1. We use this method to access the overall importance of features within each DR output and identify the differences between clusters.

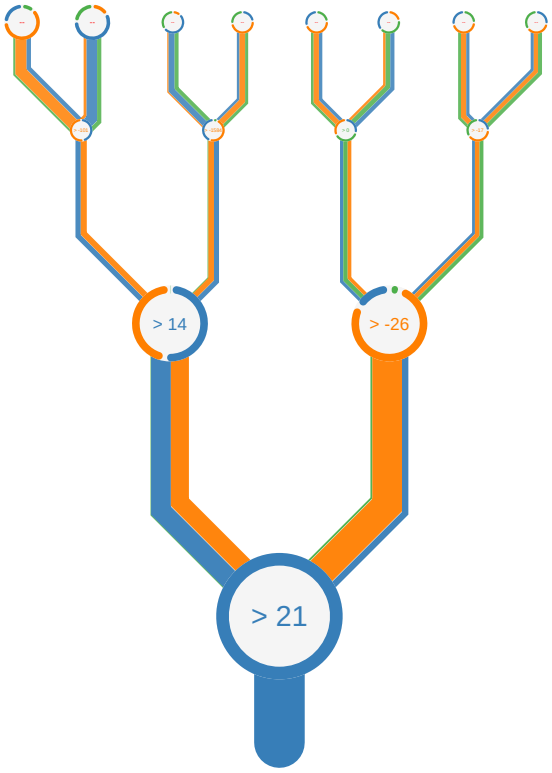


Figure 2.15: **A decision tree aggrogate from a random forrest plotted with the Epiphyte version of the TreeSurgeon program [Ellis and Sherwen, 2019].** The data originates from Sherwen et al. [2019] and the imporance of Tempearature (blue), Depth (orange) and Chlorophyll *a* (green).

NOTE: The only downside is that Random Forrests are in themselves ML techniques which also need to be evaluated. To do this, as they are simply being used as indicators of cluster properties which we are to explore further, we can initiate a collection of 300 random Forrest classifiers, from which we take the median. A sort of ensemble learning from an ensemble.

2.6 Results

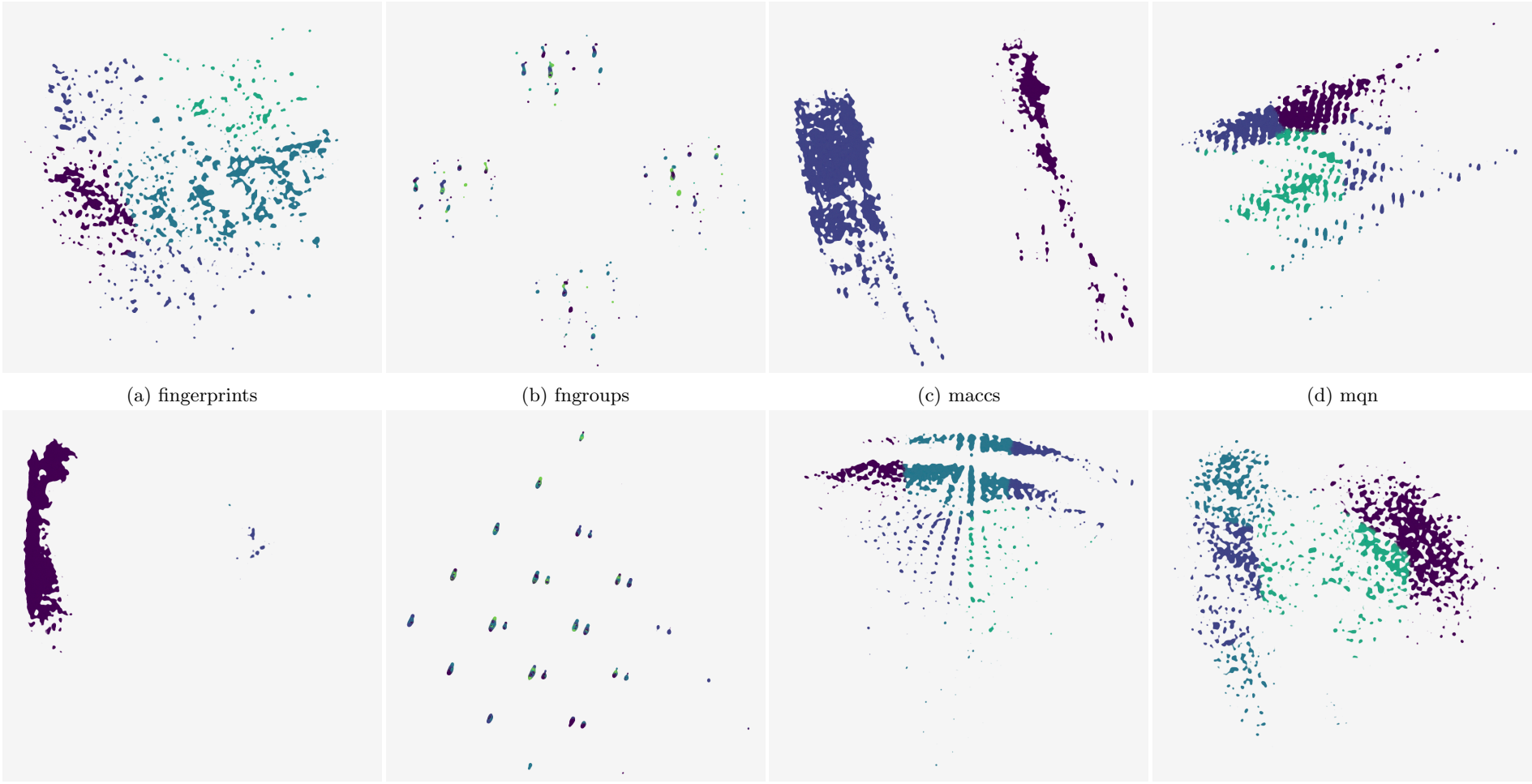
There exist many methods to of defining the chemical structure of species within the MCM. This section evaluates the different structural representations (Section 2.2) and the ability of DR algorithms to separate the chemical space within which these lie into a two dimensional scatterplot.

2.6.1 Visual Overview

Explorative data analysi invloves a degree of figure interativity. **??** described the imporance of visualisation in employing the pattern cognitive functions of the human brain, and **??** explained the importance of having an evenly distributed datapoints to aid the understanding of graphs. This subsection combignes the two ideas in the usage of dimsenionality recduction to exploit patterns within a dataset.

Using the techneques in we explore the visual distribution of different dimensionality reduction methods across all input types. This subsection explores the spatial distribution of groups (blobs) in the figures. The colours represent the automatically calculated clusters, which shall be looked at in a more analytical method within Subsection 2.6.2. The individial meanings of clusters are also explored in Subsection 2.6.3.

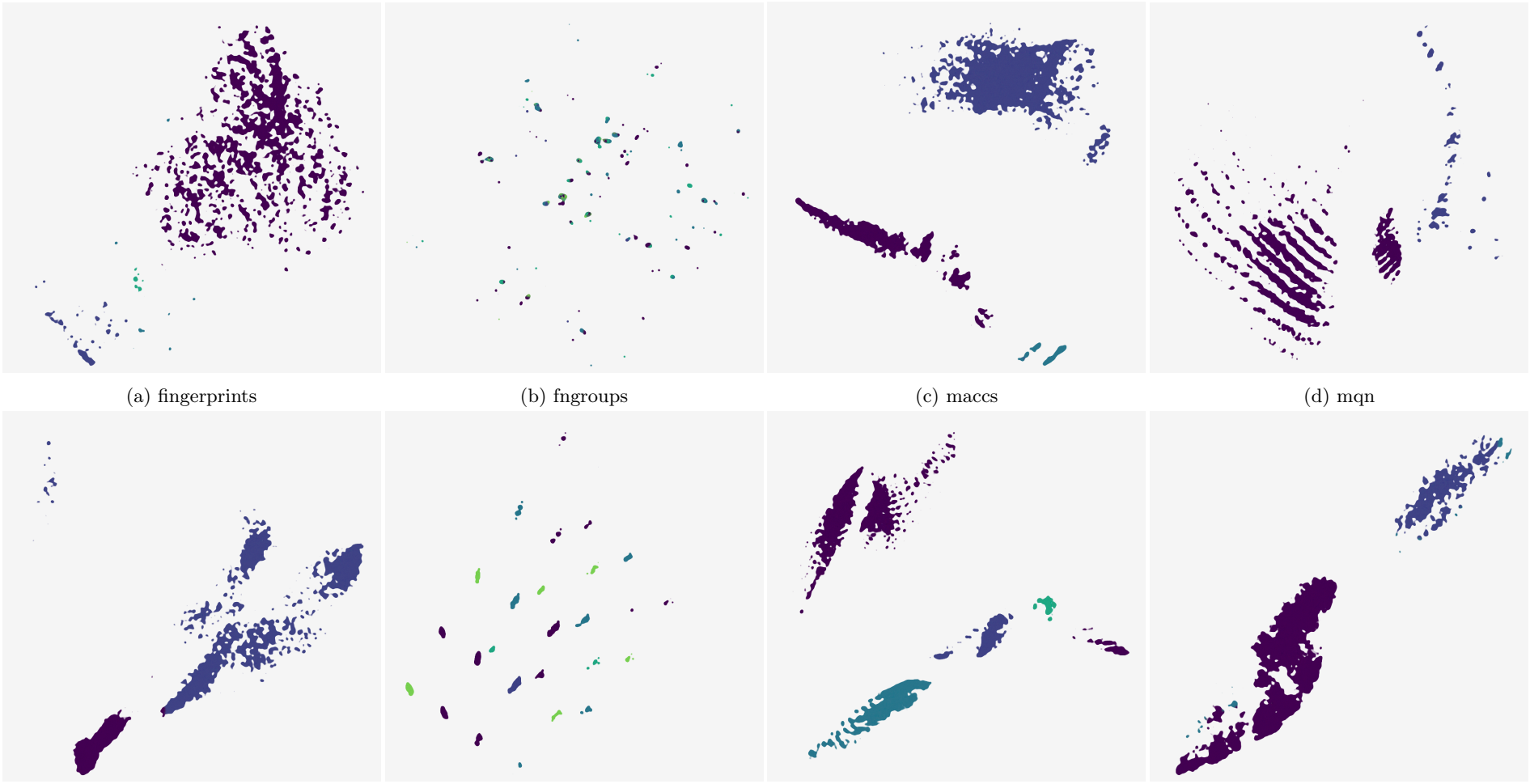
Using priciple component analysis, the linear decomposition results in a uniform distribution of datapoints, often forming lines of simialar properties - **??**. Since an AutoEncoder with linear activation functions will produce identical results to the PCA algorithm, it is not surprising that some of the linear separation (especially for ordinal data such as (b) functional groups and (f) protocol groups) can be seen across the two algorithms. However since compression using the AutoEncoder does not discard data, and can allow for non-linear relationships, this produces better grouped clusters, especially for (g) smiles and (d) MQN inputs - Figure 2.17. Whereas the spatial positioning in both PCA and AE contains an inherent meaning about the data (distances have a meaning), t-SNE is a non-linear DR method, where the distance and positioning of a cluseter holds no such information. Instead it uses a graph-like force model to group features of similarity. The result of this is one where items are better separated, but their separation contains little to no meaning about the dataset. Out of the three discussed methods, t-SNE makes it the easiest to visually isolate clusters from their neighbours (Figure 2.18). This has its application in data exploration and communication of the results.



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

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

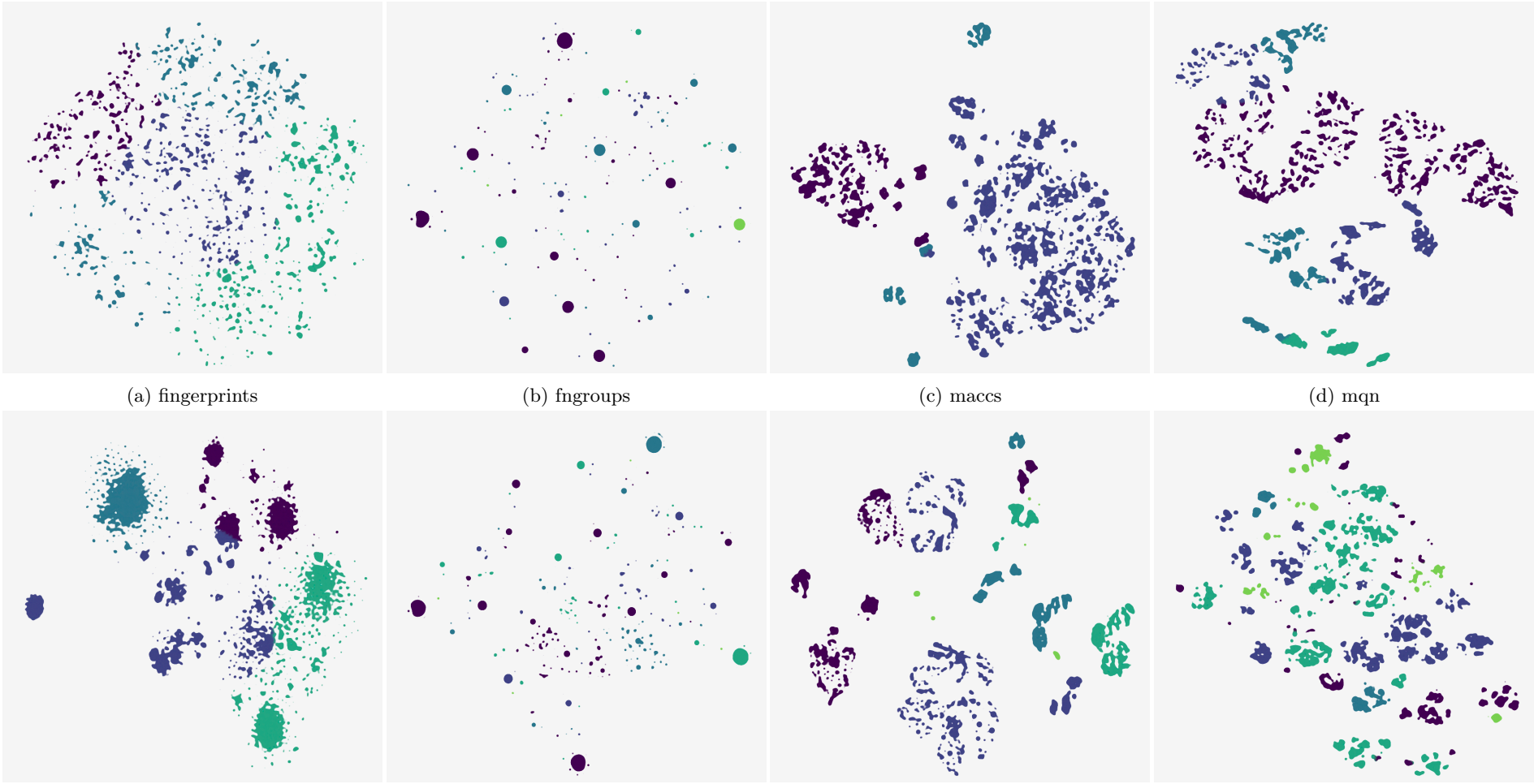
Figure 2.16: **Comparing clusters for all inputs after a reduction to 2 dimensions using Principle Component analysis.** Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.



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

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

Figure 2.17: **Comparing clusters for all inputs after a reduction to 2 dimensions using an AutoEncoder.** Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.



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

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

Figure 2.18: **Comparing clusters for all inputs after a reduction to 2 dimensions using t-SNE.** Each graph has undergone several clustering algorithms under a range of parameters. The result with the best silhouette coefficient has been chosen. Colours follow the greedy four colour theorem and are there only to indicate the contrast between cluster boundaries.

2.6.2 Mathematical Cluster Analysis

Subsection 2.6.3 explores the distribution of functional groups within automatically selected clusters. Here the silhouette coefficient (Subsubsection 2.5.1.1) is used to determine the vector clustering technique and its input configuration for the output of a DR algorithm. The silhouette coefficient can be used to give a mathematical representaion of the goodness of fit of points within a specific cluster. Using the best selected method of auto classification (Subsection 2.5.1) we compare the silhouette coefficients and the number of clusters for each input and DR algorithm.

Tables 2.2 - 2.4 show the silhouette coefficients and number of groups for each DR algorithm. Inputs for each algorithm are ranked in order of their silhouette coefficient, where the closer to 1 the value, the better the clustering.

Ordinal inputs such as functional groups or the protocol categories consistently rank the highest within each algorithm. This is beacuse the algorithms only have to idenitfy permutations of each category and classify the species into these. It is noted that the t-SNE silhouette coefficient for these is 30% lower than for PCA and AE algorithms. However the number of groups located is also greatly reduced from 140 to 106. This suggests that the vector clustering algorithms have tried to combign datapoints into a group, which has come at a cost to the silhouette value.

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

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

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

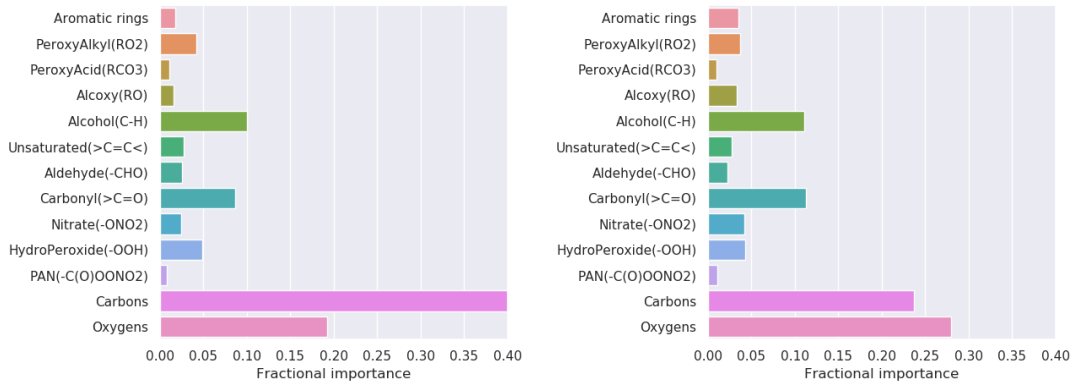
2.6.3 Feature Selection Comparison

In the previous subsection we acessed the ability of 3 DR algorithms to separate the chemistry of a mechanism into distinct, well defined, clusters. This subsection looks explores the functional groups which are responsible for the greatest variation between clusters for each input method (across all clustering algorithms). To determine the importance of each group, the output of a random forrest classifer is plotted in Figures 2.19-2.21.

The first thing that is apparent is that group importance is persistant across all the different dimensionality reduction techneques. Although values may vary, the most prevalent features seem to exist for all DR algorithms. This suggests feature importance is related to species representation (input) and not the machine learning (ML) algorithm itself. The positive outcome of this suggests that should an input which separates species in the style we are most interested in be foumd, this can be directly applied into other ML methods - such as random forrests or graph neural networks.

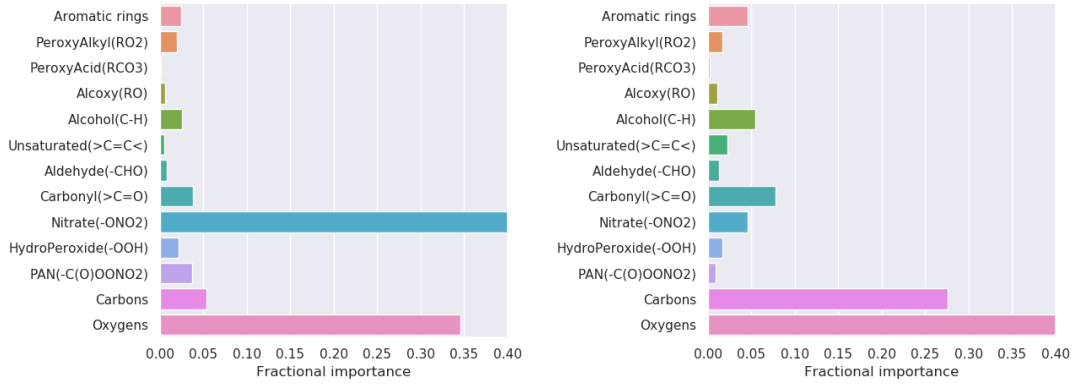
Havign established that the most important functional groups are persistant across all DR algorithms, we observe which of these are are identified for each input. When discussing these, the characters (a-h) shall be used, and refer to the corresponding subplots in all Figures 2.19-2.21.

Starting with species name (h) the number of Oxygens and Carbons are



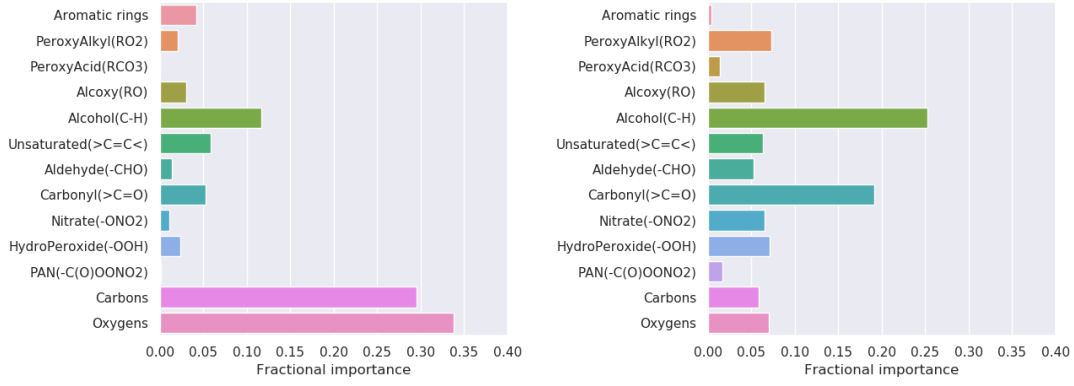
(a) fingerprints

(b) fngroups



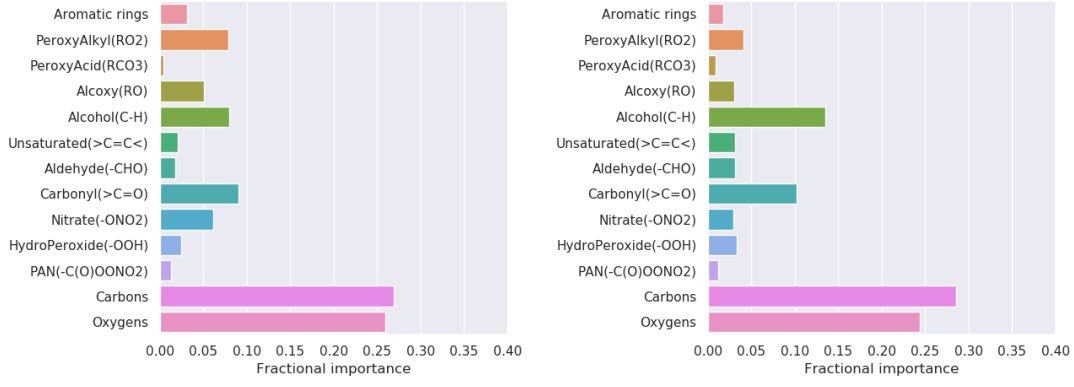
(c) maccs

(d) mqn



(e) node2vec

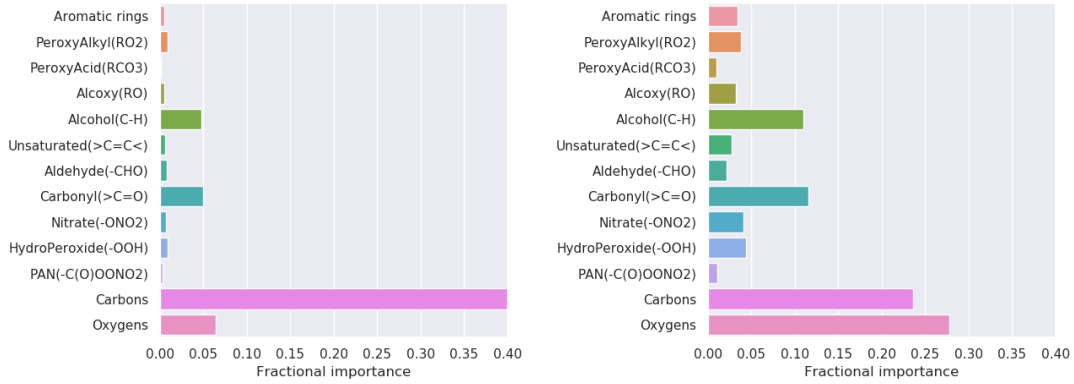
(f) protocol



(g) smiles

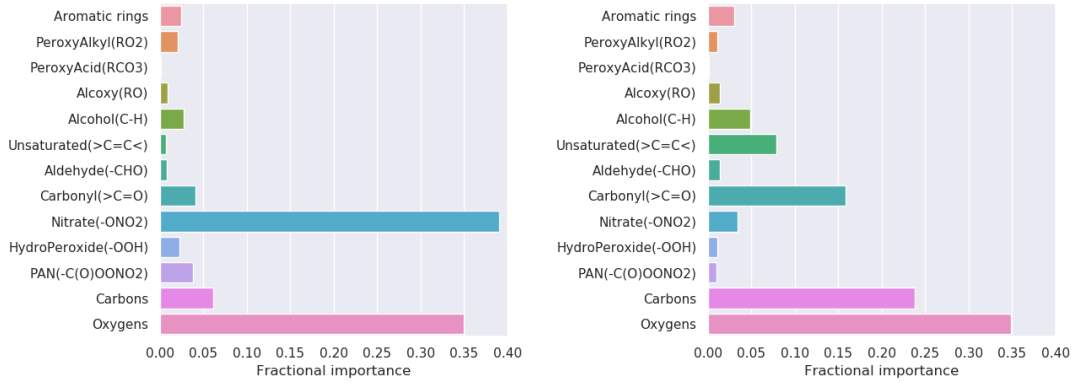
(h) spec

Figure 2.19: **Comparing feature importance for PCA clusters.**



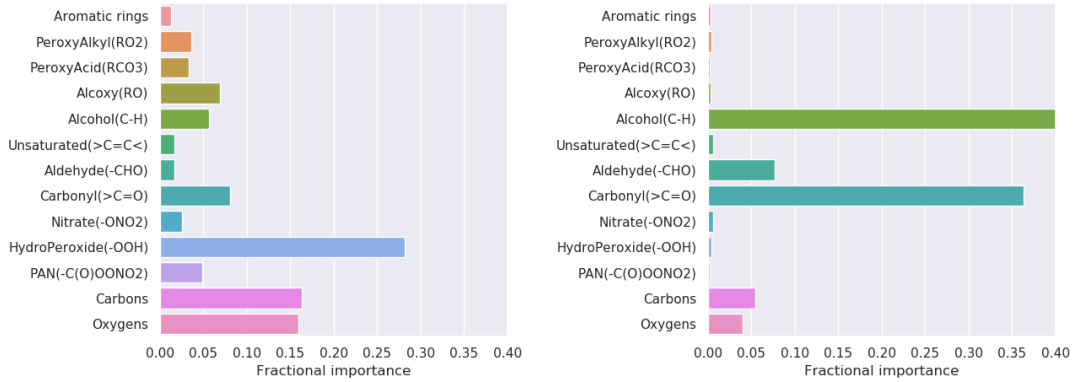
(a) fingerprints

(b) fngroups



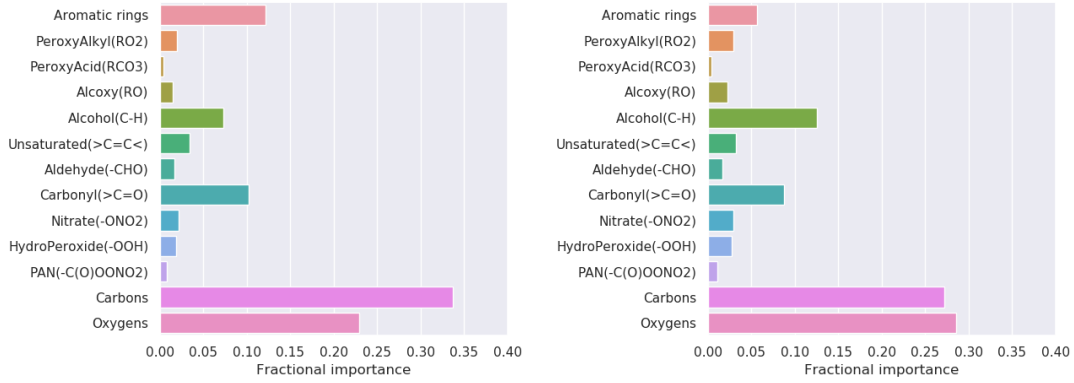
(c) maccs

(d) mqn



(e) node2vec

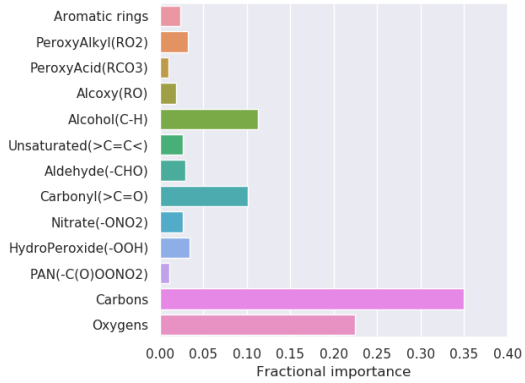
(f) protocol



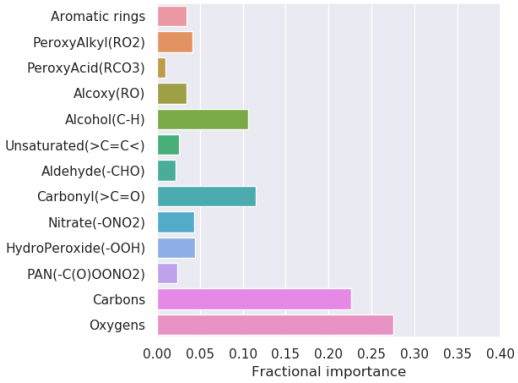
(g) smiles

(h) spec

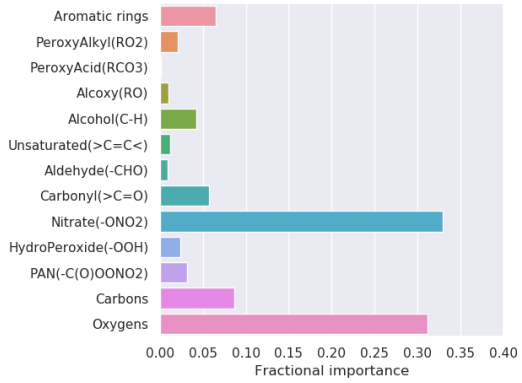
Figure 2.20: **Comparing feature importance for AE clusters.**



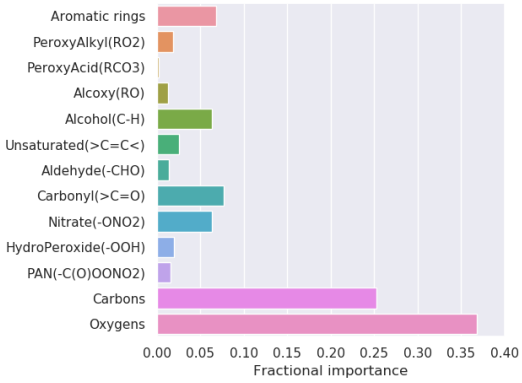
(a) fingerprints



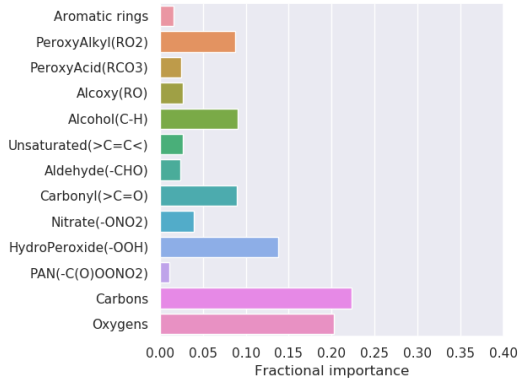
(b) fngroups



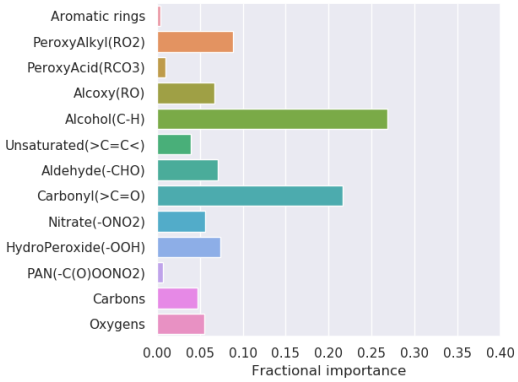
(c) maccs



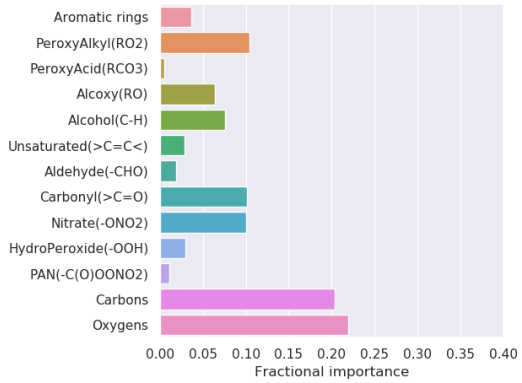
(d) mqn



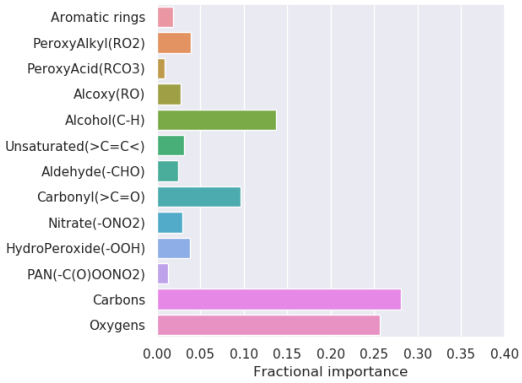
(e) node2vec



(f) protocol



(g) smiles



(h) spec

Figure 2.21: **Comparing feature importance for t-SNE clusters.**

2.6.4 Individual Cluster Comparison

Figure 2.22: **Case Study 1: t-SNE MQN.** We compare the functional group distribution for individual clusters within the t-SNE 2D representation of the Mollecular Quantum Number fingerprint.

2.7 Conclusions

tsne best - but has to be rerun

others can have data fed into them and reconstructed.

DR can be used to find patterns in dataset which is best

interaction

There are a range of inputs each showing a few different things.

Depending on what properties we are intereste dwe may select accordingly

Bibliography

Alon, U., Zilberstein, M., Levy, O., and Yahav, E. (2019). Code2Vec: Learning Distributed Representations Of Code. http://dl.acm.org/citation.cfm?doid=3302515.3290353.

Anderson, C. (2008). The End Of Theory: The Data Deluge Makes The Scientific Method Obsolete. *online*. http://www.wired.com/print/science/discoveries/magazine/16-0.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60. https://doi.org/10.1145/304181.304187.

Appel, K. and Haken, W. (1976). Every planar map is four colorable. *Bull. Amer. Math. Soc.*, 82(5):711–712. https://projecteuclid.org:443/euclid.bams/1183538218.

Aumont, B., Szopa, S., and Madronich, S. (2005). Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: Development of an explicit model based on a self generating approach. *Atmospheric Chemistry and Physics*, 5(9):2497–2517. https://www.atmos-chem-phys.net/5/2497/2005/.

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

Bostock, M. (2012). D3.js - data-driven documents. http://d3js.org/.

Box, G. E. P. (1976). Science And Statistics. *Journal of the American Statistical Association*, 71(356):791–799. https://www.tandfonline.com/doi/abs/10.1080/01621459.1976.10480949.

Breiman, L. (2001). Random Forests. *Machine learning*, 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Cohen, E. (2018). Node2Vec: Embeddings For Graph Data. https://towardsdatascience.com/node2vec-embeddings-for-graph-data-32a866340fef.

Daniel Ellis (2019). D3-Fourcolour Voronoi. https://observablehq.com/@wolfiex/d3-fourcolour-voronoi.

Dataman (2019). Convolutional Autoencoders For Image Noise Reduction. https://towardsdatascience.com/convolutional-autoencoders-for-image-noise-reduction-32fce9fc1763.

Descartes, R. and Lafleur, L. J. (1960). *Meditations On First Philosophy*. Bobbs-Merrill New York. http://selfpace.uconn.edu/class/percep/DescartesMeditations.pdf.

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., and Zhi, D. (2019). Gene2Vec: Distributed Representation Of Genes Based On Co-Expression. *BMC genomics*, 20(Suppl 1):82. http://dx.doi.org/10.1186/s12864-018-5370-x.

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. https://www.ncbi.nlm.nih.gov/pubmed/12444722.

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.

Ellis, D. and Sherwen, T. (2019). Wolfiex/treesurgeon: Wollemia. https://doi.org/10.5281/zenodo.3346817.

Ester, M., peter Kriegel, H., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press.

F.R.S., K. P. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572. https://doi.org/10.1080/14786440109462720.

Géron, A. (2017). *Hands-On Machine Learning With Scikit-Learn And Tensorflow: Concepts, Tools, And Techniques To Build Intelligent Systems*. O'Reilly Media. https://books.google.co.uk/books?id=khpYDgAAQBAJ.

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

Hamadache, M. and Lee, D. (2017). Principal Component Analysis Based Signal-To-Noise Ratio Improvement For Inchoate Faulty Signals: Application To Ball Bearing Fault Detection. *International journal of control, automation, and systems*, 15(2):506–517. https://doi.org/10.1007/s12555-015-0196-7.

Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev (2013). Inchi - The Worldwide Chemical Structure Identifier Standard. *Journal of cheminformatics*, 5(1):7. http://dx.doi.org/10.1186/1758-2946-5-7.

Hernandez, W. and Mendez, A. (2018). Application Of Principal Component Analysis To Image Compression. In Göksel, T., editor, *Statistics - Growing Data Sets and Growing Demand for Statistics*. InTech. http://www.intechopen.

com/books/statistics-growing-data-sets-and-growing-demand-for-statistics/application-of-principal-component-analysis-to-image-compression.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441. https://doi.org/10.1037%2Fh0071325.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., and Ermon, S. (2018). Tile2Vec: Unsupervised Representation Learning For Spatially Distributed Data. http://arxiv.org/abs/1805.02855.

Jolliffe, I. T. and Cadima, J. (2016). Principal Component Analysis: A Review And Recent Developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202. http://dx.doi.org/10.1098/rsta.2015.0202.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). Scipy: Open source scientific tools for Python. http://www.scipy.org/.

Keller, C. A. and Evans, M. J. (2019). Application of random forest regression to the calculation of gas-phase chemistry within the geos-chem chemistry model v10. *Geoscientific Model Development*, 12(3):1209–1225. https://www.geosci-model-dev.net/12/1209/2019/.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet Classification With Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

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. https://doi.org/10.5281/zenodo.2864247.

Leite, N. M. N., Pereira, E. T., Gurjão, E. C., and Veloso, L. R. (2018). Deep convolutional autoencoder for eeg noise filtering. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2605–2612.

Lynch, H. (2011). *Infant Places, Spaces And Objects: Exploring The Physical In Learning Environments For Infants Under Two*. PhD thesis. http://dx.doi.org/10.21427/D73W37.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing Data Using T-Sne. *Journal of machine learning research: JMLR*, 9(Nov):2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512992.

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

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation Of Word Representations In Vector Space. http://arxiv.org/abs/1301.3781.

Morozov, A. (2016). Modelling biological evolution: Linking mathematical theories with empirical realities. *Journal of Theoretical Biology*, 405:1 – 4. http://www.sciencedirect.com/science/article/pii/S0022519316301849.

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. http://dx.doi.org/10.1002/cmdc.200900317.

Noble, C. E. (1957). Human Trial-And-Error Learning. *Psychological reports*, 3(2):377–398. https://doi.org/10.2466/pr0.1957.3.h.377.

Oliphant, T. (2006). Guide to numpy.

Oliveira, B., Pereira, F., de AraÃ°jo, 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. http://www.sciencedirect.com/science/article/pii/S000926140600861X.

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. http://dx.doi.org/10.1002/jcc.20237.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011a). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830. http://dl.acm.org/citation.cfm?id=1953048.2078195.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-Learn: Machine Learning In Python . *Journal of Machine Learning Research*, 12:2825–2830.

People2Vec (2019). People2Vec. http://people2vec.org/.

Powell, V. (2020). Principal Component Analysis Explained Visually. http://setosa.io/ev/principal-component-analysis/.

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. http://dx.doi.org/10.1021/acs.jcim.7b00425.

rdkit (2019). Rdkit. https://github.com/rdkit/rdkit/blob/24f1737839c9302489cadc473d8d9196ad9187b4/rdkit/Chem/MACCSkeys.py.

Reinhardt, A. (1975). *Art-As-Art: The Selected Writings Of Ad Reinhardt*. Documents of 20th-century art. Viking Press. https://books.google.co.uk/books?id=zyK4AAAAIAAJ.

Roberts, R. (1989). *Serendipity: Accidental Discoveries In Science*. Wiley Science Editions. Wiley. https://books.google.co.uk/books?id=hf57X0s4aPwC.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326. https://science.sciencemag.org/content/290/5500/2323.

Sherwen, T., Chance, R. J., Tinel, L., Ellis, D., Evans, M. J., and Carpenter, L. J. (2019). A machine-learning-based global sea-surface iodide distribution. *Earth System Science Data*, 11(3):1239–1262. https://www.earth-syst-sci-data.net/11/1239/2019/.

sklearn (2019). Comparing Different Clustering Algorithms On Toy Datasets — Scikit-Learn 0.21.3 Documentation. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html.

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. http://dx.doi.org/10.1126/science.aai8636.

T. Leube, B., Inglis, K., J. Carrington, E., and Sharp, P. (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.

Turanyi, T. and Tomlin, A. (2015). *Analysis Of Kinetic Reaction Mechanisms*. Springer. http://eprints.whiterose.ac.uk/84294/.

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. http://dx.doi.org/10.1002/anie.200800827.

Watson, D. F. (1981). Computing The N-Dimensional Delaunay Tessellation With Application To Voronoi Polytopes*. *The Computer Journal*, 24(2):167–172. https://doi.org/10.1093/comjnl/24.2.167.

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. https://pubs.acs.org/doi/abs/10.1021/ci00057a005.

Yu-ChenLo (2018). Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538 – 1546. http://www.sciencedirect.com/science/article/pii/S1359644617304695.

Appendices

Appendix A

Supplementary Mathematics

A.1 PCA

A.1.1 Statistics

Firstly we define the variance:

$$\sigma = \frac{\sum_{i=1}^{N}(X - \mu_X)(X - \mu_X)}{n - 1} \tag{A.1}$$

where X is the dataset, μ the mean and n the number of datapoints.

If we wish to then compare dataset X with dataset Y we may use the covarance:

$$cov(X, Y) = \frac{\sum_{i=1}^{N}(X - \mu_X)(Y - \mu_Y)}{n - 1} \tag{A.2}$$

For n distinct variables we may construct an $n \times n$ matrix containing $n!/(n-2)! \times 2$ different combinations of covariences:

$$C = \begin{pmatrix} \sigma_X & cov(X,Y) & cov(X,Z) & \cdots & cov(X,n) \\ cov(Y,X) & \sigma_Y & cov(Y,Z) & \cdots & cov(Y,n) \\ cov(Z,X) & cov(Z,Y) & \sigma_Z & \cdots & cov(Z,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(n,X) & cov(n,Y) & cov(n,Z) & \cdots & \sigma_n \end{pmatrix}$$

A.1.2 Matrices and Eigenvectors

An eigenvector is a vector $\mathbf{v}$, that when operated on by a given operator produces a scalar multiple of itself (Equation A.3) - this scalar multiple is called the eigenvalue λ. Eigenvectors can only be found for square matrices and are perpendicular to the matrix regardless of their dimension. A $n \times n$ matrix will produce n eigenvectors. Conventionally these are scaled to unity, which may be done by dividing the eigenvector by the pythoagorean distance of each element.

$$C\mathbf{v} = \lambda\mathbf{v} \tag{A.3}$$

An example of an eigenvector/value pair is shown in the following equations:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 4 \begin{pmatrix} 3 \\ 2 \end{pmatrix} \tag{A.4}$$

One property of the eigenvalue/eigenvector pair is that the square matrix acts as a transformation on the eigenvector. This means that we may treat the eigenvector as a direction from the origin, whose magnitude we can scale. The eigenvalue however remains scale independent and is the same value as before:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \times \begin{pmatrix} 6 \\ 4 \end{pmatrix} = 4 \begin{pmatrix} 6 \\ 4 \end{pmatrix} \tag{A.5}$$

A.2 t-SNE

A.2.1 Student T distribution

Created by William Gosset and published under the pseudonym student [1] **?**.

The distribution consists of a family of continuous probability distributions which may be used when sample size is small and the standard deviaton is unknown. The curve itself resembles that of a normal distribution, just with a shorter amplitude and greater full width at half maximum (FWHM).

A.2.1.1 T-Score

Much like the z-score mentioned earlier [ref standardiz], t-scores also convert individual values to a standard form. This is genearlly ised when you dont know the population standard deviation (often due to having too few datapoints). At greater than 30 datapoints this resembles the equation of the z-score, and will often give you the same result.

$$t(x_i) = \frac{x_i - \mu_x}{S_{sample}/\sqrt{n}} \tag{A.6}$$

A.2.2 Kullback-Leiber (KL) divergence

KL divergence (also known as relative entropy) is a measure of distance between two distributions. It arises **?**.

`https://medium.com/syncedreview/kullback-leibler-divergence-explained-e358fbacf046`

[1] At the time Gosset was employed by Guinness Brewries in Dublin. This meant that chemists were forbidden from publishing their findings. After explaining that his mathematical and philosphical conculaions were of no use to competing brewries he was finally allowed to publish under the pseudonym 'student'. This was mainly to avoid difficulties with the rest of the staff.

Appendix B

Neural Network Activation Functions

B.1 Binary Step

.

This is a simple threshold function. If the input is above the threshold, the message is passed on. This makes it efficient, but unable to classify a single input into multiple categories. This can be likened to a yes|no decision tree.

$$f(x) = \begin{cases} 1, & \textbf{if } x < threshold \\ 0, & \text{otherwise} \end{cases} \tag{B.1}$$

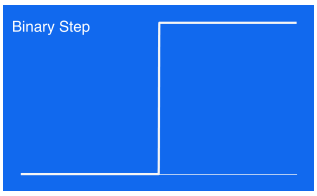


Figure B.1: Binary Step activation function.

B.2 Linear

.

This produces a signal proportional to the input multiplied by each neurons weight. It is an improvement over the step function as it allows for multiple outputs. It does however mean that we are unable to use backpropagation (gradient descent) to train the model. In adition to not being able to improve a model, all the layers in the neural network collapse into a single layer. This means that the final layer will always be a linear function of the first layer. This eliminates all the merits which may be gained from deep learning. A neural network with a linear activation function is simply a linear regression model.

$$f(x) = m(x) \tag{B.2}$$

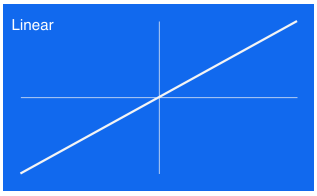


Figure B.2: Linear activation function.

B.3 Sigmoid / Logistic

.

The first of the non-linear activation functions. It has a smooth gradient providing smooth output values which are bound between 1 and 0, normalising the output of each neuron. The main disadvantage is that is falls foul the vanishing gradient problem - for extreme values of x there is close to no change in the prediction. This may result in either early termination of the training, or a slow training cycle in obtaining adequate precision. The activations is computationally expensive and the outputs are not zero centred.

$$f(x) = 1/1 + e^x \tag{B.3}$$

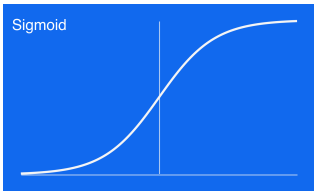


Figure B.3: Sigmoid activation function.

B.4 Hyperbolic Tangent

.

Much like the sigmoid function in both advantages and disadvantages. The hyperbolic tangent function provides a smooth curve which is zero centred. It is however computationally expensive and suffers from the vanishing gradient problem.

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{B.4}$$

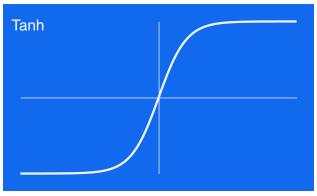


Figure B.4: Tanh activation function.

B.5 Rectified Linear Unit

.

A commonly used activation for large deep neural networks, due to its computational efficiency and quick convergence. It is non-linear although it appears like a linear function, and allows for back propagation. It does however suffer from the dying ReLU problem - when inputs tend to zero or below, the gradient of the function becomes zero and the network cannot perform backpropagation to learn.

$$f(x) = \begin{cases} 0, & \textbf{if } x < threshold \\ x, & \text{otherwise} \end{cases} \tag{B.5}$$

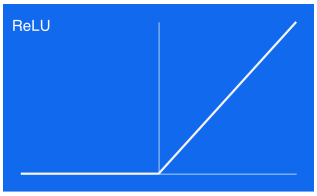


Figure B.5: ReLU activation function.

B.6 Swish

.

https://arxiv.org/abs/1710.05941v1 ***a new, self-gated activation function discovered by researchers at Google. According to their paper, it performs better than ReLU with a similar level of computational efficiency. In experiments on ImageNet with identical models running ReLU and Swish, the new function achieved top -1 classification accuracy 0.6-0.9% higher.***

$$f(x) = x/1 - e^{-x} \tag{B.6}$$

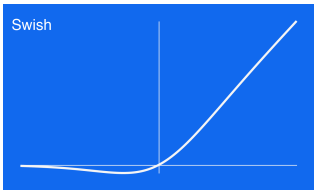


Figure B.6: Swish activation function.

B.7 A note on backpropagation

As it has not been explicitly explained before back-propagation is an algorithm used to train neural networks. The derivative (or gradient) of an activation function is important in the use of back propagation. Here the model weights are adjusted, and improved, by tracing back all the connections in network, suggesting an optimal weight of each neuron.

Appendix C

Miscellaneous

C.1 Correspondance with Mike Jenkin

Mike Jenkin 11th September 2019

Note on naming conventions in the CRI mechanism

The lumped or ‘‘common’’ species in the CRI mechanism are, by definition, used to represent a set of real species with different structures and properties. The criterion for lumping is the maximum number of NO-to-NO2 conversions (i.e. maximum number of ozone molecules) that the subsequent degradation can produce - and lumped species can therefore represent a large number of real species with different structures and properties.

In later expansions of the mechanism, the chemistry for species such as isoprene and terpenes defined intermediates that are representative of more restricted sets of real species. For these, it is possible to relate them to more restricted sets of MCM species that are the main contributors.

Although I tried to be logical in naming, the mechanism was developed over many years with little or no funding and may therefore not be fully transparent and foolproof throughout. However, I think quite a lot of the naming is logical, as expanded on below.

1) The numbers in most of the species names (the ‘‘CRI index’’) are the number of NO-to-NO2 conversions that can result from the subsequent OH-initiated NO-propagated chemistry. For radical termination products (e.g. hydroperoxides formed from RO2 + HO2 and nitrates formed from RO2 + NO), this is a grey area, and the number is therefore the same as that for the precursor RO2 radical. In these cases it is simply a convenient label.

2) There are a number of series of peroxy radicals, which are denoted RNxxO2, RIxxO2, RAxxO2, RExxO2, RUxxO2, RTNxxO2, RTXxxO2. These represent peroxy radicals with different structural features or formed from different types of precursor, as indicated below. Occasionally, extra peroxy radicals with the same CRI index are included by inserting a letter after the index (e.g. RNxxAO2) to increase flexibility of the mechanism. Peroxy radicals formed specifically from addition of NO3 to an alkene/diene are prefixed by ‘‘N’’.

RNxxO2: These were originally representative of peroxy radicals formed from linear or ‘‘n-‘‘ alkanes and their carbonyl products. They are also used for peroxy radicals formed from slightly-branched precursors (e.g. 2-methylhexane), and are formed as a convenient default intermediate with the correct CRI index in the latter stages of degradation of other precursor classes.

RIxxO2: These were originally representative of peroxy radicals formed from branched or ‘‘i-‘‘ alkanes and their carbonyl products, but tend to be used only for smaller branched precursors that can produce acetone as a major product from their subsequent degradation. This is because acetone is a particularly unreactive carbonyl, the formation of which can interrupt the ozone formation processes under typical regional-scale photochemical episode conditions in north-west Europe.

RAxxO2: These peroxy radicals are formed from the addition of OH to aromatic compounds, and are complex bicyclic structures containing a peroxide bridge (e.g. like BZBIPERO2 in the MCM).

RExxO2: These peroxy radicals are formed from ether degradation, and allow the formation of unreactive formate ester products to be represented.

RUxxO2: These peroxy radicals are formed from degradation of conjugated dienes (currently only isoprene and 1,3-butadiene). Those formed initially (e.g. RU14O2) contain allyl functionalities (i.e. a specific unsaturated linkage), although the terminology is also used for some peroxy radicals formed from subsequently-formed unsaturated products.

Related to this, the species CRU14O2 and TRU14O2 in the EMEP variant of CRI v2.2 (described in https://doi.org/10.1016/j.atmosenv.2019.05.055) were specifically introduced to represent the cis- and trans- isomers required for the Peeters (LIM) reaction framework. CRU14O2 represents CISOPAO2 and CISOPCO2 in MCM v3.3.1 and TRU14O2 represents ISOPAO2 and ISOPCO2 in MCM v3.3.1. However, CRI v2.2 itself uses a different approach where the chemistry is represented by a conditions-dependent rate coefficient for the single peroxy radical, RU14O2.

RTNxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an endocyclic double bond. This is currently limited to α-pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other endocyclic monoterpenes by simply adding new sets of initiation reactions.

RTXxxO2: This terminology is used for peroxy radicals formed from monoterpenes containing an exocyclic double bond. This is currently limited to β-pinene in CRI, although the original idea was that the mechanism could be used as a surrogate for other exocyclic monoterpenes by simply adding new sets of initiation reactions.

Finally, the species DHPR12O2 in CRI v2.2 is a peroxy radical containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C536O2 and C537O2 in MCM v3.3.1 (these species being referred to as ‘‘di-HPCARPs’’ by Peeters et al., 2014: https://doi.org/10.1021/jp5033146).

3) Hydroperoxides formed the reactions of HO2 with the above peroxy radicals have ‘‘OOH’’ in place of ‘‘O2’’. Nitrates formed the reactions of NO with the above peroxy radicals have ‘‘NO3’’ in place of ‘‘O2’’.

4) There are a number of series of carbonyl compounds, which are denoted CARBxx, UCARBxx, UDCARBxx, TNCARBxx and TXCARBxx.

CARBxx: These are used to represent carbonyls and hydroxycarbonyls. Occasionally, extra carbonyls/hydroxycarbonyls with the same CRI index are included by inserting a letter after the index (e.g. CARBxxA) to increase the flexibility of the mechanism.

Related to this, the species DHPCARB9 in CRI v2.2 is a carbonyl containing two hydroperoxy groups. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species DHPMEK and DHPMPAL in MCM v3.3.1 in MCM v3.3.1.

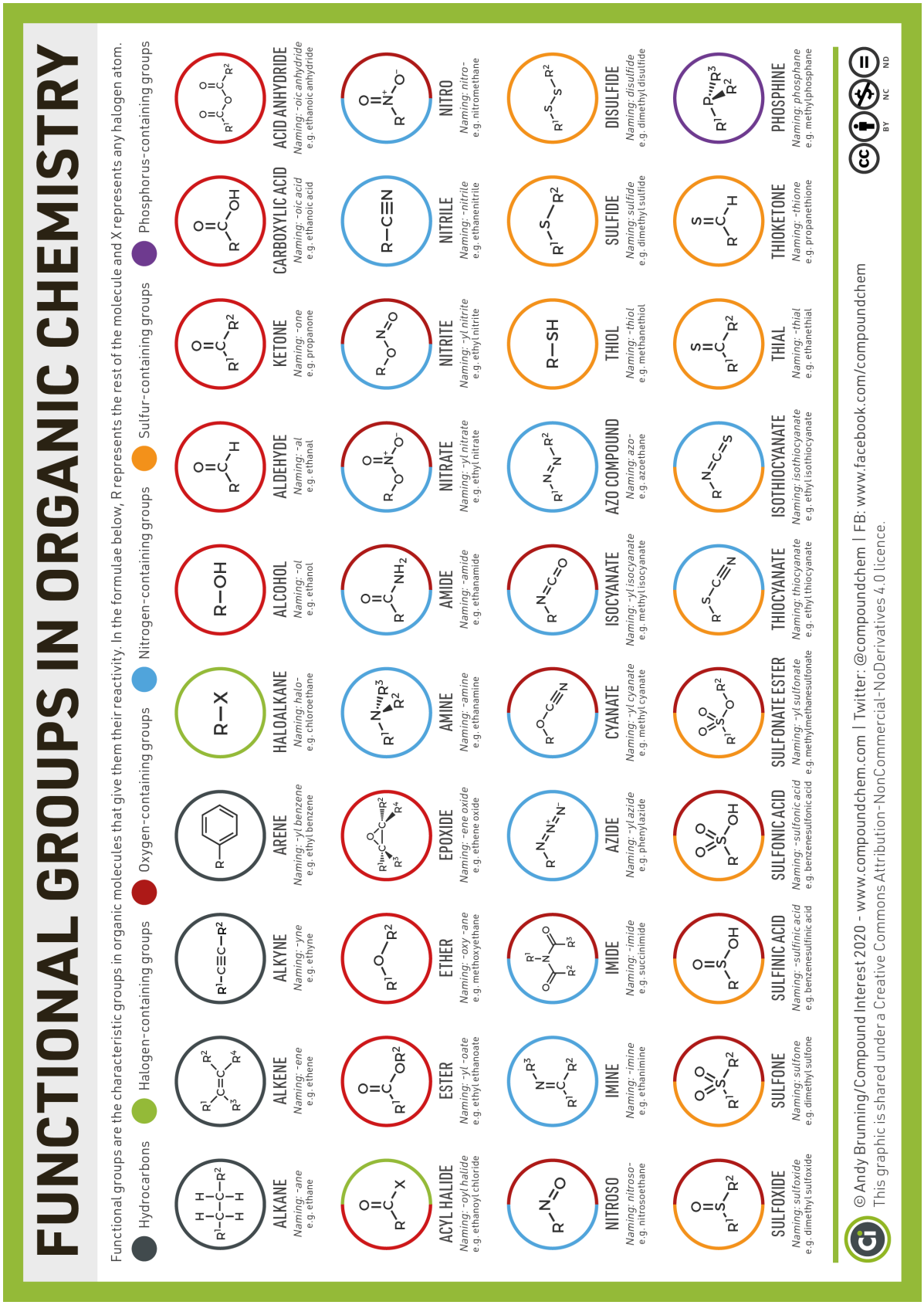
UCARBxx: This terminology is used for unsaturated carbonyls/hydroxycarbonyls, formed for example from isoprene (although one of the main ones, UCARB10, has been ‘‘unlumped’’ into MVK and MACR in the EMEP CRI v2.2 variant).

Related to this, the species HPUCARB12 in CRI v2.2 is an unsaturated carbonyl containing a hydroperoxy group. Again, it is required for representation of the Peeters (LIM) mechanism, and is representative of the species C5HPALD1 and C5HPALD2 in MCM v3.3.1.

UDCARBxx: This terminology is used for unsaturated dicarbonyls, formed from aromatics.

TNCARBxx and TXCARBxx: This terminology is used for carbonyl compounds, formed from monoterpenes with endocyclic and exocyclic double bonds, respectively.

C.2 Functional Groups



Source:?

Appendix D

Chapter Keywords

This section uses the Term Frequency Inverse Document Frequency to determine the keywords of each chapter - a techneque which has been described in **??** and **?**. Text size corresponds to the importance of each word.

D.1 Introduction

AIR SPECIES POLLUTION ATMOSPHERE EARTH ATMOSPHERIC MODEL CHEMISTRY EQUATION DT CHEMICAL OZONE REACTION MECHANISM CLIMATE OXYGEN NUMERICAL EQN DEVELOPMENT HUMAN CYCLE SYSTEM NITROGEN GAS AGO CONCENTRATIONS GCM TIMESCALES ALSO POSSIBLE EVENTUALLY HUMANITY MODELLING DEVCYCLE INDOOR AMBIENT PLANET INT POLLUTANTS HEALTH CHINANOX LIFE EFFECTS ESM RATES MANY CHANGE SOURCE MODELS REPRESENTATION INCREASE RANGE CHANGING COMPLEX UNDERSTANDING GRAPH EXAMPLE YEARS SCIENTIFIC DATA QUALITY BOX GLOBAL OH REACTIONS CELLS SERIES QUICKLY BECOME INTEGRATION TIME PRODUCTION GEOSCHEM NOX THESIS ACT TRAGEDY RADIATION FLAW LIVED DISEASE DRAG HOMO RESPIOTORY WEATHER DEATHS BILLION INDUSTRIAL SE PREMATURE ESMS EXPERIMENTS LIFETIME CHINA ACTUAL LIMITS TRANSFER FLUSHLEFT WHODATA OZONEROLE

D.2 Applying Visual Analytics to the Atmospheric Chemistry Network

GRAPH NODES LAYOUT NODE GRAPHS EDGES SPECIES EDGE LAYOUTS REPRESENTATION USING DISTRIBUTION ALGORITHM CONFLUENT DENSITY DESIGN FORCEDIRECTED NETWORK MECHANISM MERCATOR ANGLE REACTIONS INFORMATION MAY CHEMISTRY MERC STRUCTURE VISUALISATION POSSIBLE FORCE CHEMICAL DATA LINKS MCM APHH SEMANTIC CM DIFFERENT DEGREE ALTHOUGH TSNET OPENORD SHOWS USER ONE REPRESENT BUTANE DRAWING HU ROUTING CROSSING BEZIER ITEMIZE REPRESENTING NUMBER SYSTEM EXAMPLE VISUAL BUNDLING YIFAN ATLAS SHAPE ADDITION AREA PROCESS MANY POINTS BEIJING BEST FORCEATLAS COLOR QUADTREE ORTHOGONAL DIRECTED HIGH PRIMARY SET DESCRIBED ADDITIONAL SELECTION ITEMS REDUCE HOWEVER METHODS NETWORKS SINCE SYNTACTIC GENERAL CURVES GENERATED RESOLUTION STUDY FOUND LARGE ALSO ENERGY DIRECTION CARBON USEFUL SIZE

D.3 Computational Learning, Visualisation and Clustering:

SPECIES PCA DATA CLUSTERS DATASET TSNE SMILES STRUCTURE ALGORITHM GRAPH USING VEC METHODS RANDOM NODE POINTS GROUPS ALGORITHMS DR NUMBER REDUCTION CLUSTERING VERB CLUSTER DIMENSIONALITY AUTOENCODER DIFFERENT SILHOUETTE EQUATION FUNCTIONAL LINEAR DIMENSIONS SET MATRIX VECTOR INPUT ORIGINAL FEATURES COLOURS DISTRIBUTION TABLES SUBIMPORT QUANTUM PRINCIPAL PROBABILITY ACTIVATION TEX MAY POSSIBLE REPRESENTING SINCE MCM METHOD COMPONENT ST FINGERPRINTS RESULT TWO INPUTS FEATURE DESCRIBED STRING MOLECULAR ALTHOUGH CONSTRUCTION OUTPUT ONE CHEMICAL NEW DECISION KEYS PARAMETERS BEST COEFFICIENT NETWORK SPACE RANGE EXAMPLE COLOUR CHEMISTRY HOWEVER REDUCED NODES VALUE NONLINEAR RESULTS GRADIENT GAUSSIAN GREEDY SYSTEM EQN FOUR WORKS STEP MAPPING COMPARING DISTANCE SEVERAL PROPERTIES LEARNING