DanEllisThesis - Change Log

Thesis.tex

```
diff --git a/thesis.tex b/thesis.tex

index 992121e..f497259 100644

--- a/thesis.tex

+++ b/thesis.tex
```

-37,6 +37,9

```
\bibliography{bibtex}

+\urlstyle{same} % do not use typewriter font for urls

%\pagenumbering{roman} http://www.markschenk.com/tensegrity/latexexplanation.html

%A4 (210 mm x 297 mm) https://tex.stackexchange.com/questions/20538/what-is-the-right-order-when-using-frontmatter-tableofcontents-mainmatter

%\addtolength{\textwidth}{12mm}%210
```

-272,9 +275,11

```
\chapter*{Abstract}

\parbox{.75\textwidth}{

-Atmospheric chemistry mechanisms play a pivotal role in our understanding of societal problems such as air pollution, climate change and stratospheric ozone loss. This thesis explores the benefits of representing these mechanisms in terms of a mathematic graph (or network) which connects species (nodes) through reactions (edges). Using the Master Chemical Mechanism run using the Dynamically Simple Box Model of Atmospheric Chemical Complexity we run simulations under a number of different representative scenarios and use graph theory and machine learning to visualise, understand and analyse the underlying chemical processes in the atmosphere.\

+Atmospheric chemistry mechanisms play a pivotal role in our understanding of societal problems such as air pollution, climate change and stratospheric ozone loss. This thesis explores the benefits of representing these mechanisms in terms of a mathematic graph (or network) which connects species (nodes) through reactions (edges). We use the Dynamically Simple Model of Atmospheric Chemical Complexity and the Master Chemical Mechanism to explore the a number of real world senarios - using graph theory and machine learning to visualise, understand and analyse the underlying chemistry of the lower atmosphere.\

-Chapter one discusses the use of various methods in the presentation of complex datasets. Chapter two applies the sociograph framework to atmospheric mechanisms and determines the best way in which to present these. Chapter three takes a more mathematical approach, comparing the results of graph centrality metrics applied to model simulation resuts against more traditional diagnostic methods. The use of graph theory is continued in Chapter four, where graph clustering and natural language processing is used to identify pairs of nodes with similar patterns. Finally Chapter five ventures into the field of chemical informatics, and looks at the use of different representations of species structure within machine learning models (PCA, t-SNE and AutoEncoders) with an aim to merging the content of this thesis into a Graph Convoluted Neural Network in future work.\

+We begin by exploring different visualisation techniques to depict chemistry within the atmosphere. It is found that the sociograph framework provides the most (visually) intuitive delineation of the species and their reactions. For large, complex systems, this type of qseudo-qualitative analysis has its limitations - physical and cognitive. Instead, the relationships between species in the network are quantified using graph centrality metrics and then compared against well-established methods such as the jacobian and rate of production analysis. Further development of graph theory allows us to couple natural language processing, network decomposition, and clustering to identify species with similar lifetimes, reaction styles, or temporal profiles. \

+Having explored aspects of mechanism analysis, visualisation and reduction, we examine how varying representations of species structure can affect the patterns highlighted by unsupervised machine learning models. This is done by visualising them in 2D space and serves as a precursor to potential future work involving Graph Convoluted Neural Networks - thus consolidating the contents
```

of this thesis.\

Ultimately it is found that using a graph-theory approach can prove highly beneficial in the understanding and explanation of chemical mechanisms, but should not (as of yet) be used in substitution of existing investigation and reduction methods.\

-309,6 +314,8

```
\tableofcontents

+\newpage

+\include{./glossary}

\listoffigures

\listoftables

\newpage
```

-318,27 +325,27

```
% Introduction 0

\include{./0_intro}

% Ch1 1

\include{./1_visual}

% Chapter 2

\include{./2_graphs}

+% %

+% % Chapter 3

+% \include{./3_centrality}

-% Chapter 3

-\include{./3_centrality}

-% Chapter 4 - done

- \include{./4_lumping}

-% Chapter 5

- \include{./5_DR}

+% %

+% % Chapter 4 - done

+% \include{./4_lumping}

+% % Chapter 5

+% \include{./5_DR}

+% %

+% %
```

```
% Conclusion

\include{./6_conclusion}

\cleardoublepage\makeatletter@openrightfalse\makeatother

\begin{appendices}
```

-357,7 +364,7

```
% \bibliographystyle{apalike}

% \bibliography{bibtex}

-%% \bibliographystyle{unsrt}

+% \bibliographystyle{unsrt}
```

0_intro.tex

```
diff --git a/intro/combigned.tex b/intro/combigned.tex

index 26376fe..c2b2316 100644

--- a/intro/combigned.tex

+++ b/intro/combigned.tex
```

-20,16 +20,16

```
A change of diet \citep{diet} soon addressed this energy imbalance, provisioning and sharing (cooperative breeding) and tool-assisted processing such as cooking \citep{cooking} - the first known case of anthropogenic indoor air pollution. The increase of cerebral power eventually led to the agricultural revolution\footnote{Domestication of plants and animals.} (12,000 years ago) and the scientific revolution\footnote{ humankind admits ignorance and gain unprecedented control} (500 years ago), \citep{sapiens}.

-As technology improved, so did the anthropogenic emissions to the atmosphere. With 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 with the ancient greeks \citep{skeptical} and Romans \citep{roman} - where it was reported that Rome had a `stink of soot and heavy air'. Similarly, in 1285 the smell of burning jet drove the Queen of England to leave Nottingham and 22 years later King Edward released the first air pollution act \citep{coal1}.

+ Air pollution and climate have always been a concern for the human race. Such disquietude was first documented 6000 years ago with the ancient greeks (lead in the air) \citep{skeptical} and the Romans (Rome was reported to have a `stink of soot and heavy air') \citep{roman}. In 1285 the smell of burning jet\footnote{The lowest rank of coal and very common at the time.} drove the Queen of England to leave Nottingham and 22 years later King Edward released the first air pollution act \citep{coal1}. In the 18th century the United Kingdom entered the Industrial age, here combustion was used to power machines and replace hand tools with mechanical ones. With this started the age of technology and automation - a process requiring energy, and thus increasing emissions to the atmosphere. In the present day technology is ever increasing in efficiency - however the rate of this is not yet suficient to mitigate any damage already caused.

\section{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 (\autoref{sec:airq}), and then relate this to the different roles of ozone (\autoref{sec:ozonerole}), concluding on changing climate and radiative forcing, for with OH plays a vital role (\autoref{sec:climatechange}).

+The atmosphere constitutes an integral part of the Earth system. It is responsible for shielding the planatary surface 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 (\autoref{sec:airq}), and then relate this to the different roles of ozone (\autoref{sec:ozonerole}), concluding on changing climate and radiative forcing, for with OH plays a vital role (\autoref{sec:climatechange}).
```

\subsection{Air Quality - It Is The Air We Breathe}\label{sec:airq}

-The atmosphere consists mainly of Nitrogen and Oxygen (forming 99% of its total mass), as well as a vast range of other species \citep{ac}. Human beings rely on oxygen to convert sugars and 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 (\ch{O3}), nitrogen (\ch{NO2}) and sulphur (\ch{SO2}) dioxides can cause respiratory problems, heart disease, strokes, cancer and chronic obstructive pulmonary disease \cite{who}. Over 80% of people who live in urban environmets\footnote{Which measure the levels of air pollution.} are exposed to poor air quality levels exceeding the recommended 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\footnote{A similar number can also be attributed to indoor air pollution - which also falls under the umbrella term of Air-Quality.} (\autoref{fig:who}).

+The atmosphere consists mainly of nitrogen (\ch{N2}) and oxygen (\ch{O2})\footnote{These form 99% of its dry-air total mass}, in addition to a vast range of other species \citep{ac}. Human beings rely on oxygen to convert sugars and 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), ozone (\ch{O3}), nitrogen dioxide (\ch{NO2}) and sulphur (\ch{SO2}) dioxide can cause respiratory problems, heart disease, strokes, cancer and chronic obstructive pulmonary disease \cite{who}. Over 80% of people who live in urban environmets\footnote{Which measure the levels of air pollution.} are exposed to poor air quality levels exceeding the recommended 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\footnote{A similar number can also be attributed to indoor air pollution - which also falls under the umbrella term of Air-Quality.} (\autoref{fig:who}).

\begin{figure}[H]

\centering

-39,7 +39,7

\end{figure}

\subsection{Stratospheric Ozone - The Protective Barrier}\label{sec:ozonerole}

-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 \citep{ozonehole}\footnote{Here the chlorine attacks the double bond and 'steals' an oxygen atom from the \ch{O3} molecule.}. This resulted in an increase in UV-B radiation, and in consequence skin cancers, immune suppression and disorders of the eye \citep{o3damage}. However, since their ban in the Montreal Protocol, the atmospheric hole in the ozone has recently recovered to levels similar to its discovery 35 years ago \citep{ozonerepair}.

+Ozone plays a vital role in the stratosphere. This was seen in the 1980s where the use of Cloro Fluro Carbon (CFC) aerosols resulted in the thinning of the atmospheric ozone \citep{ozonehole}\footnote{Here the chlorine attacks the double bond and 'steals' an oxygen atom from the \ch{O3} molecule.}. This resulted in an increase in UV-B radiation, and in consequence skin cancers, immune suppression and disorders of the eye \citep{o3damage}. Due to this, the Montreal Protocol on Substances that Deplete the Ozone Layer was put into place to reduce the adverse effects experienced by humans and the Earths surface \citep{montreal}. As part of this, CFCs are still being phased out resulting in a gradual decrease in the damage of the ozone hole.

\subsection{Changing Climate} \label{sec:climatechange}

-230,7 +230,7

\subsection{The Dynamically Simple Model Of Atmospheric Chemical Complexity}

-Within this thesis, the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) shall be used to run model simulations. This a simple box model designed for the comparison of a range of gas-phase chemical schemes under different conditions \citep{dsmacc}.

+Within this thesis, the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) was used to run model simulations. This a simple box model designed for the comparison of a range of gas-phase chemical schemes under different conditions \citep{dsmacc}.

The DSMACC model uses the Kinetic PreProcessor (KPP) to convert a chemical mechanism into the set of ordinary differential equations which can be solved using a suite of FORTRAN numerical integrators it provides \citep{kpp}. The Tropospheric and Ultraviolet (TUV) model from \cite{tuv} is used to calculate the strengths of different photolysis reactions for the mechanism. These are determined at the start of a simulation and then predicted using cubic splines \citep{dsmaccgit}. This is the model setup that will be used to propagate the chemistry forwards in time using the Rosebrock integrator.

\section{Thesis Layout}

1_visual.tex

diff --git a/visintro/combigned.tex b/visintro/combigned.tex

index 8026d7b..71c389f 100644

--- a/visintro/combigned.tex

+++ b/visintro/combigned.tex

-8,10 +8,10

In nature, animals rely on the propagation of DNA to encode information critical to their survival. Examples of these are found in hives (where an insects role is defined by its genetic composition), or in Oscines (songbirds) which have an inherent predisposition to learn species-specific songs, \citep{modelingpythonbees,genomics,birds,birdsongs,sapiens}. For humans; however, this process is highly impractical due to the vast and varied nature of the information need to process. Instead, we have developed a predisposition to learning language at an early age. In essence, a skill allowing for the effective communication of ideas, conditions and dangers between a large number of people\footnote{Several studies, exploring the ratio of the neocortex to the rest of the brain, suggest that the number of relationships a human can successfully monitor is limited to ~150. It is suggested that ideas of gossip and common metaphysical beliefs are the reason for this \citep{sapiens,neo,gossip}. This limit is still seen in social networks today \citep{social}.}.

-The downside to this is that communicatory patterns are limited to only the people they have been taught to. Here problems of differing language and dialect significantly reduce the amount of information which may be passed between groups/tribes. Such issues were quickly overcome through the use of visualisation in the form of pictographs (cave paintings - e.g. \autoref{cave}). Such methods complement our ability to both detect shapes and spot patterns within nature\footnote{It has been found that 10,000 year-old pictographs show hints of a shared cultural background between spatially different groups of humans \citep{cave}.} as well as providing an intuitive method of communication between separate groups.

+The downside to learnt behaviours, such as language, is that communicatory patterns are limited to only the people they have been taught to. Here problems of differing language and dialect significantly reduce the amount of information which may be passed between groups/tribes. Such issues were quickly overcome through the use of visualisation in the form of pictographs (cave paintings - e.g. \autoref{cave}). Such methods complement our ability to both detect shapes and spot patterns within nature\footnote{It has been found that 10,000 year-old pictographs show hints of a shared cultural background between spatially different groups of humans \citep{cave}.} as well as providing an intuitive method of communication between separate groups.

-As communities continue to increase in size, problems of accounting and resource management start to emerge. Here the ability to store large amounts of data had not been previously required by a hunter-gatherer species. This problem was again solved by the samaritans ($\tilde 3500$BC) with the creation of writing - a system for coordinating affairs and storing information external to a humans brain \citep{archaic,beforeCuneiform}. Using this quantities and items are depicted using a system of signs and shapes (cuneiform\footnote{This is often mistaken for hieroglyphics. Although both are forms of logographic script, hieroglyphs are restricted to the ancient Egyptian sociolinguistic context. }) - a practical and intuitive way for us to apply the pattern recognition and analytical parts of our brain while reducing the cognitive load by breaking up the problem into manageable parts.

+As communities continue to increase in size, problems of accounting and resource management start to emerge. Here the ability to store large amounts of data had not been previously required by a hunter-gatherer species. This problem was again solved by the Samaritans ($\tilde 3500$BC) with the creation of writing - a system for coordinating affairs and storing information external to a humans brain \citep{archaic,beforeCuneiform}. Using this quantities and items are depicted using a system of signs and shapes (cuneiform\footnote{This is often mistaken for hieroglyphics. Although both are forms of logographic script, hieroglyphs are restricted to the ancient Egyptian sociolinguistic context. }) - a practical and intuitive way for us to apply the pattern recognition and analytical parts of our brain while reducing the cognitive load by breaking up the problem into manageable parts.

Throughout history, we have continued to apply this system of intertwining data information with visual artefacts to enable people to cope with the complexities of the information provided, \citep{tufte}. It is for this reason that visualisation can be used as a means of enhancing the reader's ability to understand the large-scale complexities of scientific data.

-86,7 +86,7

\caption{\textbf{Two tree-inspired visualisations. }\

(a) shows the decisions made on a single decision tree within a random Forrest. Hear each branch split corresponds to a decision and the node/leaf colour represents the category of the decision. Stronger and more important decisions correspond to larger leaves and thicker branches. \

- (b) shows a radial plot in the shape of a tree trunk. Here time is shown radiating outwards from the centre. This allows us to spot any changes in evens - much like the rings of a tree can be used to identify when natural disasters (such as tsunamis or avalanches) have struck them. This specific visualisation shows the net flux of species from a chemical simulation.

+ (b) shows a radial plot in the shape of a tree trunk. Here time is shown radiating outwards from the centre. This allows us to spot any changes in events - much like the rings of a tree can be used to identify when natural disasters (such as tsunamis or avalanches) have struck them. This specific visualisation shows the net flux of species from a chemical simulation.

These are coloured from low fluxes (blue) to high fluxes (red). The abrupt changes here show the diurnal cycle where photochemical reactions stop and then start up again. }

\label{fig:trees}

\end{figure}

-408,17 +408,61

\subsubsection{The Traditional Network Graph}\label{sec:tradnetconc}

-Finally, we have the traditional network representation in the form of a mathematical graph. Here species are represented as nodes (circles) and reactions as the links (lines) between them. This analogy has its roots in social representation and can be described using the metaphor of people holding hands - a concept familiar to most people. Graph representations allow for an overview of the structural relationships within the MCM network, and even to compare it against other reduced mechanisms. \autoref{fig:graphc1} shows the comparison of the MCM against the reduced Common Representative Intermediates (CRI) \citep{cri} mechanism. In fixing common species (generally the primary emitted VOCs) between both mechanisms, we can use the graph as a fingerprint to compare changes in network structure. The CRI mechanism reduces the number of species within the MCM based on their ozone-forming potential. This is seen within the enclosed polygons in \autoref{fig:graphc1}, where the messy structure of the MCM (top) is greatly reduced, forming clusters of lumped species with similar ozone-forming potential (bottom). This form of representation is the most intuitive and commonly used sociograph, and therefore shall further be explored in \autoref{ch2}.

+Finally, we have the traditional network representation in the form of a mathematical graph. Here species are represented as nodes (circles) and reactions as the links (lines) between them. This analogy has its roots in social representation and can be described using the metaphor of people holding hands - a concept familiar to most people. Graph representations allow for an overview of the structural relationships within the MCM network, and even to compare it against other reduced mechanisms, \autoref{fig:graphc1}

+Here two

+ shows the comparison of the MCM against the reduced Common Representative Intermediates (CRI) \citep{cri} mechanism. In fixing common species (generally the primary emitted VOCs) between both mechanisms, we can use the graph as a fingerprint to compare changes in network structure. The CRI mechanism reduces the number of species within the MCM based on their ozone-forming potential. This is seen within the enclosed polygons in \autoref{fig:graphc1}, where the messy structure of the MCM (top) is greatly reduced, forming clusters of lumped species with similar ozone-forming potential (bottom). This form of representation is the most intuitive and commonly used sociograph, and therefore shall further be explored in \autoref{ch2}.

\begin{figure}[H]

\centering

- \includegraphics[width=1.1\textwidth]{fingerprintposter.pdf}

- \caption{\textbf{Two node-link graphs comparing the MCM and the reduced CRI mechanism for an n-heptane subset.} The top graph shows the MCM v3.2 subset for n-heptane. Here features of the chemistry (e.g. anthropogenic and biogenic emitted species) are seen within the graph structure. The bottom graph shows the reduced Common Representative Intermediates (CRI) v2.1. Species that exist in both mechanisms are fixed, allowing us to compare the differences in structure between both. Node colours represent modules of clusters of the chemistry and hold no further meaning for this section. }

+ \includegraphics[width=1\textwidth]{poster.png}

+ \caption{\textbf{Comparing a range of MCM and CRI mechanims using their graph shape and structure.} Source: \cite{mcmblue}}

```
\label{fig:graphc1}
\end{figure}
+\begin{figure}[H]
+ \centering
+ \begin{subfigure}[b]{.49\textwidth}
+ \centering \includegraphics[width=\textwidth]{m31_m32.png}
+ \caption{MCM v3.1 vs MCM v3.2}
+ \label{fig:m1to2}
+ \end{subfigure}
+ \begin{subfigure}[b]{.49\textwidth}
+ \centering \includegraphics[width=\textwidth]{m32_m33.png}
+ \caption{MCM v3.2 vs MCM v3.3.1}
+ \label{fig:m2to3}
+ \end{subfigure}
+ \begin{subfigure}[b]{.49\textwidth}
+ \centering \includegraphics[width=\textwidth]{cr1_cr5.png}
+ \caption{CRI v2.0(r1) vs CRI v2.0(r5)}
+ \label{fig:cr1tocr5}
+ \end{subfigure}
+ \begin{subfigure}[b]{.49\textwidth}
+ \centering \includegraphics[width=\textwidth]{cr1_m32.png}
+ \caption{MCM v3.2 vs CRI v2.0(r1)}
+ \label{fig:ts}
+ \end{subfigure}
+ \hfill
+ \caption{ \textbf{Voronoi cells of each node from the graph layout - used to identify changes in mechanisms.} A difference plot between the different graphs in \autoref{fig:graphc1}. These use colours to show us species that are added or taken away between different versions. Subplots (a) and (b) show the increasesin mechanism size of the MCM whilst (c) and show the reduction from MCM v3.2 to CRI v2.0(r1), and followed by the fith reduction to CRI v2.0(r5).
+ Figure colouring: purple cells only exist within the first mechanism, pink only exist within the second, and blue are present in both. Source: \cite{mcmblue}}
+ \label{fig:mcmchange}
+\end{figure}
```

2_graphs.tex

```
diff --git a/visanalytics/combigned.tex b/visanalytics/combigned.tex
index f6c9ce1..15cd3f8 100644
```

--- a/visanalytics/combigned.tex

+++ b/visanalytics/combigned.tex

-24,7 +24,7

\includegraphics[width=\textwidth]{C141CO33d.png}

\caption{3D}

\end{subfigure}

- \caption{\textbf{The molecule \ce{C141CO3} shown in both 2D and 3D node-link structures.} This is a the result of a series of inorganic species reactions and a desocciation from BCARY - the only sesqueterpine in the MCM. 3D visualisation by \citep{mol3d}. }

+ \caption{\textbf{The molecule C141CO3 (MCM name) shown in both 2D and 3D node-link structures.} This is a the result of a series of inorganic species reactions and a desocciation from BCARY - the only sesqueterpine in the MCM. 3D visualisation by \citep{mol3d}. }

\label{fig:mol}

\end{figure}

-53,7 +53,7

\centering

\includegraphics[width=\textwidth]{figures_c1/butane.png}

- \caption{\textbf{A systematic representation of the degregation of butane.} Using this we are able to see the process \ce{C4H10} undergoes before its ultimate demise as carbon monoxide and water. Source: \citep{butane} }

+ \caption{\textbf{A systematic representation of the degregation of butane.} Using this we are able to see the process \ce{C4H10} undergoes before its ultimate demise as carbon dioxide and water. Source: \citep{butane} }

\label{fig:butane}

\end{figure}

-68,7 +68,7

Historically it is shown that the graph format has proven to be an efficient means of understanding the reactions within a mechanism. Traditionally these are constructed manually, with the designer making a series of choices on how best to place, and simplify the chemistry based on their application. As our understanding of chemistry improves and we have started to progress into automated and semi-automated mechanism construction. This makes the construction of mechanisms with tens of millions of species and billions of reaction possible (\citep{protocol}) and is the point where the manual design/simplification of reaction networks becomes infeasible.

-Today automatic graph layouts allow us to generate multivariate and complex graphs quickly \citep{ch3} -This means that, much like in the construction of a mechanism, we can rely on computer-aided design to generate a directed graph representation of the chemistry. \cite{sciamerican} states that "The beauty of a good information graphic is that it can tell a whole story in a single unit of visual content". This is particularly true for the use of directed graphs in chemistry where we can compare different mechanism structures.

+Today automatic graph layouts allow us to generate multivariate and complex graphs quickly \citep{ch3}. This means that, much like in the construction of a mechanism, we can rely on computer-aided design to generate a directed graph representation of the chemistry. \cite{sciamerican} states that "The beauty of a good information graphic is that it can tell a whole story in a single unit of visual content". This is particularly true for the use of directed graphs in chemistry where we can compare different mechanism structures.

However, several problems emerge from the complete automation of a task. Firstly real-world data very rarely reacts how it is expected to. Here networks of high edge density often obfuscate the graph data and produce what is only described as a birds nest',hairball' or ‘ball of yarn' within the literature \citep{ch7}. Although such problems can be shown as moments of turbulence, they encourage a greater understanding of the graphic design process and can catalyze to merge unique ideas into an effective visualisation \citep{goodideas} - much like the composite metaphors in

\autoref{ch1}.

-77,7 +77,7

-\section{Graph Syntatics}\label{syntatic}

+\section{Graph Syntactics}\label{syntatic}

Syntactic representation considers how best to distribute information on a page for maximum impact. This can be seen between the force-directed graph (top) and geographical location (bottom) layouts in \autoref{fig:worldmap}. Although the geographical layout gives a more accurate representation of the distances between unconnected nodes (airports), a force-directed graph provides greater insight into the relationships (flights) between each airport. This highlights the importance of choosing a suitable syntactic representation to highlight the features of interest. The remainder of this section discusses the syntactic choices required for the visualisation of a complex chemical mechanism.

-339,7 +339,7

\subsubsection{Distribution Of Primary Emitted VOCs}

-Within the construction of an atmospheric chemical mechanism, a chemist first begins with a primary emitted species. This is then broken down to produce other species, depending on its structure and functional groups (\autoref{fig:protocol}). This process suggests that in constructing a network from such a mechanism, this structure will be prominent. Knowledge dictates that a chemical graph should start from a large emitted species, and aim towards carbon monoxide (and ultimately \ch{co2} although this is not included in the MCM). To show such a structure, we expect any primary emitted species to be evenly distributed and the chemistry to tend towards the location of CO (the centre). In searching for a layout that satisfies this requirement, the tsNET graph (\autoref{dfig:ts}) is found to be the best, followed by the OpenOrd and ForceAtlas2. Yifan Hu (\autoref{fig:yfan}) and Mercator (\autoref{fig:merc}b) both contain areas where many of the primary emitted (orange) species are grouped and are therefore unsuitable for the representation of the MCM structure.

+Within the construction of an atmospheric chemical mechanism, a chemist first begins with a primary emitted species. This is then broken down to produce other species, depending on its structure and functional groups (\autoref{fig:protocol}). This process suggests that in constructing a network from such a mechanism, this structure will be prominent. Knowledge dictates that a chemical graph should start from a large emitted species, and aim towards carbon monoxide (and ultimately \ch{co2} although this is not included in the MCM). To show such a structure, we expect any primary emitted species to be evenly distributed and the chemistry to tend towards the location of CO (the centre). In searching for a layout that satisfies this requirement, the tsNET graph (\autoref{fig:ts}) is found to be the best, followed by the OpenOrd and ForceAtlas2. Yifan Hu (\autoref{fig:yfan}) and Mercator (\autoref{fig:merc}b) both contain areas where many of the primary emitted (orange) species are grouped and are therefore unsuitable for the representation of the MCM structure.

\subsubsection{Calculation Of Spatial Clustering}\label{sec:nodedensitya}

-791,7 +791,7

\includegraphics[width=\textwidth]{figures_c1/tap3/ch4_weighted_s1-eps-converted-to.pdf}

\caption{Connected weighted (flux)}

\end{subfigure}

- \caption{\textbf{A weighted and unweighted force diagram of the methane mechanism.} Here it is seen that upon weighting, edges with a faster flux (pink) are drawn closer than those of a weaker one (blue).}

+ \caption{\textbf{A weighted and unweighted force diagram of the methane mechanism.} Here it is seen that upon weighting, edges with a larger flux (pink) are drawn closer than those of a weaker one (blue).}

\label{fig:resmeth}

\end{figure}

3_centrality.tex

diff --git a/model_diagnostics/combigned.tex b/model_diagnostics/combigned.tex

```
index c021b18..72ec6b0 100644

--- a/model_diagnostics/combigned.tex

+++ b/model_diagnostics/combigned.tex
```

-1,6 +1,6

```
\section{Introduction}

-The node-link (ball-stick) style structure has long been used to represent real-world relationships between items (\autoref{sec:chemgraph}). Such a structure is complementary to our cognitive disposition towards pattern recognition, and it is for this reason that the node-link visualisation format has been used for anything ranging from transportation maps \citep{beck} to the differentiation of ancestorial lineages of the human race (\autoref{fig:skulls}). However, the abundance and complexity of real-world data often present us with difficulties in manually representing it in a useful form. In \autoref{syntatic}, it is suggested this may be overcome with the use of computational analysis and automated visualisation tools. Such methods usually require a level of data manipulation to transform the data into a machine parseable form.

+The node-link (ball-stick) style structure has long been used to represent real-world relationships between items (\autoref{sec:chemgraph}). Such a structure is complementary to our cognitive disposition towards pattern recognition, and it is for this reason that the node-link visualisation format has been used for anything ranging from transportation maps \citep{beck} to the differentiation of ancestorial lineages of the human race (\autoref{fig:skulls}). However, the abundance and complexity of real-world data often present us with difficulties in manually representing it in a useful form. In \autoref{syntatic}, it was suggested this may be overcome with the use of computational analysis and automated visualisation tools. Such methods usually require a level of data manipulation to transform the data into a machine parseable form.

\begin{figure}[H]

\centering
```

-110,7 +110,7

```
\textbf{A note on unintentional filtering}\

\textit{

-The script used for web scraping extracts author names directly from the google scholar page, and no the articles themselves. This means some author names can be omitted and replaced by ellipses - producing an inaccurate graph. Therefore the results in this section are not explicit, but rather a demonstration of graph theory on a real-world dataset.

+The script used for web scraping extracts author names directly from the google scholar page, and not the articles themselves. This means some author names can be omitted and replaced by ellipses - producing an inaccurate graph. Therefore the results in this section are not explicit, but rather a demonstration of graph theory on a real-world dataset.
```

-218,13 +218,13

```
\input{tables/Out-Degree_Citation.tex}

-\subsection{Closness Centrality}\label{sec:closeness}

-Often within a network, we are interested in how easy it is to to get information from one node to every other node. This is what the closeness centrality tells us. To calculate a nodes closeness, we begin by taking the reciprocal sum of all the Dijkstra paths (The shortest available path.) to every other node \citep{closeness-book,closeness}.

+\subsection{Closeness Centrality}\label{sec:closeness}

+Often within a network, we are interested in how easy it is to to get information from one node to every other node. This is what the closeness centrality tells us. To calculate a nodes closeness, we begin by taking the reciprocal sum of all the Dijkstra paths\footnote{The shortest available path.} to every other node \citep{closeness-book,closeness}.

This gives a representation of how far information from a particular person (node) will need to travel to reach every other node. Such a metric has applications in intelligence gathering,
```

telecommunications and word importance within key-phrase extraction \citep{terror,examples_centrality,phrase}.

\begin{quote}

\textit{

-\textbf{Example analogy:} If we take the UK rail network as an example, York station will have a high closeness value as it is well connected and central in location. This means it is easy to reach every other location when compared to other stations.\autoref{appendix:rail}

+\textbf{Example analogy:} If we take the UK rail network as an example, York station will have a high closeness value as it is well connected and central in location. This means it is easy to reach every other location when compared to other stations, \autoref{appendix:rail}

\end{quote}

-250,7 +250,7

\end{quote}

-Authors with a high betweenness in \autoref{fig:betauth} are seen to lie along the joints between clusters. Here we can imagine that removing Li, Griffin or Liu can disrupt the overall flow of collaboration, potentially isolating the work of the Max Planck from that of everyone else. Similarly, Jenkin and Pilling can be seen as holding much of the Leeds cluster together. In removing them from the network (if for example, the refused to collaborate) it is possible to see how many of groups within the Leeds environment may not have worked together, with the cluster potentially separating into several smaller groups. Finally, we see Saunders (Australia), who served to introduce the MCM to the Chinese atmospheric community. In removing her from the network, it can be seen that much of the collaboration which exists would have been significantly less likely.

+Authors with a high betweenness in \autoref{fig:betauth} are seen to lie along the joints between clusters. Here we can imagine that removing Li, Griffin or Liu can disrupt the overall flow of collaboration, potentially isolating the work of the Max Planck for Chemistry from that of everyone else. Similarly, Jenkin and Pilling can be seen as holding much of the Leeds cluster together. In removing them from the network (if for example, the refused to collaborate) it is possible to see how many of groups within the Leeds environment may not have worked together, with the cluster potentially separating into several smaller groups. Finally, we see that Saunders (Australia) is highlightes as an important node - an action which can be attibuted her introducing the Chinese atmospheric community to the MCM. In removing her from the network, it can be seen that much of the collaboration which exists would have been significantly less likely.

\begin{figure}[H]

\centering

-356,7 +356,7

This is repeated until a pre-defined tolerance, ϵ is reached. For best results, this can be set to just under the numerical precision of the programming language/hardware.

-For smaller systems, it is possible to use the LAPACK \citep{lapack} library, as used by \cite{numpy}. For a vast network, however, the computation of a $n \times n $ matrix can be very memory inefficient for small machines. It is then possible to apply the methods as described above using a sparse matrix on per-node bases as can be seen within the Python's SciPy implementation of the Networkx source code \citep{scipy,networkx}.

+For smaller systems, it is possible to use the LAPACK \citep{lapack} library, as used by \cite{numpy}. For a vast network, however, the computation of a $n \times n $ matrix can be very memory inefficient for small machines. It is then possible to apply the methods as described above using a sparse matrix on per-node bases as can be seen within the Python SciPy implementation of the Networkx source code \citep{scipy,networkx}.

\subsubsection{Prediction}\label{sec:applypr}

As the PageRank algorithm loos at how quantities flow' within a network, it can be used to identify not only the bottlenecks (betweenness centrality) but also any nodes which are connected well within the network. As the flows between a node are somewhat governed by the number of links it contains, the PageRank algorithms tend to correlate, but not dependence, on the betweenness of a node. \autoref{fig:pagerankauth} uses the PageRank algorithm to identify important authors within eachcluster' or research group. Due to its propagating nature, authors connected to these important nodes are often also of greater importance. An application of this can again be the determination of how to best spread new results or information with the least number of people. \textit{Note: if we only had one person we would probably use the node with the highest closeness centrality.}

-366,7 +366,7

\includegraphics[width=.8\textwidth]{figures_c3/pagerankauthor.png}

\input{tables/pagerank_Author.tex}

- \caption{ \textbf{Page Rank centrality within the co-Author network}. Node size and colour represent the ranking of each node from the page rank algorithm. Bigger,lighter nodes are more important.}

+ \caption{ \textbf{Page Rank centrality within the co-Author network}. Node size and colour represent the ranking of each node from the page rank algorithm. Larger, lighter coloured nodes are more important.}

\label{fig:pagerankauth}

\end{figure}

-379,7 +379,7

\section{Classifying The Master Chemical Mechanism Network}\label{sec:globalclass}

-Having shown that graph metrics can help the roles of individual nodes within the network, these are now applied to an atmospheric chemical system. Since computational efficiency and resources are often a limiting factor, many applications of the MCM only require a small subset of the entire mechanism. For this reason, it may be of interest to compare these against each other, in an attempt to classify the type of network the MCM chemistry falls under. In this section, we apply graph theory to the entire MCM network to determine its defining characteristics. This is achieved through the analysis of several hundred Monte Carlo selected subsets of the MCM. Each of these is a different combination of the primary emitted VOC's within the MCM v3.3.1.

+Having shown that graph metrics can help the roles of individual nodes within the network, these are now applied to an atmospheric chemical system. Since computational efficiency and resources are often a limiting factor, many applications of the MCM only require a small subset of the entire mechanism. For this reason, it may be of interest to compare these against each other, in an attempt to classify the type of network the MCM chemistry falls under. In this section, we apply graph theory to the entire MCM network to determine its defining characteristics. This is achieved through the analysis of several hundred Monte Carlo selected subsets of the MCM. Each of these is a different combination of the primary emitted VOCs within the MCM v3.3.1.

\subsection{Network Density}\label{sec:netdensity}

Network density is the easiest metric to understand. Visually this can induce complexity and obscure aspects in a graph; mathematically, it can greatly increase the computation time for metrics or algorithms. By definition, we can define network density as a measure of how well connected a node is to every other node. Mathematically it is the ratio of edges against the total number of possible edges for a complete graph\footnote{A complete graph is one where every node is connected to every other node.} of the same size. In chemical terms, we can use this to determine the sparsity of the graph (which has applications on model integrator selection) and give us insights on the chemical structure. In \autoref{fig:density}, higher numbers of species (nodes) results in an overall decrease in the node-edge ratio - its density. This suggests a modular or hierarchical structure, where new species directly react only with a set number of species, and not the entire mechanism. An explanation for this is that the addition of larger species introduce new branches within the chemistry, which then need to be oxidised before they are small enough to react with the species from a different branch. Since these branches are somewhat isolated from the rest of the chemistry, they decrease the network density, even though their addition may increase the amount of chemistry that occurs within it.

-387,7 +387,7

\begin{figure}[H]

\centering

\includegraphics[width=.7\textwidth]{figures_c3/sparcity.png}

- \caption{\textbf{How the MCM graph density scales with number of species.} A figure showing that an creasing the number of species within a mechanism subset results in an increased model sparsity (decreasing density).}

+ \caption{\textbf{How the MCM graph density scales with number of species.} A figure showing that an increasing number of species within a mechanism subset results in an increased model sparsity (decreasing density).}

\label{fig:density}

\end{figure}

-404,7 +404,7

Here C is the average clustering coefficient and L, the shortest path length of the graph. Comparing these with the average shortest path length, L_R, and clustering coefficient C_l (as calculated using an equivalent random and lattice graph) gives the above equation. The output is a result between positive and negative one {-1,1}, where a value of 0 suggests the graph exhibits perfect small world-ness.

-In assessing the network structure of the MCM, a Monte Carlo (random) approach was taken to extract several hundred subsets from the entire mechanism. For each of these, the omega coefficient was calculated and plotted in \autoref{fig:smw}. Here it is seen that subsets with a small number of species (for example those derived only from Methane or Ethane) exhibit a more lattice-style (grid) graph, with the majority of the networks showing a more random network structure \autoref{fig:gstructure}. All the results, however, show a prevalence of small-world features over any of the alternative network structures - they are closer to 0 than 1 or -1. This reflects the idea that large species react locally, forming branches (\autoref{ch2}), before oxidising to smaller species with more reactions. This result is also seen within the Reaxys chemical database \citep{rscgraph}.

+In assessing the network structure of the MCM, a Monte Carlo (random) approach was taken to extract several hundred subsets from the entire mechanism. For each of these, the omega coefficient was calculated and plotted in \autoref{fig:smw}. Here it is seen that subsets with a small number of species (for example those derived only from methane or ethane) exhibit a more lattice-style (grid) graph, with the majority of the networks showing a more random network structure \autoref{fig:gstructure}. All the results, however, show a prevalence of small-world features over any of the alternative network structures - they are closer to 0 than 1 or -1. This reflects the idea that large species react locally, forming branches (\autoref{ch2}), before oxidising to smaller species with more reactions. This result is also seen within the Reaxys chemical database \citep{rscgraph}.

-429,7 +429,7

\label{fig:gstructure}

\end{figure}

-To assess the best distribution for describing the monte carlo subsets of the MCM I use the Kolomogorov-Smirnov statistic \citep{ks} to analyse the goodness of fit of the ω coefficient in \autoref{fig:smw} to a number of distributions. This calculates the maximum distance D between the selected cumelative distribution function $S(x)$ (In our case the Logarithmic, Exponential and Power Law) of the data and the fitted model $P(x)$:

+To assess the best distribution for describing the Monte Carlo subsets of the MCM, the Kolomogorov-Smirnov statistic \citep{ks} was used to analyse the goodness of fit of the ω coefficient in \autoref{fig:smw} to a number of distributions. This calculates the maximum distance D between the selected cumelative distribution function $S(x)$ (In our case the Logarithmic, Exponential and Power Law) of the data and the fitted model $P(x)$:

\begin{equation}

D = \smash{\displaystyle\max_{x \ge x_{min}}} |{S(x) - P(x)}|

-461,7 +461,7

Using the species concentration as a metric, we can map how it changes over time, and how in changing the initial concentrations of a simulation can produce different results. This can be useful for looking at a range of possible scenarios and evaluating the potential outcome after a pre-determined amount of time. An example would be through the use of policy-based simulations to predict changes in air composition over cities.

-Using a simple example from a Methane only subset of the MCM (\autoref{fig:concentration}), it is possible to observe the inverse relationship between \ch{NO2} and \ch{NO} using only their concentration profiles. Here nitrogen monoxide reacts with a \ch{Ro2} species to produce an RO and nitrogen dioxide.

+Using a simple example from a methane only subset of the MCM (\autoref{fig:concentration}), it is possible to observe the inverse relationship between \ch{NO2} and \ch{NO} using only their concentration profiles. Here nitrogen monoxide reacts with a \ch{Ro2} species to produce an RO and nitrogen dioxide.

This then photolyses back to nitrogen oxide, releasing oxygen which may go on to form ozone (\autoref{sec:o3prod}). The latter part of this reaction is dependant on photons and therefore can only occur during daytime (mostly).

\begin{figure}[H]

-474,7 +474,7

\subsubsection{Rate Of Production And Loss}\label{sec:ropa}

-Analysing the concentration-time profiles allows the comparison of how a series of scenarios or runs change concerning their initial conditions and simulation length. Although these can tell us how, and how much, each species changes over time, it does not rank or quantifies the specific reactions to which this may be attributed. Rate of Production Analysis (ROPA)\footnote{and loss} provides a method for establishing the total contribution from each reaction by calculating the change of concentration (concerning time) for the produced species - the instantaneous reaction Flux.

+Analysing the concentration-time profiles allows the comparison of how a series of scenarios or runs change concerning their initial conditions and simulation length. Although these can tell us how, and how much, each species changes over time, it does not rank or quantifies the specific reactions to which this may be attributed. Rate of Production Analysis (ROPA)\footnote{and loss} provides a method for establishing the total contribution from each reaction by calculating the change of concentration (concerning time) for the produced species - the instantaneous reaction flux.

\begin{eqnarray}

r_1 = A + B \overset{\kappa_1}{\xrightarrow{\hspace*{7mm}}} \eta C & & \text{ Reaction 1}\\[15pt]

-547,25 +547,25

Having covered the general definition of a Jacobian matrix and how it is constructed, we can now apply it to the context of mechanism analysis and comprehension. The first analogy that needs to be made is that for the flux is the change of a species concentration in time (the first differential with respect to time, d/dt). If we consider the change in a species concentration as a displacement', we can think of the flux as itsvelocity'.

Similarly, the Jacobian provides us with a description of how the individual flux of a species changes concerning the concentration (or displacement) or another species (the second-order partial differential). This is analogous to the acceleration of the object or particle we first displaced. In using the Jacobian, we have constructed a relational matrix which outlines the effect a nominal change of a species has on all other species - a concept which is the foundation of the connectivity method (a mechanism reduction technique where all but essential species are removed) \citep{connectivity}.

-Since the format of a jacobian is already in the form of a relational matrix, it can easily be converted to a weighted adjacency matrix, and then directly into the graph format. Since it only considers the aggregated influence between species, much of the work that would otherwise be needed to convert a mechanism into a graph format has already been done. To make use of the Jacobian matrix, several extraction algorithms were written for an updated version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) \citep{dsmacc,dsmaccgit}, as discussed in \autoref{ch0}. Here we edit the kinetic pre-processor output, \citep{kpp} to release the values of the Jacobian Matrix and return them at each model timestep for analysis. The process for how this is done is described in \autoref{sec:jacpractical}.

+Since the format of a Jacobian is already in the form of a relational matrix, it can easily be converted to a weighted adjacency matrix, and then directly into the graph format. Since it only considers the aggregated influence between species, much of the work that would otherwise be needed to convert a mechanism into a graph format has already been done. To make use of the Jacobian matrix, several extraction algorithms were written for an updated version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) \citep{dsmacc,dsmaccgit}, as discussed in \autoref{ch0}. Here we edit the kinetic pre-processor output, \citep{kpp} to release the values of the Jacobian Matrix and return them at each model timestep for analysis. The process for how this is done is described in \autoref{sec:jacpractical}.

\subsubsection*{ A Note On Using The Flux Instead Of The Jacobian }

\textit{

-Depending on the model setup or the users' capabilities, extraction of the jacobian matrix for each timestep may not be possible. In many cases, the reaction rates and concentration may still be available, allowing for the calculation of reaction fluxes throughout the simulation. If this is the case, the total flux can be calculated using the method described in \autoref{eqn:ode}. From this, an edge-weighted by a reaction flux can be created from every reactant to each product. This generates a multi-graph (A graph with multiple edges between nodes) which may be simplified by taking the net flux value for all edges between two nodes. \

+Depending on the model setup or the users' capabilities, extraction of the Jacobian matrix for each timestep may not be possible. In many cases, the reaction rates and concentration may still be available, allowing for the calculation of reaction fluxes throughout the simulation. If this is the case, the total flux can be calculated using the method described in \autoref{eqn:ode}. From this, an edge-weighted by a reaction flux can be created from every reactant to each product. This generates a multi-graph (A graph with multiple edges between nodes) which may be simplified by taking the net flux value for all edges between two nodes. \

However, the potential for human/coding error, additional simplification and a non-explicit definition of the contribution of each species make the use of a Jacobian much more efficient in network generation from a chemical mechanism.

\subsection{A Practical Example Using The MCM}\label{sec:jacpractical}

-Taking a single equation from the MCM, we may calculate the jacobian relationships between species and convert them into a graph. A randomly chosen ethane reaction (\autoref{eqn:line}) from a simple mechanism was chosen. It must be noted that in general, it is unusual in the MCM that alkyl radicals react rapidly and extremely well with \ce{O2} to from stabilised peroxy radicals, \citep{mcmorigin}. In general, the reaction would consist of the following two steps:

+Taking a single equation from the MCM, we may calculate the Jacobian relationships between species and convert them into a graph. A randomly chosen ethane reaction (\autoref{eqn:line}) from a simple mechanism was chosen. It must be noted that in general, it is unusual in the MCM that alkyl radicals react rapidly and extremely well with \ce{O2} to from stabilised peroxy radicals, \citep{mcmorigin}. In general, the reaction would consist of the following two steps:

\ce{C2H6 + OH ->[\kappa_1] C2H5. + H2O}

-and \ce{C2H5. + O2 -> [\kappa_2] CH2H5O2}.

+and \ce{C2H5. + O2 ->[\kappa_2] CH5O2}.

\begin{equation}

\label{eqn:line}

-\ce{C2H6} + \ce{OH} ->[\kappa_3] \ce{C2H5O2}

+\text{ \ce{C2H6 + OH ->[\kappa_3] C2H5O2}}

\end{equation}

For simplicity, in this example, this will be the only equation for our mechanism. The resultant Flux \autoref{eqn:exflux} and resultant Jacobian \autoref{eqn:exjac} may be calculated.

-605,7 +605,7

\end{eqnarray}

-This forms a 'sparse' jacobian. Substituting numbers from subset mechanisms containing the methane and ethane precursors, we get \autoref{eqn:exjacsp}.

+This forms a 'sparse' Jacobian. Substituting numbers from subset mechanisms containing the methane and ethane precursors, we get \autoref{eqn:exjacsp}.

-817,7 +817,7

\subsubsection{Extracting The Required Results}

Model diagnostics such as concentration and the net flux passing through a species may be extracted directly from the DSMACC box model. These provide the baseline comparison and can be directly compared to the graph metrics. Species concentration tells us the abundance of different species, and the net-flux tells us how fast this is changing in time.

-As some species may have a fast inwards and outwards flux (low net-flux), the absolute flux is also included. Finally, the sensitivity of each species for other species is also extracted (the jacobian matrix). This serves to not only generate the graph used to represent the chemistry, (\autoref{sec:graphconstruction}) but also to identify the overall influence a species has on others in the network. This can be calculated by taking the net sum of the influence a species has on every other from the Jacobian. This is analogous to calculating the out-degree of a node in the jacobian network.\

+As some species may have a fast inwards and outwards flux (low net-flux), the absolute flux is also included. Finally, the sensitivity of each species for other species is also extracted (the Jacobian matrix). This serves to not only generate the graph used to represent the chemistry, (\autoref{sec:graphconstruction}) but also to identify the overall influence a species has on others in the network. This can be calculated by taking the net sum of the influence a species has on every other from the Jacobian. This is analogous to calculating the out-degree of a node in the Jacobian network.\

\input{metricics.tex}

-1110,7 +1110,7

\end{figure}

-Using a graph with reversed links weighted by model results of a jacobian is equivalent to a network created by an adjoint matrix (which is used to run models backwards). With this network, we run the PageRank algorithm with a 'personalised' initiated ranking vector of 1000000 for \ch{NC101CO} and -1 for everything else (A damping factor value of 0.01 is also used for the algorithm). This produces the results in \autoref{tab:nc101}. Here although all nodes receive a ranking value due to transportation within the PageRank algorithm, there is a distinct split between highly ranked values and the rest. It is found that \ce{NC101CO} has the strongest influence on itself (which makes sense), followed by that of α-pinene. Other more direct influences are seen from NAPINBOOH, NAPINBO, \ch{NAPINBO2}, from which NAPINBO has twice the influence from the other two. This is most likely as this has the highest net-flux from the model (\autoref{tab:nc101vdot}).

+Using a graph with reversed links weighted by model results of a Jacobian is equivalent to a network created by an adjoint matrix (which is used to run models backwards). With this network, we run the PageRank algorithm with a 'personalised' initiated ranking vector of 1000000 for \ch{NC101CO} and -1 for everything else (A damping factor value of 0.01 is also used for the algorithm). This produces the results in \autoref{tab:nc101}. Here although all nodes receive a ranking value due to transportation within the PageRank algorithm, there is a distinct split between highly ranked values and the rest. It is found that \ce{NC101CO} has the strongest influence on itself (which makes sense), followed by that of α-pinene. Other more direct influences are seen from NAPINBOOH, NAPINBO, \ch{NAPINBO2}, from which NAPINBO has twice the influence from the other two. This is most likely as this has the highest net-flux from the model (\autoref{tab:nc101vdot}).

\begin{table}[H]

\centering

4_lumping.tex

diff --git a/mechanism_lumping/combigned.tex b/mechanism_lumping/combigned.tex

index 62d358e..40c4bd6 100644

--- a/mechanism_lumping/combigned.tex

+++ b/mechanism_lumping/combigned.tex

-4,9 +4,11

In the previous chapters, we have discussed visualisation and its role in bridging the gap between data and understanding. We have applied centrality metrics to a chemical network to tell us what species are of importance and experimented in getting machine learning models to learn the chemical structure of the species in a mechanism. This final research chapter provides a (brief) overview of current mechanism reduction techniques while providing two novel alternatives to aid the process.

-Science often deals with the problem of understanding complexity. Such a task may be accomplished through organisation and partitioning, for example, the learning of a new skill through chunking (breaking up a problem into manageable chunks), or the parallelisation of a sizeable mathematical problem. In cases where such methods fail, we are forced to 'disregard' complexity. It is common to approximate an atom as a sphere or the value π as 3 with little consequence to the overall result of a calculation. The process of lumping has long been used to replace a complex, changing process (e.g. Quantum Mechanics or Boundary Layer Fluid Dynamics) with a more straightforward constant process, \citep{approx}. In such cases, an approximation may be far more useful than a lengthy exact solution, or none at all provided the primary criteria/outcome is identified and optimised for (evaluated against a benchmark or standard).

+Science often deals with the problem of understanding complexity. Such a task may be accomplished through organisation and partitioning (e.g. chunking a problem into smaller problems) and processing these at the the same time using many workers (parallelism). In cases where such methods fail, we are forced to disregard' complexity. To do this physical processes may be simplified\footnote{It is common to approximate a year as 365 days, an atom as a sphere and replace the Van der Walls equation with the ideal gas law (for normal pressures).}, or described using mathematics. Theorems and ideas may be applied to emulatereal-world' outcomes based on the platonian concept of an abstract ‘Ideal' world \citep{platoform, physapprox}.

-Similar problems of complexity are seen within the chemistry of the atmosphere. An example is seen within the Master Chemical Mechanism (MCM v3.3.1), \citep{mcm}, this contains 1228 \ch{RO2} reactions. If written explicitly, all \ce{RO2-RO2} (gross and self) interactions would result in a total of 1,507,984 reactions. Instead, the MCM overcomes this problem by creating a \ch{ro2} pool, with which all \ch{RO2} species react. This results in a mechanism which preserves the quality of science (the primary goal of the MCM is to preserve \ch{o3} prediction) with only 0.000814 of the total possible \ch{ro2+ro2} reactions.

+ The process of lumping has long been used to replace a complex, changing process (e.g. Quantum Mechanics or Boundary Layer Fluid Dynamics) with a more straightforward constant process, \citep{approx}. In such cases, an approximation may be far more useful than a lengthy exact solution, or none at all provided the primary criteria/outcome is identified and optimised for (evaluated against a benchmark or standard).

+Similar problems of complexity are seen within the chemistry of the atmosphere. An example is seen within the Master Chemical Mechanism (MCM v3.3.1), \citep{mcm}, this contains 1228 \ch{RO2} reactions. If written explicitly, all \ce{RO2-RO2} (gross and self) interactions would result in a total of 1,507,984 reactions. Instead, the MCM overcomes this problem by creating a \ch{ro2} pool, with which all \ch{RO2} species react. This results in a mechanism which preserves the quality of science (the primary goal of the MCM is to preserve \ch{o3} prediction) with only 0.000814 of the total possible \ch{ro2} - \ch{ro2} reactions.

However, even with such simplifications, atmospheric chemical mechanisms have been increasing in size over the last ten years (\citep{defra1},\autoref{fig:webmcm}). With the ability to automate their construction, mechanisms with species numbers of the millions become possible. Although the existence of more-explicit mechanisms may improve the quality of science produced, they can cause problems for efficient computation, diagnosis and analysis. This chapter shall look at two methods in which we may simplify a mechanism by grouping species with similar reaction patterns together. These are through the use of species lifetime (\autoref{sec:lifetime}) and graph-based clustering (\autoref{sec:graphreduction}).

-37,7 +39,7

\subsection{Species Removal}

-Similar to reaction removal, the removal of species is useful because the removing or combining of species inherently reduces or simplifies the reactions within a mechanism. This method also has added benefit of reducing the size of the jacobian matrix used to propagate the chemical system forwards. For large systems which do not use a sparse framework, storing a n^2 matrix in memory can prove difficult.

+Similar to reaction removal, the removal of species is useful because the removing or combining of species inherently reduces or simplifies the reactions within a mechanism. This method also has added benefit of reducing the size of the Jacobian matrix used to propagate the chemical system forwards. For large systems which do not use a sparse framework, storing a n^2 matrix in memory can prove difficult.

Many methods of species reduction are possible. The simplest of these is through the use of trial and error \citep{tur1990} (Method 1). Here the consuming reactions for a species are removed, and if the resulting deviation in results between the full and reduced mechanism is small within a certain threshold, their results are retained. The main downside to this is that it only works on a per-species level, which may be very resource-consuming for large mechanisms.

-80,12 +82,10

Reductions have been made on a compound-by-compound basis and compared to the MCM using a series of 5-day box-model simulations, \citep{cri}.

\paragraph*{Why further simplify the CRI network?}\label{sec:whycri}

-5809 species and 17224 reactions

-CRI v2.2 \citep{cri} is a mechanism of 422 species and 1261 reactions - that is 7% of the species and 7% of the reactions of the full MCM. Although this is significantly smaller than the full MCM, it may still prove problematic if used within a global model - for comparison the GEOS-Chem\footnote{A global 3D model of atmospheric chemistry driven by meteorology from NASA's Goddard Earth Observing System (GEOS), \citep{geos}.} standard chemistry is approximately half the size of this, \citep{geosgit}.

+CRI v2.2 \citep{cri} is a mechanism of 422 species and 1261 reactions - that is ~7% of the full MCM (5809 species and 17224 reactions). Although this is significantly smaller than the full MCM, it may still prove problematic if used within a global model - for comparison the GEOS-Chem\footnote{A global 3D model of atmospheric chemistry driven by meteorology from NASA's Goddard Earth Observing System (GEOS), \citep{geos}.} standard chemistry is approximately half the size of this, \citep{geosgit}.

\subsection{The Box-Model}

-The box model is an adapted version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) \citep{dsmacc,dsmaccgit}. Recent updates allow for multiple parallel runs, easy extraction of rates, fluxes and the jacobian matrix as well as a simple Ncurses (a command like semi-graphic interface) interface for loading and parsing new files.

+The box model is an adapted version of the Dynamically Simple Model of Atmospheric Chemical Complexity (DSMACC) \citep{dsmacc,dsmaccgit}. Recent updates allow for multiple parallel runs, easy extraction of rates, fluxes and the Jacobian matrix as well as a simple Ncurses (a command like semi-graphic interface) interface for loading and parsing new files.

The DSMACC model works by using the Kinetic PreProcessor (KPP), \citep{kpp}, to generate Fortran code, which can then be used to integrate the provided mechanism. As there were some issues presented a pre-pre parser code is used before running KPP. Occasionally a post parser may be required on some of the files to produce the desired output.

-103,7 +103,7

\label{eqn:icslhs}

\end{equation}

\section{Graph Based Reduction}\label{sec:graphreduction}

-It has been shown that a graph-based representation of the atmospheric chemical network proves useful in both the visual and mathematical analysis of simulation results (\autoref{c2,c3}). It, therefore, follows that the network representation of mechanism may also have its uses in the simplification, and thus reduction, of chemical complexity. This section will outline the basic methods of modularity (cluster) detection with the graph framework, the different methods in which this may be done and eventually apply it to a case example representative of the chemistry within the London environment.

+It has been shown that a graph-based representation of the atmospheric chemical network proves useful in both the visual and mathematical analysis of simulation results (\autoref{ch2,ch3}). It, therefore, follows that the network representation of mechanism may also have its uses in the simplification, and thus reduction, of chemical complexity. This section will outline the basic methods of modularity (cluster) detection with the graph framework, the different methods in which this may be done and eventually apply it to a case example representative of the chemistry within the London environment.

-202,13 +202,13

\subsection{Species Type And Clustering}

-The bubble chart provides an intuitive way to represent groups for interactive or small systems but is less useful for larger numbers of species and print (\autoref{fig:imbubble}). Instead, a tree approach is better suited to revealing the hierarchical structure of the network, as shown in \autoref{fig:imap2page}. Here branches are numerically labelled on each level, allowing us to navigate the structure using a sequence of numbers (e.g. to get to \ch{c4h6} we take the first branch from the centre, followed by the fifth branch after that resulting in the notation 1 . 5 . \ch{c4h6}).

+The bubble chart provides an intuitive way to represent groups for interactive or small systems but is less useful for larger numbers of species and print (\autoref{fig:imbubble}). Instead, a tree approach is better suited to revealing the hierarchical structure of the network, as shown in \autoref{fig:imap2page}. Here branches within \autoref{fig:imap2page} are numerically labelled for each level. This allows us to navigate the hierarchy using a sequence of numbers (e.g. to get to \ch{c4h6} we take the branch 1 from the centre, followed by branch 5 - resulting in the notation 1.5.C4H6).

This split notation allows a general overview of the mechanism structure, as well as the reasoning/process of the clustering algorithm. The first level split in \autoref{fig:iml1} shows branches 1,2 and 5 to have origins in the linear (n-) alkane species. This can be seen through both the emitted species (bold) and the \emph{RN} prefix of the species. Here the linear alkanes can react with OH to extract hydrogen and then from a \ce{RO2}, or produce a carbonyl \emph{\ce{CARBxx}}, which can then go on to produce the \emph{\ce{RNxxO2}} peroxy radical.

-Except for benzine in 2.14, branches 3 and 4 contain the aromatic species in the network. Branches 4.{2,5,9,11} all consist of \emph{\ce{RAxxO2}} species, which are the product of the addition of OH to toluene/benzine ringed species. 4.{1,7,8} and 1.5 contain peroxy radicals formed from the degradation of conjugated dienes \emph{\ce{RUxxO2}}. For the CRI v2.2 mechanism these are only isoprene and 1,3-butadiene. Such peroxy radicals often go on to form unsaturated carbonyls, as denoted by \emph{\ce{UCARBxx}}.

+Except for benzene in 2.14, branches 3 and 4 contain the aromatic species in the network. Branches 4.{2,5,9,11} all consist of \emph{\ce{RAxxO2}} species, which are the product of the addition of OH to toluene/benzene ringed species. 4.{1,7,8} and 1.5 contain peroxy radicals formed from the degradation of conjugated dienes \emph{\ce{RUxxO2}}. For the CRI v2.2 mechanism these are only isoprene and 1,3-butadiene. Such peroxy radicals often go on to form unsaturated carbonyls, as denoted by \emph{\ce{UCARBxx}}.

-Branch 3 contains the monoterpenes. This can be seen in 3.{2,5} ($\alpha-$pinene) and 3.6 ($\beta-$pinenen). Here peroxy radicals formed from the reaction with the e\textbf{n}docyclinc\footnote{Inside the pinene ring.} and e\textbf{x}docyclinc\footnote{Outside the pinene ring.} double bonds of $\alpha-$ and $\beta-$ pinene are denoted with the prefix \emph{\ce{RTN}} and \emph{\ce{RTX}}.

+Branch 3 contains the monoterpenes. This can be seen in 3.{2,5} ($\alpha-$pinene) and 3.6 ($\beta-$pinene). Here peroxy radicals formed from the reaction with the e\textbf{n}docyclinc\footnote{Inside the pinene ring.} and e\textbf{x}docyclinc\footnote{Outside the pinene ring.} double bonds of $\alpha-$ and $\beta-$ pinene are denoted with the prefix \emph{\ce{RTN}} and \emph{\ce{RTX}}.

The \emph{\ce{RIxxO2}} prefix was used initially for the peroxy radicals iso (`i-') alkanes and their carbonyl products - branches 3.{1,4}, however, they tend to mainly be used for smaller branched precursors which produce acetone (\ch{CH3COCH3}) as a significant product in their oxidation chain (branch 3.1). Acetone is a relatively unreactive carbonyl, the fact that it is water-soluble means that they may be washed out of the atmosphere by precipitation, \citep{acetonerain}. This may have been seen to interrupt the ozone formation process under regional-scale photochemical smog conditions in north-western Europe.

-283,7 +283,7

v2 = [i,j,k, \dots z]

\end{equation}

- This can be done using pythoagoras' theorem in \autoref{euclid}:

+ This can be done using Pythagoras' theorem in \autoref{euclid}:

\begin{equation}

e_{dist} = \sqrt{(a-i)^2 + (b-j)^2 + (c-k)^2 + \dots + (n-z)^2}

-354,8 +354,8

The agreement of both metrics suggests a similarity between the lifetime values and their change in time for simulation. This is in agreement of with the $x-y$ plot of the species. In selecting species that are part of the same initial cluster and have a high agreement between both similarities, it is possible to gauge the suitability for two species to be lumped together.

- \subsection{A Quick Concentration Comparison}

- Having described how the similarity distances work, \autoref{fig:metric} showed the locations of the best and worst matched pairs. This subsection looks a the differences between these using a log10 ensemble of the concentrations for the 300 simulations used in the results section. \autoref{fig:bestworst}(a,b) show that the best matching pairs contain an easy to match flat decay curve, with the worst \autoref{fig:bestworst}(c,d) often containing a combination of a species which decays with one which undergoes a photolytic reaction.

+ \subsection{A Quick Comparison}

```
+ Having described how the similarity distances work, \autoref{fig:metric} showed the locations of the best and worst matched pairs. This subsection looks a the differences between these using a log10 ensemble of the mixing ratios for the 300 simulations used in the results section. \autoref{fig:bestworst}(a,b) show that the best matching pairs contain an easy to match flat decay curve, with the worst \autoref{fig:bestworst}(c,d) often containing a combination of a species which decays with one which undergoes a photolytic reaction.

\begin{figure}[H]
```

-380,7 +380,7

```
\includegraphics[width=\textwidth]{ensemble/C2H5CO3-CH3NO3.pdf}

\caption{}

\end{subfigure}%\

- \caption{\textbf{Comparing the best (a-b) and worst (c-d) species combinations using the combigned similarity metrics.}Here species which only undergo a simple decay seem to be the easiest to group together. Species pairs between an photolytic and non photolytic species produce different profiles at differing magnitudes and are therefore difficult to match.}

+ \caption{\textbf{Comparing the best (a-b) and worst (c-d) species combinations using the combined similarity metrics.} Here species which only undergo a simple decay seem to be the easiest to group together. Species pairs between an photolytic and non photolytic species produce different profiles at differing magnitudes and are therefore difficult to match.}

\label{fig:bestworst}

\end{figure}
```

-394,7 +394,7

```
\section{Results}

-In order to get a representation of the mechanism, we run 300 randomly initiated scenarios (\autoref{sec:lumpinputs}). The experimental setup is one such that it is possible to add more data points at a later date. From each simulation, the no diagonal elements of the jacobian are used to construct a graph representative of the aggregated hourly means of the simulation output. Each of these graphs is then run through the infomap algorithm, and a grouping/clustering produced. Each infomap is run 100 times, where the result with the best fit (shortest code length) is taken - this is an optional parameter on the algorithm.

+In order to get a representation of the mechanism, we run 300 randomly initiated scenarios (\autoref{sec:lumpinputs}). The experimental setup is one such that it is possible to add more data points at a later date. From each simulation, the no diagonal elements of the Jacobian are used to construct a graph representative of the aggregated hourly means of the simulation output. Each of these graphs is then run through the infomap algorithm, and a grouping/clustering produced. Each infomap is run 100 times, where the result with the best fit (shortest code length) is taken - this is an optional parameter on the algorithm.

\subsection{The Co-Grouping Network}
```

-426,7 +426,7

```
\includegraphics[width=\textwidth]{fig/c4.png}

\caption{>40% of graphs}

\end{subfigure}%

-\caption{\textbf{Filetering the infomap clustering relationship matrix/graph} How the clustering relationship network changes as weak links (links between species which do not appear in many of the infomap groupings) are removed. }

+\caption{\textbf{Filtering the infomap clustering relationship matrix/graph} How the clustering relationship network changes as weak links (links between species which do not appear in many of the infomap groupings) are removed. }

\label{fig:infomapprune}

\end{figure}
```

-498,7 +498,7

\includegraphics[width=\textwidth]{ensemble/NRI12OOH-NRI1202.pdf}

\caption{\ce{ NRI12OOH \ \ \ NRI1202 }}

\end{subfigure}%\

-\caption{\textbf{Comparing the best and worst pairs from \autoref{tab:lumppair}}Time is in the format DD-MM HH}

+\caption{\textbf{Comparing the best (a-b) and worst (c-d) species pairs from \autoref{tab:lumppair}}. Species which make a good candidate for reduction have a similar diurnal profile and production/loss patterns as well as ranges of magnitude in which the concentration lies. This is seen in subplots (a) and (b). Bad pairings either cover very different magnitude ranges (d) or have dice different temporal profiles (c and d). Time is in the format DD-MM HH}

\label{fig:lumppair}

\end{figure}

-542,7 +542,7

\newpage

-\section{Conculsions}

+\section{Conclusions}

\autoref{ch2} discussed graphs as a useful method for representing the chemistry within a mechanism. Building on that \autoref{ch3} showed that graph centrality metrics could be used to mathematically locate nodes (species) of importance from the chemical network from a chemical simulation. This chapter explores the chemical structure of the MCM network and uses graph clustering methods to locate groups of similar chemistry (\autoref{fig:imap2page}).

5_DR.tex

diff --git a/dr/combigned.tex b/dr/combigned.tex

index fcd93ea..32f0323 100644

--- a/dr/combigned.tex

+++ b/dr/combigned.tex

-2,7 +2,9

\section{Introduction}

\subsection{Historical Significance}

-The established process of trial and error has always underpinned our survival \citep{TrialandError}. Babies are born to rely on a set of sensory reflexes and a framework for physical reasoning \citep{pr}, and with these, we develop methods to navigate the influence of change within a physical, and auditory space \citep{objects}. This method of decision making is reflected in our adult lives with ideas and actions being limited in choice by our intuition and experience \citep{descartes}. 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. Nevertheless, even with all prior experience, the discovery of new and untested techniques involve the tortuous traipsing through a sea of uncertainty. Such methods sometimes prove fruitful, through accidental discoveries of items such as x-rays, penicillin... \citep{accidental}; finding novel applications for existing methods such as optical tweezers for chemistry or the abstract field of maths utilised by Einstein, but more often than not end in the constant evolution of a pre-existing project with no apparent result.

+The established process of trial and error has always underpinned our survival \citep{TrialandError}. Babies are born to rely on a set of sensory reflexes and a framework for physical reasoning \citep{pr}, and with these, we develop methods to navigate the influence of change within a physical, and auditory space \citep{objects}. This method of decision making is reflected in our adult lives with ideas and actions being limited in choice by our intuition and experience \citep{descartes}. 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. Nevertheless, even with all prior experience, the discovery of new and untested techniques involve the tortuous traipsing through a sea of uncertainty.

+ Such methods sometimes prove fruitful, through accidental discoveries of items such as polyetheylene, penicillin, x-rays, nylon, teflon, velcro etc. \citep{accidental,serendipity}; finding novel applications for existing methods such as optical tweezers for chemistry or the abstract field of maths utilised by Einstein, but more often than not end in the constant evolution of a pre-existing project with no apparent result.

\subsection{Theory And Simulation In Science}

-90,17 +92,13

\subsubsection{Species Names}

-In \autoref{ch4} it was shown that the dedicated species names for species in the CRI mechanism were often representative of their structural properties. This also applies for the MCM, where an intuitive naming convention following the FACSIMILE format 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.

+In \autoref{ch4} it was shown that the dedicated species names for species in the CRI mechanism were often representative of their structural properties. This also applies for the MCM, where an intuitive naming convention following the FACSIMILE format 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.

\subsubsection{SMILES Strings}\label{sec:SMILES}

- SMILES (‘Simplified Molecular-Input Line-Entry System') provide a human-readable representation of the molecular structure,

- \citep{smiles}. 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 \autoref{sec:fingerprints}, SMILES strings provide a useful tool for quickly comparing species structure.

+ SMILES (‘Simplified Molecular-Input Line-Entry System') provide a human-readable representation of the molecular structure, \citep{smiles}. 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 \autoref{sec:fingerprints}, SMILES strings provide a useful tool for quickly comparing species structure.

\paragraph*{Construction Methodology of SMILES strings}

The construction of a SMILES string happens in three parts:

-206,7 +204,7

\subsubsection{Node Embeddings (Node2Vec)}\label{sec:n2vec}

\autoref{ch2} and \autoref{ch3} showed that the underlying structure of a chemistry mechanism graph contains information about the species and reactions within it. Here as a species is oxidised the O-C ratio increases. Long-chain VOCs are likely to fragment into two radicals, producing smaller more oxidised species. Eventually, this process leads to the production of carbon dioxide and water. \autoref{fig:vk} shows a subset of the MCM representing the chemistry in Beijing. Node colour and size show the increase of oxidation as species head towards CO at the centre) - lighter colour and larger node.

-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 \citep{node2vec}. Since this may also be used for dimensionality reduction, it is described within the next section

(\autoref{sec:n2v}).

+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 \citep{node2vec}. Since this may also be used for dimensionality reduction, it is described within the next section (\autoref{sec:n2v}).

\begin{figure}[H]

-238,7 +236,7

In this section, we begin by explaining the data preparation required for dimensionality reduction (\autoref{sec:prep}) before describing the different possible methods of reducing the dimensions of a dataset through Principle Component Analysis, Auto Encoders and t-Distributed Stochastic Neighbor Embedding.

-\subsection{Preperation Of The Data}\label{sec:prep}

+\subsection{Preparation Of The Data}\label{sec:prep}

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 (i) normalisation and (ii) standardisation.

-262,7 +260,7

\end{equation}\

-\subsection{Principle Component Analysis (Pca)}

+\subsection{Principle Component Analysis (PCA)}

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 \citep{pcareview}.

-279,7 +277,7

\end{figure}

-\subsubsection{Mathematical Explanation Of Pca}

+\subsubsection{Mathematical Explanation Of PCA}

\emph{\textbf{Note:} The basic statistics/mathematics required to understand this section is shown in \autoref{apendix:pca}. Please read this if you are not familiar with any of the terms below.

-289,7 +287,7

-\subsection{T-Distributed Stochastic Neighbor Embedding (t-SNE)}\label{sec:overcrowd}

+\subsection{t-Distributed Stochastic Neighbor Embedding (t-SNE)}\label{sec:overcrowd}

t-SNE is an algorithm designed with visualisation in mind \citep{tsne}. 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, as shown in \autoref{fig:tsneforcegraph}. This allows the ability to capture non-linear structures in the data which cannot be accomplished through linear mapping methods (e.g. PCA).

-356,7 +354,7

-\subsection{Pca Vs t-SNE, A Quick Comparison.}

+\subsection{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. In essence, an example of this give by \cite{wyche} where mechanisms can be separated into different pathways (on account of the underlying chemistry) and \cite{kinetics} where sensitivity analysis is used within mechanism reduction. 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.

-401,7 +399,7

-\subsection{The Auto-Encoder (Ae)}\label{sec:ae}

+\subsection{The Auto-Encoder (AE)}\label{sec: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 \citep{aenoise,aeim} and as an input to more complex machine learning models.\

-479,7 +477,7

\autoref{fig:n2vedge} shows the return and input parameters ($p\ &\ q$) determine how fast we explore the network and our probability to leave the neighbourhood. 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 $q \ge 1$ and $p > q$ our parameters as $p = 2.0,q=1.1$. In the case of a weighted graph (something that we are \textit{not} exploring within this chapter) the resultant $alpha$ value calculated is further multiplied by the edge weight.

-To generate the node2vec embeddings for each species, we use the python2 code provided by the original paper by \cite{node2vec} with a set of 50000 random walks, each of length 9 product/reaction generations. The reasoning behind this is that we have a large graph, with a power-law like structure (where species are often heavily connected, \autoref{ch3}).

+To generate the Node2Vec embeddings for each species, we use the python2 code provided by the original paper by \cite{node2vec} with a set of 50000 random walks, each of length 9 product/reaction generations. The reasoning behind this is that we have a large graph, with a power-law like structure (where species are often heavily connected, \autoref{ch3}).

\textit{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. }

-722,7 +720,7

As was touched on in \autoref{sec:mathclustanalysis} the MACCS input consists of a series of logical questions about a species structure. Since many of those questions regard the existence of a Nitrogen atom, data was separated species with a Nitrate or PAN group, and those without. In making a series of decisions on which cluster a species falls under, this largest most recurring branch for the RandomForrestClassifier (imagine of temperature in \autoref{fig:iodenetree}) falls under the existence of a Nitrate group.

- The main inconsistency between clusters and DR algorithms comes from the node2vec embedding (e) - much of which can be explained by the poor performance of the DR and clustering algorithms of separating the chemistry into groups (see plots in \autoref{sec:cldist}). \autoref{sec:selectcomp} continues this analysis by comparing output with {{content}}lt;3$ clusters each against the graph plots presented in this subsection. The content of individual groupings is explored for an output with multiple clusters.

+ The main inconsistency between clusters and DR algorithms comes from the Node2Vec embedding (e) - much of which can be explained by the poor performance of the DR and clustering algorithms of separating the chemistry into groups (see plots in \autoref{sec:cldist}). \autoref{sec:selectcomp} continues this analysis by comparing output with {{content}}lt;3$ clusters each against the graph plots presented in this subsection. The content of individual groupings is explored for an output with multiple clusters.

\begin{figure}[H]

-752,7 +750,7

Using the DR output where only two/three groups are located by the clustering algorithms we have (\autoref{fig:biMACCS} and \autoref{fig:biN2V}). In exploring the MACCS key input for the PCA and t-SNE DR algorithms (\autoref{fig:biMACCS}) we find that for the cumulative importance bar charts we know that the existence of Nitrates is vital in the split determining which group a species falls into. This manifests itself as having a single cluster containing PAN and Nitrate species, with others not. In the t-SNE plot (\autoref{fig:biMACCS}b) we see that there exists a third group which is missing both Aldehyde and PAN functionalisation for each species. This is shown by the teal colour in \autoref{fig:tsnevis}c and resides between the Nitrogen-containing and Nitrogen-deficient groups.

-\autoref{fig:biN2V} shows the comparison of the Node2Vec embedding using PCA and the AE DR algorithms. In \autoref{fig:pcavis}e and \autoref{fig:aevis}e, it is seen that these are generally not separated into well-partitioned clusters. Both groups consist of one large cluster (shown by the second bar chart of each row which contains all functional groups) and one or two fragment ones. In exploring the AE plot (\autoref{fig:biN2V}b), it is seen that as part of the cumulative plot (right), the -OOH functional group is an important separatory factor since the smaller of the two groups does not contain any species which contain a hydroperoxy functional group. In the PCA plot, although providing different cumulative results, again shows species within the smaller groups not containing any \ce{RO, RCO3, OOH, ONO2} or OOH functional groups. This can potentially be due to the graph structure, where the random walker (which generates the node2vec embedding) has become trapped by a group of non-oxidised species.

+\autoref{fig:biN2V} shows the comparison of the Node2Vec embedding using PCA and the AE DR algorithms. In \autoref{fig:pcavis}e and \autoref{fig:aevis}e, it is seen that these are generally not separated into well-partitioned clusters. Both groups consist of one large cluster (shown by the second bar chart of each row which contains all functional groups) and one or two fragment ones. In exploring the AE plot (\autoref{fig:biN2V}b), it is seen that as part of the cumulative plot (right), the -OOH functional group is an important separatory factor since the smaller of the two groups does not contain any species which contain a hydroperoxy functional group. In the PCA plot, although providing different cumulative results, again shows species within the smaller groups not containing any \ce{RO, RCO3, OOH, ONO2} or OOH functional groups. This can potentially be due to the graph structure, where the random walker (which generates the Node2Vec embedding) has become trapped by a group of non-oxidised species.

\begin{landscape}

-777,7 +775,7

\hfill

\includegraphics[width=1.6\textheight]{./outputs/AE/node2vec/group.png}

\ (b) AE

- \caption{ \textbf{Comparing individual clusters between node2vec for PCA and t-SNE algorithm output.} The bar chart to the right is the cumelative chart which represents the splits in deciding the cluster a species falls into from \autoref{sec:fsclust}. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.}

+ \caption{ \textbf{Comparing individual clusters between Node2Vec for PCA and t-SNE algorithm output.} The bar chart to the right is the cumelative chart which represents the splits in deciding the cluster a species falls into from \autoref{sec:fsclust}. Unlabeled bar charts to the left represent the partitioning of species within an individual cluster.}

\label{fig:biN2V}

\end{figure}

\end{landscape}

6_conclusion.tex

diff --git a/conclusion/combigned.tex b/conclusion/combigned.tex

index 9761db7..7295fe4 100644

--- a/conclusion/combigned.tex

+++ b/conclusion/combigned.tex

-13,5 +13,56

Finally, in preparation for future research, the use of different species structure representations was run through a number of dimensionality reduction algorithms. Here the different inputs were reduced to two dimensions and plotted in a $x-y$ scatterplot. Analysis of these scatterplots showed that the t-SNE algorithm provided the best spacing between clusters. Additionally, it is found that the type of input can influence the features that are obtained as part of the dimensionality reduction process. It is suggested that if using a neural network, the molecular quantum number or tokenised SMILES input are likely to produce the best results.

-\section{Future Work}

+\section{Future Work}\label{sec:futurework}

With the newly emerging age of big data', the fields of data analysis and graph theory are ever-improving. An example of this is the development and use of graph convolutional neural networks in 2016. Here a neural network receives not only information about the structure of an item, but also the relationships it has with everything else. Theoretically, this framework may allow the artificial network tolearn' the relationships and protocols of a chemical mechanism and generate the correct chemical pathways based on the structure of a new (and unseen species).

+\newpage

+\section*{Reproducability}

+\addcontentsline{toc}{section}{\protect\numberline{}Reproducability}%

+The code used within this thesis is provided ‘as is' within the relevant repositories. There will be an attempt to make it more presentable and fully documented within the near future, but this has not yet happened. For many of the tasks it is possible to download a clean repository and implement any relevant changes yourself.

+\subsection*{The Box Model}

+Most of the work in this thesis relies on the use of the DSMACC Box model \citep{dsmacc}. In order to reproduce it the specific code I have used can be found in \citep{dsmaccgit}, however any box model which allows you to extract both the fluxes and Jacobian matrix may be used.

+\subsection*{Photolysis Calculations}

+Photolysis rates are calculated with version 5.2 of the Tropospheric and Ultraviolet and Visible codebase. Photolysis rates are calculated once at the start of each box model run and then interpolated with the use of cubic splines to provide the values required throughout the day. This can be located at \citep{tuv}, Photolysis rates within the J array correspond to the lines outlined in \verb|./INPUTS/MCMTUV| and are hard wired within the \verb|./MCMvXX.inc| include files.

+\subsection*{The Master Chemical Mechanism}

+For the work, we have made use of various versions of the master chemical mechanism \citep{mcm}. Different versions of this and its reduced component (CRI) can be obtained from the MCM website: \url{mcm.york.ac.uk}. Alternatively the KPP presentation of all the mechanisms I have used are located within the \verb|./mechanisms| folder in the DSMACC repository.

+\subsection*{Kinetic Pre-Processor}

+To transpose the chemical mechanism into a usable format, the Kinetic Pre-Processor rewrites the human readable first order ordinary differential equations into FORTRAN95 code. The version of this originates from FlexChem - the KPP rewrite used in GEOSChem (KPP 2.3.01). This is located at \url{https://github.com/wolfiex/kpp_2.3.01_gc/}

+\subsection*{ML libraries}

+Simple processing tasks as clustering , PCA and t-SNE generally make use of the Scikit-Learn package \citep{sklearn}.

+Graph Layouts such at TSNET and Mercator can be found in \url{https://github.com/wolfiex/tsNET} and \url{https://github.com/networkgeometry/mercator}.

+The AutoEncoder code can be found within the DSMACC repository at \url{https://github.com/wolfiex/DSMACC-testing/blob/master/dsmacc/examples/rate_ae.py} and the Graph AutoEncoder at \url{https://github.com/tkipf/gae}.

+Although not documented, the aim of this thesis was to work up to the use of a graph convolutional network such as the one in \url{https://github.com/wolfiex/gcn}.

```
+\subsection*{Chemial representation and Molecular Keys}

+Chemical species representation for SMILES and INCHI strings are taken directly from the MCM. Additional conversions into MACCS and MQN keys make use of the RDKIT python package: \citep{rdkit}.

+\subsection*{Observation and model run reproducibility}

+To reproduce the results made from field campaigns it is possible to extract the data directly from the Centre for Environmental Data Analysis. The four field campaigns used are provided below.

+\begin{itemize}

+ \item{\url{https://catalogue.ceda.ac.uk/uuid/648246d2bdc7460b8159a8f9daee7844}}

+ \item {\url{https://catalogue.ceda.ac.uk/uuid/81892deb2dd5e7f0d26b9c587af45f3d}}

+ \item{\url{https://catalogue.ceda.ac.uk/uuid/a457d9715f3c4bc295ef975932e491d9}}

+ \item {\url{https://catalogue.ceda.ac.uk/uuid/cee49a1f044b79d5413b7a0282467508}}

+\end{itemize}

+Once downloaded, these are wrangled into the initial conditions CSV format for the use in model runs - some of which are spun up to steady state based on the users preference and aim of the study.

+Non-observational runs are initated through the use of a Latin hypercube format to provide a random assortment of initial concentrations within a pre-defined limit. An example of the intial conditions output for one run of these can be found in \url{https://github.com/wolfiex/DSMACC-testing/blob/master/InitCons/lhs_spinup.csv}.

+\bibliographystyle{apalike}

+\bibliography{bibtex}
```

Glossary

```
diff --git a/glossary.tex b/glossary.tex

ew file mode 100644

index 0000000..8e46b64

--- /dev/null

+++ b/glossary.tex

-0,0 +1,55

+\section*{List of Abbreviations}

+\addcontentsline{toc}{section}{\protect\numberline{}List of Abbreviations}%

+\subsection*{Atmosphere}

+ \begin{center}

+ \begin{tabular}{ p{.18\textwidth}p{.65\textwidth} }

+ \textbf{HOx } & OH + \ce{HO2}\

+\textbf{NOx } & NO + NO2\

+\textbf{NOy } & $\Sigma$ oxidized atmospheric odd-nitrogen species\

+\textbf{NOz } & NOy - NOx\

+\textbf{PAN } & PeroxyAcyl Nitrate\
+\textbf{pp{m,b,t}v } & parts per {million, billion, trillion} by volume\

+ \end{tabular}

+ \end{center}

+\subsection*{Modelling}

+ \begin{center}

+ \begin{tabular}{ p{.18\textwidth}p{.65\textwidth} }

+ \textbf{DSMACC } & Dynamically Simple Model of Atmospheric Chemical Complexity\

+\textbf{GEOSChem } & Chemistry component of NASA's Goddard Earth Observing System\

+\textbf{KPP } & Kinetic Pre Processor\

+\textbf{ROPA } & Rate of Production (and Loss) Analysis\

+\textbf{TUV } & Tropospheric, Ultraviolet and Visible Radiation Model\

+ \end{tabular}

+ \end{center}

+\subsection*{Artificial Intelligence}

+ \begin{center}

+ \begin{tabular}{ p{.18\textwidth}p{.65\textwidth} }

+ \textbf{CRI } & Common Representative Intermediates\

+\textbf{INCHI } & International Chemical Identifier (developed by IUPAC)\

+\textbf{IUPAC } & International Union of Pure and Applied Chemistry\

+\textbf{MACCS } & Molecular ACCess System\

+\textbf{MCM } & Master Chemical Mechanism\

+\textbf{MQN } & Molecular Quantum Number\

+\textbf{SMARTS } & SMILES arbitrary target specification\

+\textbf{SMILES } & Simplified Molecular-Input Line-Entry System\

+ \end{tabular}

+ \end{center}

+\subsection*{Artificial Intelligence}

+ \begin{center}

+ \begin{tabular}{ p{.18\textwidth}p{.65\textwidth} }

+ \textbf{AE } & Auto Encoder\

+\textbf{DBSCAN } & Density-Based Spatial Clustering of Applications with Noise\

+\textbf{DR } & Dimensionality Reduction\

+\textbf{GMM } & Gaussian Mixture Model\

+\textbf{GNN } & Graph Neural Network\
+\textbf{ML } & Machine Learning\

+\textbf{OPTICS } & Ordering Points To Identify the Clustering Structure\

+\textbf{PCA } & Principle Component Analysis\

+\textbf{t-SNE } & t-distributed Stochastic Neighbor Embedding\

+ \end{tabular}

+ \end{center}
```

Bibliography

```
diff --git a/bibtex.bib b/bibtex.bib

index 8b8f0b8..f907065 100644

--- a/bibtex.bib

+++ b/bibtex.bib
```

-767,7 +767,7

```
author = {Joshua Stevens},

language = {en},

month = {February},

- note = {\url{https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china?fbclid=IwAR1z9jXZfY8xNZsCCRRo8Eor2hCjbNDIV70wXGOlzmNyFPkFBesURDCAwB4}},

+ note = {\url{https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china}},

publisher = {NASA Earth Observatory},

title = {{Airborne Nitrogen Dioxide Plummets Over China}},

year = {2020}
```

-825,7 +825,7

```
@article{closeness-book,

- author = {poliaktiv},

+ author = {Poliaktiv},

note = {\url{https://www.politaktiv.org/documents/10157/29141/SocNet_TheoryApp.pdf}},

title = {{Social Network Analysis: Theory And Applications}},

year = {2011}
```

-1083,7 +1083,7

```
note = {\url{http://www.sciencedirect.com/science/article/pii/S1352231008006742}},

number = {31},

pages = {7185 - 7195},

- title = {A Common Representative Intermediates (Cri) Mechanism For Voc Degradation. Part 1: Gas Phase Mechanism Development},
+ title = {A Common Representative Intermediates (Cri) Mechanism For VOC Degradation. Part 1: Gas Phase Mechanism Development},

volume = {42},

year = {2008}
```

-1252,7 +1252,7

```
note = {\url{https://www.atmos-chem-phys.net/5/641/2005/}},

number = {3},

pages = {641--664},

- title = {Development Of A Detailed Chemical Mechanism (Mcmv3.1) For The Atmospheric Oxidation Of Aromatic Hydrocarbons},

+ title = {Development Of A Detailed Chemical Mechanism (MCMv3.1) For The Atmospheric Oxidation Of Aromatic Hydrocarbons},

volume = {5},

year = {2005}
```

-1377,7 +1377,7

```
author = {Ellis, Dan},

institution = {Github},

note = {\url{https://github.com/wolfiex/DSMACC-testing}},

- title = {{Dsmacc-Testing}},

+ title = {{DSMACC-Testing}},

year = {2020}
```

-1645,6 +1645,23

```
year = {2000}

+@article{fixation,

+ author = {Over, Eelco A. B.

+and Hooge, Ignace T. C.

+and Erkelens, Casper J.},

+ day = {01},

+ doi = {10.3758/BF03192777},

+ issn = {1554-3528},

+ journal = {Behavior Research Methods},

+ month = {May},

+ note = {\url{https://doi.org/10.3758/BF03192777}},

+ number = {2},

+ pages = {251--261},
+ title = {A Quantitative Measure For The Uniformity Of Fixation Density: The Voronoi Method},
+ volume = {38},
+ year = {2006}
@misc{forrester,
author = {Dan Ellis and
Tomás Sherwen},
```

-1823,7 +1840,7

```
month = {February},
note = {\url{https://doi.org/10.5281/zenodo.3676008}},
publisher = {Zenodo},
- title = {Geoschem/Geos-Chem: Geos-Chem 12.7.1},
+ title = {GEOSChem/Geos-Chem: Geos-Chem 12.7.1},
version = {12.7.1},
year = {2020}
```

-2017,7 +2034,7

```
@article{hufftree,
- autor = {{Sad CRUD Developer}},
+ author = {{Sad CRUD Developer}},
journal = {{StackOverflow}},
note = {\url{https://i.stack.imgur.com/9T1Am.png}},
title = {{The Huffman Tree}},
```

-2200,7 +2217,7

```
note = {\url{https://www.atmos-chem-phys.net/15/11433/2015/}},
number = {20},
pages = {11433--11459},
- title = {The Mcm V3.3.1 Degradation Scheme For Isoprene},
+ title = {The MCM V3.3.1 Degradation Scheme For Isoprene},
volume = {15},
year = {2015}
```

-2297,7 +2314,7

```
note = {\url{https://www.atmos-chem-phys.net/6/187/2006/}},
number = {1},
pages = {187--195},
- title = {Technical Note: Simulating Chemical Systems In Fortran90 And Matlab With The Kinetic Preprocessor Kpp-2.1},

+ title = {Technical Note: Simulating Chemical Systems In Fortran90 And Matlab With The Kinetic Preprocessor KPP-2.1},

volume = {6},

year = {2006}
```

-2607,7 +2624,17

```
@misc{mcm,

author = {Andrew Rickard},

note = {\url{http://mcm.york.ac.uk/}},

- title = {{Mcm Website}},

+ title = {{MCM Website}},

+ year = {2020}

+@article{mcmblue,

+ author = {Dan Ellis},

+ doi = {10.5281/zenodo.4294816},

+ note = {\url{https://doi.org/10.5281/zenodo.4294816}},

+ publisher = {Zenodo},

+ title = {Wolfiex/MCM-Blueprint: Thesisref},

+ version = {v0.0.3},

year = {2020}
```

-2627,7 +2654,7

```
author = {Mike Jenkins},

howpublished = {slide deck},

note = {Presentation for the EPSR group, Imperial Collage},

- title = {{History Of The Master Chemical Mechanism (Mcm) And Its Development Protocols}},

+ title = {{History Of The Master Chemical Mechanism (MCM) And Its Development Protocols}},

year = {2002}
```

-2655,7 +2682,7

```
pages = {161-180},

pdf = {https://hal.archives-ouvertes.fr/hal-00295229/file/acp-3-161-2003.pdf},

publisher = {{European Geosciences Union}},

- title = {{Protocol For The Development Of The Master Chemical Mechanism, Mcm V3 (Part A): Tropospheric Degradation Of Non-Aromatic Volatile Organic Compounds}},

+ title = {{Protocol For The Development Of The Master Chemical Mechanism, MCM V3 (Part A):
```

```
Tropospheric Degradation Of Non-Aromatic Volatile Organic Compounds}},

volume = {3},

year = {2003}
```

-2667,7 +2694,7

```
note = {\url{https://www.atmos-chem-phys.net/3/181/2003/}},

number = {1},

pages = {181--193},

- title = {Protocol For The Development Of The Master Chemical Mechanism, Mcm V3 (Part B):
Tropospheric Degradation Of Aromatic Volatile Organic Compounds},

+ title = {Protocol For The Development Of The Master Chemical Mechanism, MCM V3 (Part B):
Tropospheric Degradation Of Aromatic Volatile Organic Compounds},

volume = {3},

year = {2003}
```

-2790,6 +2817,14

```
year = {2019}

+@misc{montreal,

+ author = {UNEP},

+ note = {\url{https://ozone.unep.org/treaties/montreal-protocol}},

+ title = {{The Montreal Protocol On Substances That Deplete The Ozone

+Layer}},

+ year = {1987}

@inproceedings{mosaic,

address = {New York, NY},

author = {Hartigan, J. A.
```

-3220,6 +3255,7

```
@misc{numpy,

author = {Oliphant, Travis},

month = {01},

+ note = {\url{https://docs.scipy.org/doc/_static/numpybook.pdf}},

pages = {},

title = {Guide To Numpy},

year = {2006}
```

-3398,23 +3434,6

```
year = {2020}

-@article{Over2006,
- author = {Over, Eelco A. B.
-and Hooge, Ignace T. C.
-and Erkelens, Casper J.},
- day = {01},
- doi = {10.3758/BF03192777},
- issn = {1554-3528},
- journal = {Behavior Research Methods},
- month = {May},
- note = {\url{https://doi.org/10.3758/BF03192777}},
- number = {2},
- pages = {251--261},
- title = {A Quantitative Measure For The Uniformity Of Fixation Density: The Voronoi Method},
- volume = {38},
- year = {2006}
@article{oxidation,
author = {Planavsky, Noah J and Asael, Dan and Hofmann, Axel and Reinhard,
Christopher T and Lalonde, Stefan V and Knudsen, Andrew and Wang,
```

-3671,6 +3690,16

```
year = {2013}
+@article{physapprox,
+ author = {Valentin N. Ostrovsky},
+ journal = {Hyle},
+ number = {2},
+ pages = {101--126},
+ title = {Towards A Philosophy Of Approximations In The 'Exact' Sciences},
+ volume = {11},
+ year = {2005}
@article{pilot,
author = {Jeanningros, Y and Vlaeminck, S E and Kaldate, A and Verstraete,
W and Graveleau, L},
```

-3715,6 +3744,17

```
year = {2018}
+@book{platoform,
+ author = {Welton, W.A. and Benso, S. and Bowery, A.M.},
+ isbn = {9780739105146},
+ lccn = {2002117245},
+ note = {\url{https://books.google.co.uk/books?id=vbtbQk_A0YoC}},
+ publisher = {Lexington Books},
+ series = {G - Reference, Information and Interdisciplinary Subjects Series},
+ title = {Plato'S Forms: Varieties Of Interpretation},
+ year = {2002}
@article{plexp,
annote = {doi: 10.1137/070710111},
author = {Clauset, Aaron and Shalizi, Cosma Rohilla and Newman, M E J},
```

-3921,7 +3961,7

```
@article{sampling,
- author = { M. D. Mckay and R. J. Beckman and W. J. Conover },
+ author = { M. D. McKay and R. J. Beckman and W. J. Conover },
doi = {10.1080/00401706.2000.10485979},
journal = {Technometrics},
note = {\url{https://amstat.tandfonline.com/doi/abs/10.1080/00401706.2000.10485979}},
```

-4067,6 +4107,17

```
year = {1949}
+@book{serendipity,
+ author = {Roberts, R.M.},
+ isbn = {9780471602033},
+ lccn = {lc88033638},
+ note = {\url{https://books.google.co.uk/books?id=hf57X0s4aPwC}},
+ publisher = {Wiley},
+ series = {Wiley Science Editions},
+ title = {Serendipity: Accidental Discoveries In Science},
+ year = {1989}
@article{shapinginfo,
acmid = {1477423},
address = {Piscataway, NJ, USA},
```

-4625,7 +4676,7

```
author = {Br{"a}uer, Peter},
institution = {Github},
note = {\url{https://github.com/pb866/TUV_DSMACC}},
- title = {{Tuv 5.2X Dsmacc}},
+ title = {{TUV 5.2X DSMACC}},
year = {2020}
```