# DanEllisThesis - Change Log

## Thesis.tex

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

index 992121e..736606d 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 +325,25

% Introduction 0

\include{./0_intro}

- \include{./1_visual}

-% Chapter 2

-\include{./2_graphs}

+% % Ch1 1

+% \include{./1_visual}

+% % Chapter 2

+% \include{./2_graphs}

% Chapter 3

\include{./3_centrality}

-% Chapter 4 - done

- \include{./4_lumping}

+% % Chapter 4 - done

+% \include{./4_lumping}

+% % Chapter 5

+% \include{./5_DR}

+% % Conclusion

+% \include{./6_conclusion}

-% Chapter 5

- \include{./5_DR}

-% Conclusion

- \include{./6_conclusion}

\cleardoublepage\makeatletter@openrightfalse\makeatother

\begin{appendices}

## -357,7 +362,7

% \bibliographystyle{apalike}

% \bibliography{bibtex}

-%% \bibliographystyle{unsrt}

+% \bibliographystyle{unsrt}

### 0_intro.tex

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

index 26376fe..f75ed41 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}

-89,19 +89,19

\label{eqn:oo3}

\end{equation}

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

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

-\subsection{The Nox Cycle}\label{sec:noxcycle}

-Ozone production/loss in the troposphere is directly dependant on the concentration of available Nitrogen Oxides (NOx) (\autoref{sec:o3prod}). These are predominantly emitted by motor vehicles and power stations and can are known to cause respiratory problems in children and asthmatics as well as disrupting terrestrial and aquatic ecosystems \citep{eea}. Although NOx may be released naturally, the anthropogenic influence on their emissions was highlighted in early 2020 where the COVID-19 coronavirus disrupted travel across mainland china, causing a significant drop in anthropogenic emissions - \autoref{fig:chinanox}.

+\subsection{The NOx Cycle}\label{sec:noxcycle}

+Ozone production/loss in the troposphere is directly dependant on the concentration of available Nitrogen Oxides (NOx) (\autoref{sec:o3prod}). These are predominantly emitted by motor vehicles and power stations and can are known to cause respiratory problems in children and asthmatics as well as disrupting terrestrial and aquatic ecosystems \citep{eea}. Although NOx may be released naturally, the anthropogenic influence on their emissions was highlighted in early 2020 where the COVID-19 coronavirus disrupted travel across mainland China, causing a significant drop in anthropogenic emissions - \autoref{fig:chinanox}.

\begin{figure}[H]

\centering

\includegraphics[width=0.7\textwidth]{china_trop_2020056.png}

- \caption{\textbf{Changes in NOx concentrations due to anthropogenic emissions.} A reduction in activity and trasport resutls in a large decrease of Nitrogen dioxide concentrations in the troposphere. Source: \citep{chinanox}}

+ \caption{\textbf{Changes in NOx concentrations due to anthropogenic emissions.} A reduction in activity and trasport produces a notable decrease of Nitrogen dioxide concentrations in the troposphere. Source: \citep{chinanox}}

\label{fig:chinanox}

\end{figure}

-During the day nitrate (\ce{NO3}) radicals can be formed through the reaction with \ch{O3}: \autoref{eqn:ono2} and \autoref{eqn:nno2}, however this is quickly destroyed through rapid photolysis (\autoref{eqn:nno3}) \citep{nitrate}. Photolysis reactions such as \autoref{eqn:nno3} and \autoref{eqn:o2} are no longer possible and th ozone production process shuts down.

+During the day nitrate (\ce{NO3}) radicals can be formed through the reaction with \ch{O3}: \autoref{eqn:ono2} and \autoref{eqn:nno2}, however this is quickly destroyed through rapid photolysis (\autoref{eqn:nno3}) \citep{nitrate}. At night photolysis reactions such as \autoref{eqn:nno3} and \autoref{eqn:o2} are no longer possible and the ozone production process shuts down.

\begin{equation}

\ce{NO2 + O3 ->[k3] NO3 + O2}

## -109,12 +109,12

\end{equation}

\begin{equation}

- \ce{NO3 ->[hv] NO2 + O(3P)}

+ \ce{NO3 ->[hv] NO2 + O}(^3P)

\label{eqn:nno3}

\end{equation}

- The increased amount of \ce{NO3} can now react with \ce{NO2} to produce dinitric pentoxide (\ce{N2O5}) and an aqueous nitric acid (\ce{HNO3}) - \autoref{eqn:n2o5} and \autoref{eqn:hno3}. \autoref{eqn:n2o5} is a three-body forwards pressure dependant reaction and a reverse temperature dependant reaction. During the day at the lower troposphere, it is warm, and this can occur within seconds, however, at night or high altitudes it can take anywhere from hours to months

\citep{fundamentals}.

+ The increased amount of \ce{NO3} can now react with \ce{NO2} to produce dinitric pentoxide (\ce{N2O5}) and (in solution) nitric acid (\ce{HNO3}) - \autoref{eqn:n2o5} and \autoref{eqn:hno3}. \autoref{eqn:n2o5} is a three-body forwards pressure dependant reaction and a reverse temperature dependant reaction. During the day at the lower troposphere, it is warm, and the reverse reaction can occur within seconds, however, at night or high altitudes it can take anywhere from hours to months \citep{fundamentals}.

\begin{equation}

-129,31 +129,48

-\subsection{Hox Cycle}

-The hydroxyl (OH) radical is central to tropospheric chemistry and a major sink for many of the greenhouse gasses (including ozone) \citep{olson}. Its primary source of production is through the action of UV in sunlight to photolyse ozone \citep{fundamentals}:

+\subsection{HOx Cycle}

+The hydroxyl (OH) radical is central to tropospheric chemistry and a major sink for many of the greenhouse gasses (including ozone - see \autoref{eqn:o3sink}) \citep{olson}. Its primary source of production is through the action of UV in sunlight to photolyse ozone \citep{fundamentals}:

+\begin{align}

+ \ce{O3 ->[hv] O}^1D\

+ O^1\ce{D + H2O -> 2OH}\

+ \ce{OH + O3 -> HO2 + O2} \label{eqn:o3sink}

+\end{align}

+As OH is highly reactive , with a lifetime of {{content}}lt;1$ seconds - \autoref{fig:timescales}, it is not transported a long distance and only exists during daytime (when it is still being produced). In reacting with a VOC, the hydroxyl radical scavenges hydrogen to form a radical species and water (\ch{h2o}). This produced radical species can then move on to react with \ch{O2} to produce a \ch{RO2} species \autoref{eqn:rdo2}.%(\autoref{fig:hox}).

-\begin{equation}

- \ce{O3 ->[hv] O1D}

-\end{equation}

-\begin{equation}

- \ce{O1D + H2O -> 2OH}

-\end{equation}

-As OH is highly reactive ({{content}}lt;1$ seconds - \autoref{fig:timescales}) it is not transported a long distance and only exists during daytime (when it is still being produced). In reacting with a VOC, the hydroxyl radical scavenges hydrogen to form a radical species and water (\ch{h2o}). This produced radical species can then move on to react with \ch{O2} to produce a \ch{RO2} species (\autoref{fig:hox}). Additionally, reaction with OH can lead to the catalytic destruction of \ch{o3}. This provides the hydroperoxide radical (\ch{HO2}).

-\begin{equation}

- \ce{OH + O3 -> HO2 + O2}

-\end{equation}

-Unlike OH, \ch{HO2} can exist both during daytime and night. It can further react with ozone to reproduce the hydroxyl radical and create two \ch{o2} molecules:

+\begin{align}

+ OH + RH \ce{ -> } R. + \ch{h2o} \

+ R. + \ch{o2} \ce{ -> } \ch{RO2} \label{eqn:rdo2}\

+ \ch{ro2} + NO \ce{ -> } RO + \ch{no2}

+\end{align}

+The created \ce{RO2} can then convert NO to \ch{NO2} producing an \ch{RO} (\autoref{eqn:ro2no}) which also does the same via the hydroperoxide radical \ch{HO2} (\autoref{eqn:ro_1} - \ref{eqn:ro_2}). This NOx conversion is able to drive Ozone formation in the conventional way: \autoref{eqn:o1}-\ref{eqn:o3}.

+ \begin{align}

+ \ch{ro2} + NO \ce{ -> } RO + \ch{no2} \label{eqn:ro2no} \

+ RO + \ce{O2} \ce{ -> } \ch{HO2} \label{eqn:ro_1} \

+ \ce{HO2} + NO \ce{ -> } \ch{NO2} + OH \label{eqn:ro_2}

+\end{align}

+\subsubsection{The Hydroperoxide Radical}

+Unlike OH, \ch{HO2} can exist both during daytime and night. It can further react with ozone to reproduce the hydroxyl radical and create two \ch{o2} molecules - \autoref{eqn:ho22o2}.

\begin{equation}

- \ce{HO2 + O3 -> OH + O2 + O2}

+ \ce{HO2 + O3 -> OH + O2 + O2} \label{eqn:ho22o2}

\end{equation}

-Its loss depends on the NO mixing ratio, where if NO {{content}}gt;10$ pptv, \ch{HO2} will react predominantly with the NOx species. At lower concentrations (3-10 pptv) \ch{HO2} reacts mainly with ozone, and at deficient concentrations, it reacts mostly with itself \citep{finlayson}. Combined OH and \ch{ho2} form the HOx species, and the cycle in \autoref{fig:hox}.

+The loss of ozone loss depends on the NO mixing ratio, where if NO {{content}}gt;10$ pptv, \ch{HO2} will react predominantly with the NOx species. At lower concentrations (3-10 pptv) \ch{HO2} reacts mainly with ozone, and at deficient concentrations, it reacts mostly with itself \citep{finlayson}. Combined OH and \ch{ho2} form the HOx species, and the cycle in \autoref{fig:hox}.

\begin{figure}[H]

\centering

## -165,15 +182,16

\begin{figure}[H]

\centering

\includegraphics[width=0.7\textwidth]{hoxcycle.png}

- \caption{\textbf{The HOx cycle.} The OH aids in the oxidation of VOCs, which makes them more water soluble - this allowing for their removal from the atmosphere. In a high NOx evironment the \ch{RO2} radicals can then reacti with NO to produce \ch{NO2}, and consequenty more ozone.}

+ \caption{\textbf{The HOx cycle.} The OH aids in the oxidation of VOCs, which makes them more water soluble - this allowing for their removal from the atmosphere. In a high NOx evironment the \ch{RO2} radicals can then reactive with NO to produce \ch{NO2}, and consequenty more ozone.}

\label{fig:hox}

\end{figure}

\section{Modelling The Earth}

In the previous section, the air quality and its detrimental effects on human health were seen to influence policy for cities and industry.

For a policy to be passed there needs to not only evidence of the problem but a strong suggestion that any proposed changes will have the desired effect. As it is not possible to perform experiments on complex, and often unknown, chemistry at every location on the planet, we are forced to rely on the numerical simulation of the Earth System, and the constituent parts within it.

\subsection{Earth System Models (ESM)}

- ESMs are models capable of predict past or future interactions of the planetary system. They represent our foremost understanding of the complex interplay between land-surface (geosphere), ocean (hydrosphere), ice (cryosphere) and the air (atmosphere), and act as a surrogate to manual experimentation - which is just not possible on the global scale.

+ ESMs are models capable of predicting past or future interactions of the planetary system. They represent our foremost understanding of the complex interplay between land-surface (geosphere), ocean (hydrosphere), ice (cryosphere) and the air (atmosphere), and act as a surrogate to manual experimentation - which is just not possible on the global scale.

ESMs can be split into individual parts. One example of this is the Chemistry section of the Goddard Earth Observing System (an integrated ESM and data assimilation model hosted by NASA's Goddard space flight centre \citep{geosgit}) - GEOS Chem. GEOS-Chem is a global 3D model of atmospheric chemistry which is driven by the meteorology provided by NASA \citep{geos}. Here the Earth is split up into cubic cells longitudinally, latitudinally, and vertically (\autoref{fig:gcm})\footnote{This image is not from GEOS-Chem.}. Each one of these cells performs several perturbations of the chemistry within them before any long-lived species are transported, and the process is repeated. If extracted separately, a single one of these cells may be used to explore the sensitivity of different species for a range of input conditions. This is the bases of the atmospheric box model.

\begin{figure}

\centering

-199,11 +217,11

\label{eqn:numerical1}

\end{equation}

\begin{equation}

-\ce{ d[N2O5]/dt -> d[NO2]/dt} + \ce{d[NO3]/dt}

+\ce{- d[N2O5]/dt = d[NO2]/dt} + \ce{d[NO3]/dt}

\label{eqn:numerical2}

\end{equation}

\begin{equation}

-\ce{ \int d[N2O5]/dt -> \int d[NO2]/dt} + \ce{\int d[NO3]/dt}

+\ce{ - \int d[N2O5]/dt = \int d[NO2]/dt} + \ce{\int d[NO3]/dt}

\label{eqn:numerical3}

\end{equation}

\subsubsection{Non-Stiff Equations}

-230,7 +248,7

\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..61c4282 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}

## -210,8 +210,7

\textit{It is worth noting that segment sizes do not represent the number of species undergoing a specific reaction pathway, but rather the percentage of all possible pathways which follow that route. This is because species often undergo a range of reactions, each of which counts as an individual weighting. It is for this reason that even though almost all\footnote{ Except for any inorganic species.} contain a C-H bond, hydrogen abstraction does not consume the whole graph. Many species have multiple possible pathways in which they may react, and the chord diagram presents the likeliness of a rection for all possible methods of reaction for all species.

-From this, we see that hydroxy reactions are the most common with C-H bonds being in abundance\footnote{This is seen within the graph layout \autoref{fig:mcmfull}}. We also see that having another type of reaction is also just as probable, with a third of the most utilised branches within the MCM protocol falling under species containing at least one Carbonyl group.

-Next, we look at the co-occurrence of branches for different species. These are represented using the area of a circle connecting two arcs (a chord). Each chord has two edges connecting two arcs\footnote{ except for self-loops, although these are addressed below.}. It is possible to discern the percentage of items going between these and other branches by comparing the width of each chord to its parent arc. Here, for example, we see a roughly even split between species with a C-H bond (i.e. all species) and every other group. This suggests an even distribution of reaction types between species.

+From this, we see that hydroxy reactions are the most common with C-H bonds being in abundance\footnote{This is seen within the graph layout \autoref{fig:mcmfull}}. Additionally we find that when applying the MCM protocol, a third of species contain at least one carbonyl group. Next, we look at the co-occurrence of branches for different species. These are represented using the area of a circle connecting two arcs (a chord). Each chord has two edges connecting two arcs\footnote{ except for self-loops, although these are addressed below.}. It is possible to discern the percentage of items going between these and other branches by comparing the width of each chord to its parent arc. Here, for example, we see a roughly even split between species with a C-H bond (i.e. all species) and every other group. This suggests an even distribution of reaction types between species.

This means that in comparing the arc length of each chord, we can visually determine the percentage of group A which relates to its partner group B. Finally it is also possible to determine the number of items in a group which contain themselves. Chemically these are species with multiples of one functional group that undergo a specific reaction pathway more than one time. Although these reactions will usually be combined within a mechanism (to avoid duplication), their

rate would be increased accordingly.

\begin{figure}[H]

## -258,9 +257,9

\autoref{fig:ho2} and \autoref{fig:oh} show arc diagrams where the reactions of interest (photolysis and OH reactions respectively) highlighted in both colour and opacity. These enable us to see patterns between the radical cylcing of \ce{OH -> HO2} chemistry (\autoref{fig:rxnho2oh}). Here the cyclic reaction shown between the dashed lines corresponds to the reaction of \ch{RO2 <-> [HO2][O2] ROOH} (\autoref{fig:rxnho2oh}).

-Applying the same methodology to photolysis and hydroxide reactions, the production of species containing fewer functional groups is seen in \autoref{fig:ohhv}. Within the highlighted reactions, it is seen that a ROOH species undergoes a reaction with OH or photolyses (\autoref{fig:rxnohhv}). In the OH reaction, Hydrogen abstraction is performed to produce an RO2 species and water, \ch{ROOH ->[OH] RO2 + H2O}. Photolysis reactions, however, photolyse the double bond, \ch{ROOH ->[hv] RO2 + HO2}, reducing the number of functional groups - producing a larger arc. It should be mentioned that the ROOH can also react with \ch{o2} to produce an RO2, although this has not been highlighted.

+Applying the same methodology to photolysis and hydroxy reactions, the production of species containing fewer functional groups is seen in \autoref{fig:ohhv}. Within the highlighted reactions, it is seen that a ROOH species undergoes a reaction with OH or photolyses (\autoref{fig:rxnohhv}). In the OH reaction, Hydrogen abstraction is performed to produce an \ch{RO2} species and water, \ch{ROOH ->[OH] RO2 + H2O}. Photolysis reactions, however, photolyse, \ch{ROOH ->[hv] RO2 + HO2}, reducing the number of functional groups - producing a larger arc.

-Finally, Peroxy Acetyl Nitrates (PANs), play a vital role in the modelling of photochemical smog (ozone events), \citep{pans}. PANs an effective reservoir species with significant importance within the production of ozone in atmospheric chemistry models (especially if transportation is involved) \citep{finlayson}. Although they are very stable at cold temperatures, these can quickly decompose (thermally) to release $NO_x$ if warmed. In the MCM the thermal decomposition of PANS is determined by the KBPAN rate constant. In comparing reactions of \autoref{fig:kbpans}, with those of \autoref{fig:no2} (at rate KFPAN), we see a cycle between two arcs forming (\autoref{fig:pandir}). This can be explained by the reactions in \autoref{fig:rxnpan} which show that \ch{RC(O)OONO2 ->[KBPAN] RC(O)O2} (+\ch{NO2}) \ch{ ->[NO2] RC(O)OONO2}.

+Finally, Peroxy Acetyl Nitrates (PANs), play a vital role in the modelling of photochemical smog (ozone events), \citep{pans}. PANs an effective reservoir species with significant importance within the production of ozone in atmospheric chemistry models (especially if transportation is involved) \citep{finlayson}. Although they are very stable at cold temperatures, these can quickly decompose (thermally) to release $NO_x$ if warmed. In the MCM the thermal decomposition of PANS is determined by the KBPAN rate constant. In comparing reactions of \autoref{fig:kbpan}, with those of \autoref{fig:no2} (at rate KFPAN), we see a cycle between two arcs forming (\autoref{fig:pansdir}). This can be explained by the reactions in \autoref{fig:rxnpan} which show that \ch{RC(O)OONO2 ->[KBPAN] RC(O)O2} (+\ch{NO2}) \ch{ ->[NO2] RC(O)OONO2}.

\textit{\textbf{NOTE:} A downside to the arc diagrams format that has been chosen is that for reactions between species of the same number of functional groups, there is no set direction. }

## -350,7 +349,7

\caption{Hydroxide}

\label{fig:oh2}

\end{subfigure}

- \caption{\textbf{ Arc diagram features for photolysis and hydroxide. reactions. } Photolysis results in species with a reduced number of functional groups, and therefore longer arcs. OH reactions for the same species do not produce such a drastic change on group number, and therefore have a smaller arc lenght.}

+ \caption{\textbf{ Arc diagram features for photolysis and hydroxide reactions. } Photolysis results in species with a reduced number of functional groups, and therefore longer arcs. OH reactions for the same species do not produce such a drastic change on group number, and therefore have a smaller arc length.}

\label{fig:wholeohhv}

\end{figure}

## -368,7 +367,7

```
\centering

\scalebox{.7}{

\schemestart [0,1,thick]

- \chemfig{R-[:30]O-[:-30]O-[:30]N(=[:90]O^{-})=[:-30]O}

+ \chemfig{R-[:30]O-[:-30]O-[:30]N(=[:90]O)=[:-30]O}

\arrow{->[\ce{}][][][.5][]}

\chemfig{R-[:30]O-[:-30]O_{.}}

\arrow{0}[,0] \chemfig{+ \ce{NO2}}
```

## -404,24 +403,69

```
+\newpage

\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 we show the growth of the MCM (left) against two versions (three variations) of the reduced Common Representative Intermediates (CRI) \citep{cri} mechanism in the same space. By fixing species which exist in mechanisms groups (generally the primary emitted VOCs) we produce a 'fingerprint'-like structure we can use to visually identify changes in their size, interconnectedness (density) and structure.

+Building on this, an ineractive visualisation (\autoref{fig:mcmchange}) was constructed to better reveal the differences between of each mechanism in (\autoref{fig:graphc1}). The code for this can be found in \citep{mcmblue}.

+\autoref{fig:m1to2} shows the expansion from MCM version 3.1 to 3.2 which included new schemes for crotonaldehyde, ethylene oxide and vinyl chloride, the introduction of methacolein and the integration of dimethyl sulphide (DMS), beta-caryophyllene and limonene \citep{mcm} - the latter of which is responsible for the additional South-West pointing branch seen within the graph representations. Similarly \autoref{fig:m2to3} shows the upgrade from MCM v3.2 to v3.3.1, the main change is the mechanism update to include the complete degregation mechanism for isoprene \citep{isopmcm}. This change results in the addition of ~100 species, many of which are mainly realated to OH initiated chemistry. However since the ratio of species to links (reactions) has now increased, these lie closer to the main body of the network - the reason for which is discussed in \autoref{ch3}.

+Similarly we can use \autoref{fig:mcmchange} to emphasise the amount that has been added (or lost) in reduction or development. \autoref{fig:m2tocri1} shows the difference between the MCM v3.2 and its reduced CRI v2.0 form, which focuses on preserving the overall ozone-forming potential of the mechanism. Simialrly \autoref{fig:cr1tocr5} shows a comparison of the CRI v2.0 after a further 5 reductions (CRI v2.0 r1). Using these two plots we can identify regions or branches of chemsitry which have been removed (namely bigonic and anthropogenic aromatic branches - bottom left and bottom right) and generate a an overview of how well the reduced mechanism structre represents all parts of the contained chemistry. We can see that on avarage the CRI mechanism does a good job at retaining the core network structure, often lumping the more esoteric (or extreme) branches into a single species at their base.

+This type of network representation is found not only simplest and most intuitve, but also the most
```

informative about what effects changing the underlying chemistry may have on a simulation. \autoref{ch2} expands on the sociograph idea, and explors the different ways in which we may tune it to maximise its potential for useful knowledge transfer.

\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_m32.png}

+ \caption{MCM v3.2 vs CRI v2.0(r1)}

+ \label{fig:m2tocri1}

+ \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}

+ \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 (d) 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}

+\newpage

\section{Conclusion}

## 2_graphs.tex

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

index f6c9ce1..c585f13 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.

```
-173,7 +173,7
```

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

\caption{Mercator}

\end{subfigure}

- \caption{\textbf{A selection of map projections.} These have been created using DataDrivenDocuments \citep{d3} and show a range of methods for mapping the spheroid shape of the Earth onto a 2D plane. }

+ \caption{\textbf{A selection of map projections.} These have been created using DataDrivenDocuments \citep{d3js} and show a range of methods for mapping the spheroid shape of the Earth onto a 2D plane. }

\label{fig:projections}

\end{figure}

```
-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}

## -630,7 +630,7

\subsubsection{Edge Bundling }

-Pioneered by \cite{bundlepioneer}, edge bundling techniques are an effective way to reduce visual clutter. Much like a force graph, edges are represented as a string of lined points. This allows for edges to be pulled together (attracted to one another) and produces a visualisation akin to moving water droplets on a hydrophobic surface. \autoref{fig:edgebundling} shows how in changing the amount of attraction between edges, it is possible to reduce clutter in a visualisation.

+Pioneered by \cite{edgebundle}, edge bundling techniques are an effective way to reduce visual clutter. Much like a force graph, edges are represented as a string of lined points. This allows for edges to be pulled together (attracted to one another) and produces a visualisation akin to moving water droplets on a hydrophobic surface. \autoref{fig:edgebundling} shows how in changing the amount of attraction between edges, it is possible to reduce clutter in a visualisation.

\begin{figure}[H]

## -791,13 +791,13

\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}

\subsection{A Model Of Beijing}

-To perform a sensitivity study on the initial positions of nodes within the force atlas algorithm a graph consisting of links and weightings is constructed using a box model simulation of the Beijing summer environment (mid-day) and feed it the gephi software \citep{gephi} - an open-source software designed for the exploration of networks. We then script the java code to perform the functions in \autoref{fig:flowrepeat}. As part of this, nodes are initiated with a random position; the ForceAtlas2 layout is then run and then the graph is rotated and translated such that it is centred around carbon monoxide and has a 45-degree angle between this and formaldehyde. This step constrains the general orientation of the graph, allowing us to analyse the generated graphs

for global and local minima. The final step is to save a copy of the generated graph layout and repeat to generate a data set, a subset of which is shown in \autoref{fig:all}. These are discussed further in \autoref{sec:patternmatch}.

+To perform a sensitivity study on the initial positions of nodes within the force atlas algorithm a graph consisting of links and weightings is constructed using a box model simulation of the Beijing summer environment (mid-day) and feed it the gephi software \citep{gephi} - an open-source software designed for the exploration of networks. We then script the java code to perform the functions in \autoref{fig:flowrepeat}. As part of this, nodes are initiated with a random position; the ForceAtlas2 layout is then run and then the graph is rotated and translated such that it is centred around carbon monoxide and has a 45-degree angle between this and formaldehyde. This step constrains the general orientation of the graph, allowing us to analyse the generated graphs for global and local minima. The final step is to save a copy of the generated graph layout and repeat to generate a data set, a subset of which is shown in \autoref{fig:allsamples}. These are discussed further in \autoref{sec:patternmatch}.

\begin{figure}[H]

\centering

-810,21 +810,21

\centering

\includegraphics[width=\textwidth]{figures_c1/beijingtest/10_900.png}

\caption{\textbf{A sample of 224 (out of the 2000) graphs generated using the ForceAtlas2 algorithm.} These represent the conditions of a spun up simulation of Beijing at noon. The shapes of each graph, and general shapes are discussed in \autoref{sec:patternmatch} and \autoref{sec:netshape}.}

- \label{fig:all}

+ \label{fig:allsamples}

\end{figure}

\subsubsection{Similarity Between Graph Shape}\label{sec:patternmatch}

-Although through the use of manual intervention, it is possible to perform a superficial level of shape analysis, our cognitive capabilities do not allow us to perform this task for all the simulations of \autoref{fig:all}- less so the entire 2000 graphs in the dataset. To overcome this problem, we rely on a method of machine learning called t-Distributed Stochastic Neighbor Embedding (t-SNE) - described in \autoref{sec:overcrowd} and is the foundation of the tsNET layout algorithm. This is a dimensionality reduction technique used in the automatic categorisation of images or photographs \citep{truthandbeauty,sketchy}.

+Although through the use of manual intervention, it is possible to perform a superficial level of shape analysis, our cognitive capabilities do not allow us to perform this task for all the simulations of \autoref{fig:allsamples}- less so the entire 2000 graphs in the dataset. To overcome this problem, we rely on a method of machine learning called t-Distributed Stochastic Neighbor Embedding (t-SNE) - described in \autoref{sec:overcrowd} and is the foundation of the tsNET layout algorithm. This is a dimensionality reduction technique used in the automatic categorisation of images or photographs \citep{truthandbeauty,sketchy}.

The input for the t-SNE for each dataset is a flattened (1 dimensional) representation of the pixels in the image - we start and by taking a binary matrix representing each image, split it up into rows, and glue these together. The pixelmap for each image is then fed into the t-SNE algorithm from the Scikit Learn package \citep{scikit-learn}. This reduces the logical list of pixels for each image into a two-dimensional representation of their similarity. We plot each file, for its $(x,y)$ coordinate, and isolate clusters of similarity using density contours in \autoref{fig:density}.

\begin{figure}[H]

\centering

\includegraphics[width=.6\textwidth]{figures_c1/beijingtest/density.png}

- \caption{\textbf{A normalised scatter plot of 2D space produced by the t-SNE algorithm.} Each triangle represents a different arrangement of the MCM nodes shown in \autoref{fig:all}, and the colours/density contours show the regions in which we find similar images/graphs. Cluster numbers correspond to the groups in \autoref{fig:densityfig}. }

+ \caption{\textbf{A normalised scatter plot of 2D space produced by the t-SNE algorithm.} Each triangle represents a different arrangement of the MCM nodes shown in \autoref{fig:allsamples},

and the colours/density contours show the regions in which we find similar images/graphs. Cluster numbers correspond to the groups in \autoref{fig:densitypic}. }

\label{fig:densty}

\end{figure}

## -860,7 +860,7

\subsubsection{Network Branch Classification}\label{sec:netshape}

-In \autoref{sec:patternmatch} it was seen that there exist a certain branch pattern that emerges from the structure of the MCM (\autoref{fig:densitypic}). Upon manual inspection of the simulations (\autoref{fig:all}) many graphs appear to contain three branches for each graph - using this it may be hypothesized that these are a result of the mechanism, and by consequence the chemistry it describes.

+In \autoref{sec:patternmatch} it was seen that there exist a certain branch pattern that emerges from the structure of the MCM (\autoref{fig:densitypic}). Upon manual inspection of the simulations (\autoref{fig:allsamples}) many graphs appear to contain three branches for each graph - using this it may be hypothesized that these are a result of the mechanism, and by consequence the chemistry it describes.

To test for this, we categorise all primary emitted species into Alkanes, Alkenes, Aromatics and Terpenes. All nodes and links in close proximity are regarded as products of these species and are placed within the same group. Using a randomly selected graph from the dataset, the network is separated spatially, and nodes within the Voronoi cell (These are described in \autoref{sec:nodedensitya}) of a primary emitted species are coloured similarly.

## -869,7 +869,7

\begin{figure}[H]

\centering

\includegraphics[width=\textwidth,trim={0 4cm 0 4cm},angle=-90]
{figures_c1/beijingtest/graphgroups.pdf}

- \caption{\textbf{Highlighting the groups of species, and their products within one of the MCM network graphs from \autoref{fig:all}} These are {\color{DarkGoldenrod} Aromatics (gold)} , {\color{DarkTurquoise} Terpenes (turquoise) } and {\color{OrangeRed} Alkane}/{\color{RoyalBlue} Alkene } carbon chains (red/blue)}

+ \caption{\textbf{Highlighting the groups of species, and their products within one of the MCM network graphs from \autoref{fig:allsamples}} These are {\color{DarkGoldenrod} Aromatics (gold)} , {\color{DarkTurquoise} Terpenes (turquoise) } and {\color{OrangeRed} Alkane}/{\color{RoyalBlue} Alkene } carbon chains (red/blue)}

\label{fig:fncolour}

\end{figure}

### 3_centrality.tex

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

index c021b18..55bd30a 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

-40,13 +40,18

\subsection{The Master Chemical Mechanism (MCM)}\label{sec:metricmcm}

-The MCM, \citep{mcm}, is a near explicit representation of our foremost understanding of gas-phase tropospheric chemistry. The mechanism describes the oxidation of 143 primary emitted VOCs and the respective rates at which this occurs. It has been tested on over 300 chamber experiments and used as a benchmarking mechanism to assist the development of reduced mechanism, providing a useful means for the evaluation of air quality models \citep{defra1}.

+The MCM, \citep{mcm}, is a near explicit representation of our foremost understanding of gas-phase tropospheric chemistry. The mechanism describes the oxidation of 143 primary emitted VOCs and the respective rates at which this occurs. It has been tested on over 300 chamber experiments and used as a benchmarking mechanism to assist the development of reduced mechanism, providing a useful means for the evaluation of air quality models \citep{defra1}. The current version (3.3.1) contains 5809 chemical species and 17224 reactions to describe them \citep{isopmcm}. However there there are still a number of weaknesses that need to be considered. Firstly there very little \ce{Cl} chemistry and no other halogens in the mechanism. Reactions with \ce{O2} are implicit as are \ce{RO2}-\ce{RO2} reactions, which are shown through the reaction with an \ce{RO2} pool.

-Version 3.3.1 of the MCM contains 5809 chemical species and 17224 reactions to describe them \citep{isopmcm}.

+\subsection{Data Collection}\label{sec:scholar}

+To generate a dataset on papers related to the MCM. The academic search engine (Google Scholar \citep{scholar}) is queried for all articles containing the words { \emph{"Master", "Chemical", "Mechanism"} and \emph{"MCM"} }. For each match, the first 100 pages of results are selected. Each of these contains ten articles, from which the first 100 pages of related articles are chosen.

+In taking the top 1000 citations for each page, a network of 15744 papers and 30178 citations\footnote{Note: this had the potential of returning up to 1000,000 nodes} is created. This process made use of an edited version of the \emph{etudier} Github repository, \citep{web}.

\begin{figure}[H]

\centering

\includegraphics[width=0.92\textheight,angle=90]{figures_c3/naturegraph.png}

-56,13 +61,6

\end{figure}

-\subsection{Data Collection}\label{sec:scholar}

-To generate a dataset on papers related to the MCM. The academic search engine (Google Scholar \citep{scholar}) is queried for all articles containing the words { \emph{"Master", "Chemical", "Mechanism"} and \emph{"MCM"} }. For each match, the first 100 pages of results are selected. Each of these contains ten articles, from which the first 100 pages of related articles are chosen.

-In taking the top 1000 citations for each page, a network of 15744 papers and 30178 citations\footnote{Note: this had the potential of returning up to 1000,000 nodes} is created. This

process made use of an edited version of the \emph{etudier} Github repository, \citep{web}.

\subsection{Visualising The Data.}

The initial visualisation of the dataset is accomplished through the use of THREE.js \citep{threejs}. This makes use of WebGL bindings and allows for the efficient viewing, querying and interacting of the data in 3 dimensions. This helped identify the temporal changes within the network by mapping a papers publication year to the $z$ direction, \autoref{fig:weball}, as discussed in \autoref{sec:filter3d}.

## -110,7 +108,7

\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.

## -171,7 +169,7

\subsection{The Co-Authorship Network}

An alternative to exploring which papers which are cited together are to look at their authors. Here undirected links are drawn between authors on the same paper. This style of analysis was used to show that the number of papers per author, and the total number of authors per paper can vary between research fields, \citep{newmancoauthor}. In combining this with a series of network centrality metrics, \citep{coauthornew} revealed that it is possible to discern promising researchers from both iter and Intra disciplinary groups.

-In building a co-authorship network for the MCM, we can identify authors who publish together\footnote{ Disclaimer: as mentioned earlier, not all authors for every paper were recorded by the web scraping algorithm} and highlight research groups who work with the MCM, \autoref{fig:authorgroup}. This shows how authors with a similar geographic location/institution are more likely to publish together. The largest cluster here falls under the MCM developer team, which resides between the University of Leeds and York. Next two German institutions which are heavily involved in the atmospheric chemistry field (Julich and Max Planck), followed by an assortment of Chinese authors, mainly centred around the Beijing or Hong Kong region.

+In building a co-authorship network for the MCM, we can identify authors who publish together\footnote{ Disclaimer: as mentioned earlier, not all authors for every paper were recorded by the web scraping algorithm} and highlight research groups who work with the MCM, \autoref{fig:authorgroup}. This shows how authors with a similar geographic location/institution are more likely to publish together. The largest cluster here falls under the MCM developer team, which resides between the University of Leeds and York. Next two German institutions which are heavily involved in the atmospheric chemistry field (FZ-Julich and Max Planck for Chemistry, Mainz), followed by an assortment of Chinese authors, mainly centred around the Beijing or Hong Kong region.

\begin{figure}[H]

## -218,13 +216,13

\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 +248,7

\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 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 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 +354,7

This is repeated until a pre-defined tolerance, $\epsilon$ 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 +364,7

\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 +377,7

\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 +385,7

\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 +402,7

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 +427,7

\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 $\omega$ 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 $\omega$ 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 +459,7

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 +472,7

\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 +545,25

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. In general, the reaction consists 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 +603,7

\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}.

## -698,8 +696,8

After reversing the links, we see that concentration for the reaction between \ch{c2h6} and OH follow the paths:

\begin{eqnarray}

- \ce{OH ->[0.1] C2H6 ->[0.1] C2H5O2} \[20pt]

- \ce{C2H6 ->[2\times10^{-7}] OH ->[2\times10^{-7}] C2H5O2}

+ \ce{OH ->[C2H6][0.1] C2H5O2} \[20pt]

+ \ce{C2H6 ->[OH][2\times10^{-7}] C2H5O2}

\end{eqnarray}

## -729,7 +727,7

\section{Case Study}\label{sec:metriccase}

-In this section, the centrality metrics discussed in \autoref{sec:graphcentrality} are applied to a range of scenarios. These range from polluted urban environments such as London \citep{clfo} and Beijing \cite{aphh}, to marine and terrestrial forest- Cape Verde \cite{capeverde} and Borneo \citep{borneo}. We determine the main drivers for the chemistry and compare the species which are important across each simulation.

+In this section, the centrality metrics discussed in \autoref{sec:graphcentrality} are applied to a range of scenarios. These range from polluted urban environments such as London \citep{clfo} and Beijing \cite{aphh}, to marine and terrestrial forest- Cape Verde \citep{capeverde} and Borneo \citep{borneo}. We determine the main drivers for the chemistry and compare the species which are important across each simulation.

\subsection{Establishing Initial Conditions From Observational Data}

Within experimental data assimilation, it is not uncommon to face problems which result in unreliable or missing data. These can range from anything as little as measuring below the instrument sensitivity to powercuts and equipment damage/theft from the local wildlife. This can result in problems when analysing the results and combining them to create a simulation of the chemistry for that environment.

## -754,7 +752,7

\begin{quote}

\textit{

-\textbf{Example analogy:} Backpropagation can be likened to the iterative calibration of scientific instrumentation. In the field of atmospheric chemistry, laser-induced fluorescence is used to calculate species concentrations and reaction rates within the troposphere, \citep{lif1,lif2}. Here the frequency of a laser can be adjusted in contrast with a known target (e.g. an amount of \ce{SO2}) to produce a response curve showing where the maximum resonance occurs.\

+\textbf{Example analogy:} Backpropagation can be likened to the iterative calibration of scientific instrumentation. In the field of atmospheric chemistry, laser-induced fluorescence is used to measure species concentrations and reaction rates within the troposphere, \citep{lif1,lif2}. Here the frequency of a laser can be tuned to a resonant frequency of a known target (e.g. \ce{OH}, \ce{NO2} and \ce{SO2}) to produce a response curve.\

Similarly, a neural network can be 'trained' (calibrated).

This is done through the use of a training dataset' - a set of input-output pairings which represent a random selection of 2/3rds of the total dataset. Next, the neurons within each layer (similar to the potentiometer dials on an instrument) are adjusted in sequence through the layers to match the known result (a standard of known concentration) to the input values provided. This process is repeated until for many iterations, or until a sufficientlygood' prediction is attained for the entire training dataset (early termination). The power of ANNs comes from the ability to adjust neuron thresholds whilst moving both forwards and backwards through the network (Note: predictions of an MLP are still only passed forwards). Finally, model performance is evaluated against the remaining 1/3rd of the total dataset.

## -769,7 +767,7

\end{figure}

\subsubsection{Applying The Mlpregressor To Observational Data}

-In the application of any type of machine aided algorithms, it is important to evaluate the results

provided. In this section, the results of 12 years of data collected as part of the [CAPE VERDE CAMPAIGN] are shown (these contain measurements spanning the entirety of 12 years, which produce the clearest tests for the algorithm). A MLPRegressor of 10 hidden layers, and a hyperbolic tan (tanh) activation function is used \autoref{sec:appendix:tanh}. Additionally, the limited-memory Broyden-Fletcher-Goldfarb-Shanno (l-BFGS) solver (a quasi-newton method which minimises the inverse of the Hessian matrix\footnote{ The hessian is a square matrix of second-order partial derivatives of a scalar-valued function/field describing the local curvature of a function (of many variables).} to steer through space and obtain a solution) and an adaptive learning rate\footnote{Each time the model improvement fails to decrease the learning loss, the learning rate is reduced by 1/5. This means smaller jumps are made towards the curve peak. } is used.

+In the application of any type of machine aided algorithms, it is important to evaluate the results provided. In this section data collected from Cape Verde (\citep{capeverde}) containing 12 years of observations are shown. A MLPRegressor of 10 hidden layers, and a hyperbolic tan (tanh) activation function is used \autoref{apx:tanh}. Additionally, the limited-memory Broyden-Fletcher-Goldfarb-Shanno (l-BFGS) solver (a quasi-newton method which minimises the inverse of the Hessian matrix\footnote{ The hessian is a square matrix of second-order partial derivatives of a scalar-valued function/field describing the local curvature of a function (of many variables).} to steer through space and obtain a solution) and an adaptive learning rate\footnote{Each time the model improvement fails to decrease the learning loss, the learning rate is reduced by 1/5. This means smaller jumps are made towards the curve peak. } is used.

The input of the regressor is in the form of a month and an hour, to represent each measurement. This allows it to find not only daily trends but also seasonal trends within the data. Once trained, the regressor is then used to predict a diurnal profile for each month based on the observational data provided. For simplicity $\log_{10}$ values of the concentrations obtained have been used. The predicted MLPRegressor line is compared to a transparent scatterplot for all the results. In addition to this, a boxplot showing the Inter Quartile Range (The range between the 25th and 75th percentile), median and mean (green line) plotted alongside to evaluate the predictor output. In this study, we only take the values for the month of June (or closest available depending on the dataset).

## -806,18 +804,18

\subsubsection{Model Initialisation Procedure}

-The aim is to generate a set of initiation concentrations which are representative of the species found for different environments around the world. In this section, we are not interested in the exact concentration modelling for specific times or scenarios. Instead, we seek to generate representative of the processed chemistry under a range of conditions.

+The aim is to generate a set of initiation conditions which are representative of the species found for different environments around the world. In this section, we are not interested in the exact concentration modelling for specific times or scenarios. Instead, we seek to generate representative of the processed chemistry under a range of conditions.

Species concentrations are extracted from an MLP regressor trained on observational data for each scenario. Each concentration is that of noon local time from the generated diurnal from summer observations at each location. This produces a monthly error of $\pm 2 months$ from June. As both nitrogen oxide and dioxide are supplied, the total NO$_x$ for each simulation are \emph{not} constrained. The initial conditions are shown in \autoref{tab:icsmetric}.

-In general observational measurements are not able to detect all the species presented within the MCM. This means that to be able to compare model scenarios, the chemistry must first be spun up. In propagating the chemistry forwards in time, primarily emitted and measured species are broken up forming the intermediate species which exist within a mechanism. To reach a steady-state, the model is initiated at noon, and the observational concentrations are rest every 24 hours. For each diurnal, the fractional difference between the concentrations at each day are compared. If the difference between these is less than 0.001, the model is left to run unconstrained for five days (right of the dashed line in \multiref{fig:ccape}{fig:cbeijing}). Model results are then taken after three days of unconstrained runs. The reason for this is that the total RO$_2$ concentration takes longer to stabilise in the polluted environments (London and Beijing). This falls into a periodic cycle beginning noon on the third day and can provide a representation of the processed chemistry within each environment.

+In general observational measurements are not able to detect all the species presented within the MCM. This means that to be able to compare model scenarios, the chemistry must first be spun up. In propagating the chemistry forwards in time, primarily emitted and measured species are broken up forming the intermediate species which exist within a mechanism. To reach a steady-state, the model is initiated at noon, and the observational concentrations are rest every 24 hours. For each diurnal, the fractional difference between the concentrations at each day are compared. If the difference between these is less than 0.001, the model is left to run unconstrained for five days (right of the dashed line in \autoref{fig:cbeijing}). Model results are then taken after three days of unconstrained runs. The reason for this is that the total RO$_2$ concentration takes longer to stabilise in the polluted environments (London and Beijing). This falls into a periodic cycle beginning noon on the third day and can provide a representation of the processed chemistry within each environment.

\textit{NOTE: It should be noted that some of the concentration plots may appear to lose their diurnal dependability. This may be attributed to the changing order of magnitude of the

concentrations, and that the species are still responding as expected. }

\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}

## -832,7 +830,7

\begin{figure}[H]

\centering

\includegraphics[width=.9\textwidth]{figures_c3/mlpregressor/conc_clfo.pdf}

-\caption{\textbf{The concentration profile for London.}This shows a the change in concentration over time for HO$_x$,NO$_x$,Ozone and RO$_2$ species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the intial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .}

+\caption{\textbf{The mixing ratio profile for London.}This shows a the change in mixing ratio over time for HO$_x$,NO$_x$, HCHO,Ozone and RO$_2$ species for a simulation run generated by the mlpregressor. Left of the dashed line shows the last 6 days of spinup, where the values are reset at noon each day until the species fractional difference is less than 0.001 .}

\label{fig:clondon}

\end{figure}

## -842,7 +840,7

\begin{figure}[H]

\centering

\includegraphics[width=.9\textwidth]{figures_c3/mlpregressor/conc_beijing.pdf}

-\caption{\textbf{The concentration profile for Beijing.}This shows the change in concentration over time for HO$_x$, NO$_x$, Ozone and RO$_2$ species for a simulation run generated by the MLPRegressor. Left of the dashed line shows the last six days of spinup, where the initial concentrations are reset at noon each day until the species fractional difference is less than 0.001 .}\label{fig:cbeijing}

+\caption{\textbf{The mixing ratio profile for Beijing.}This shows the change in mixing ratio over time for HO$_x$, NO$_x$, HCHO, Ozone and RO$_2$ species for a simulation run generated by the MLPRegressor. Left of the dashed line shows the last six days of spinup, where the initial values are reset at noon each day until the species fractional difference is less than 0.001 .}\label{fig:cbeijing}

\end{figure}

\newpage

## -949,10 +947,10

Individual categories are split between traditional metrics and graph centrality metrics. To represent the importance of a species, the following values may be extracted through the use of a simple box model:

\begin{itemize}

-\item[-] \textbf{Concentration} - This describes the abundance of a species within the atmosphere.

-\item[-] \textbf{Net flux} - This describes the rate of net (absolute) change of concentration over time for a species.

-\item[-] \textbf{Absolute flux} - Some species may have a large flux going through them (production and loss), resulting in a small net flux. This sums the production and loss fluxes.

-\item[-] \textbf{ influence} - Influence is the total magnitude of an effect that changing a species concentration by 1% would have on other species within the network. Since the graph is generated using the Jacobian matrix, an alternative method for calculating this can be by calculating the total out-degree of a node.

+\item[-] \textbf{Concentration} - this describes the abundance of a species within the atmosphere.

+\item[-] \textbf{Net flux} - this describes the rate of net (absolute) change of concentration over time for a species.

+\item[-] \textbf{Absolute flux} - some species may have a large flux going through them (production and loss), resulting in a small net flux. This sums the production and loss fluxes.

+\item[-] \textbf{ influence} - influence is the total magnitude of an effect that changing a species concentration by 1% would have on other species within the network. Since the graph is generated using the Jacobian matrix, an alternative method for calculating this can be by calculating the total out-degree of a node.

\end{itemize}

## -969,7 +967,7

Finally, the 'Metric Sum' is the sum of all the metric values scaled between 1 and zero (the mean).

-\subsection{Senario Analysis}

+\subsection{Scenario Analysis}

In selecting the top 10 ranking species for each category, it is possible to examine if the importance of a species with centrality metrics varies from the results suggested by traditional metrics. In this subsection, we explore the TF-IDF rankings of each metric and use this to decide if species importance is local to a specific metric. We look at what species are highlighted by each scenario (Figures \ref{fig:heatl} - \ref{fig:heatbj}) and compare them against the primary emitted species shown in \autoref{tab:icsmetric}. Finally, we compare the total metric sum against the traditional metrics of concentration and flux and compare the correlation.

## -979,7 +977,7

\subsection*{London}

-The London dataset (\autoref{fig:fglondon}) contains a mix of anthropogenic and biogenic aromatics and long-chain alkanes. We have a section of alkanes which have a low overall metric sum and a small value for closeness and page rank. Combined with their high net flux, absolute flux and influence values, this suggests that they have a moderate directional flux, most likely influencing the production of many other species at a consistent rate. In addition to these, we have species with a moderate closeness but a high betweenness. These are often species such as formaldehyde (\ch{HCHO}), glyoxal (\ch{c2o2}) and acetaldehyde (\ch{ch3co3}) which can serve as tracers for fast photolytic reactions. This is because on the graph structure (\autoref{fig:vk}) they sit between the dense centre of the network (high closeness) and the branches formed from each primary emitted species (a low closeness value). Their high connection density and importance in the network is also picked up by the page rank algorithm. Other species with high betweenness and a low centrality are the monoterpenes limonene and $\alpha$ pinene, as well as hexane

(\ch{nc6h14}) and butane products. These are (or are close to) primary emitted species and therefore have a low closeness. Since much of the chemistry originates with such species, the outward 'flow' of information also results in a lower page rank value.

+The London dataset (\autoref{fig:fglondon}) contains a mix of anthropogenic and biogenic aromatics and long-chain alkanes. We have a section of alkanes which have a low overall metric sum and a small value for closeness and page rank. Combined with their high net flux, absolute flux and influence values, this suggests that they have a moderate directional flux, most likely influencing the production of many other species at a consistent rate. In addition to these, we have species with a moderate closeness but a high betweenness. These are often species such as formaldehyde (\ch{HCHO}), glyoxal and acetaldehyde which can serve as tracers for fast photolytic reactions. This is because on the graph structure (\autoref{fig:vk}) they sit between the dense centre of the network (high closeness) and the branches formed from each primary emitted species (a low closeness value). Their high connection density and importance in the network is also picked up by the page rank algorithm. Other species with high betweenness and a low centrality are the monoterpenes limonene and $\alpha$ pinene, as well as hexane (\ch{nc6h14}) and butane products. These are (or are close to) primary emitted species and therefore have a low closeness. Since much of the chemistry originates with such species, the outward 'flow' of information also results in a lower page rank value.

\begin{figure}[H]

\centering

-994,7 +992,7

Similar to London, the fast photochemical tracers are identified, although some have a slightly lower flux between them (Betweenness) and page rank values for Beijing (\autoref{fig:fgbeijing}). This suggests that the network structure or weightings may have shifted slightly from London, creating more links, or importance in a specific branch of chemistry.

- Additionally, their overall metric sum is lower. Glyoxal, Methyl Vinyl Ketone (MVK) and their associated criegee configurations all feature heavily in the middle of \autoref{fig:heatbj}. These are important as they represent the fast chemistry formed by both the anthropogenic and biogenic chemistry that is within the simulation. These tend to have a high closeness and page rank centrality, a pattern that is also seen with the long-chain alkane products from Octane (\ch{NC8H18}), Hexane (\ch{nc6h14}) and Isoprene.

+ Additionally, their overall metric sum is lower. Glyoxal, Methyl Vinyl Ketone (MVK) and their associated criegee configurations all feature heavily in the middle of \autoref{fig:heatbj}. These are important as they represent the fast chemistry formed by both the anthropogenic and biogenic chemistry that is within the simulation. These tend to have a high closeness and page rank centrality, a pattern that is also seen with the long-chain alkane products from Octane (n-\ch{C8H18}), Hexane (n-\ch{c6h14}) and Isoprene.

\begin{figure}[H]

\centering

-1024,16 +1022,16

\subsection{Providing An Overall Overview Using The TF-IDF And The Metric Sum.}

-In the previous section, it was shown that centrality metrics could be used to complement the use of traditional metrics in the analysis of the chemical network. As each metric represents a different aspect of importance, should a single ranking value for a node be required, it is possible to take the average sum of all three metric values. Looking at \multiref{fig:heatcv}{fig:heatbj} it is possible to see similar trends in colour gradient between the purples of the traditional metrics of flux and concentration with the total metric sum (the blue column). This suggests that it is possible to compare each scenario with the use of the metric sum.

+In the previous section, it was shown that centrality metrics could be used to complement the use of traditional metrics in the analysis of the chemical network. As each metric represents a different aspect of importance, should a single ranking value for a node be required, it is possible to take the average sum of all three metric values. Looking at \autoref{fig:heatbj} it is possible to see similar trends in colour gradient between the purples of the traditional metrics of flux and concentration with the total metric sum (the blue column). This suggests that it is possible to compare each scenario with the use of the metric sum.

In selecting the ten highest-ranking species from the mean centrality metric table for each simulation, \autoref{tab:groupcomp} can be created. Unlike the previous method, we are now looking at species which are essential across all metrics in a simulation.

-Beijing consists mainly of Quinones and Dialdehydes, which are both derivatives of Benzene. London again has Benzine related compounds, mixed with the fast photochemical indicators, which were also ranked highly in \autoref{fig:heatl}. Looking at the highest-ranking sum (NaN-mean), it is

seen that Isoprene, hept/hexane and glyoxal products highlighted as the most consistently important across all four simulations.

+Beijing consists mainly of Quinones and Dialdehydes, which are both derivatives of Benzene. London again has Benzene related compounds, mixed with the fast photochemical indicators, which were also ranked highly in \autoref{fig:heatl}. Looking at the highest-ranking sum (NaN-mean), it is seen that Isoprene, hept/hexane and glyoxal products highlighted as the most consistently important across all four simulations.

\begin{table}[H]

\centering

\input{tables/groupmetric.tex}

-\caption{\textbf{A table of the top 10 ranked species for each simulation.} Only species that exist within at least 3 out of the four simulations are used. The Nan-Mean takes the mean of all available data, ignoring runs where a species is not present.}

+\caption{\textbf{A table of the top 10 ranked species for each simulation.} Only species that exist within at least 3 out of the four simulations are used. The Nan-Mean takes the mean of all available data, ignoring runs where a species is not present. Species presented within the table follow the MCM naming convention.}

\label{tab:groupcomp}

\end{table}

## -1044,7 +1042,7

-\section{Calculating Production Sensitivity Using Personalised Page Rank.}

+\section{Causality analysis using Personalised Page Rank.}

In \autoref{sec:pagerank}, the results of the PageRank algorithm was obtained by solving for the eigenvalues and vectors of the google matrix. It was also mentioned that an equivalent method to get a result might be obtained by propagating the one's vector (a 1D vector of unity) in small increments (Equation \autoref{eqn:forwards}). This works much like the integrator within a chemical box model, except rather than updating the species concentration with each time step, we move information between each node.

## -1110,12 +1108,13

\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 $\alpha$-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 $\alpha$-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

\begin{tabular}{p{.6\textwidth}p{.2\textwidth}}

```
\toprule

+Species & PageRank Ranking\ \midrule

NC101CO & 9.920000e-01 \

APINENE & 9.210000e-06 \

NAPINBO & 4.540000e-03 \

-1183,8 +1182,8

\begin{figure}[H]

\centering

- \includegraphics[width=.7\textwidth]{figures_c3/ch2_distance.pdf}

- \caption{\textbf{Showing total the influence from each species on HCHO for a sample MCM subset of Butane.} }

+ \includegraphics[width=1\textwidth]{newfigs/ch2_distance_links.pdf}

+ \caption{\textbf{Showing total the influence from each species on HCHO for a sample MCM subset of Butane.} Species importance is determined using the reverse pagerank algorithm. It is calculated by taking a snapshot of a chemical simulation and rendered using the transpose of the Jacobian relational matrix. We then use a customised page rank algorithm to look at where flow goes from HCHO (the large node in the midde). A larger node and thicker link suggests a greater influence on the net change of HCHO. }

\label{fig:backtrace}

\end{figure}
```

## 4_lumping.tex

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

index 62d358e..5a8e92e 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 $\pi$ 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 (Chapters \multiref{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/NRI1200H-NRI1202.pdf}

\caption{\ce{ NRI1200H \ \ \ 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..457271c 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}; 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..540d41d 100644

--- a/conclusion/combigned.tex

+++ b/conclusion/combigned.tex

## -1,17 +1,98

\section{Conclusions}

-Humans are social creatures. Our increased neocortex size allows us to interact and communicate effectively between large amounts of people. It is due to this that we created methods for storing information external to our brains and developed a hive mind. Information transfer was pivotal to our development of technology and scientific advancement, however, doing so comes at a cost - and increase of radiative forcing and climate change due to anthropogenic emissions (a problem we are now trying to reduce through the use of scientific understanding and policy).

+The topic of changing climate has been a prominent talking point in the last decade \citep{IPCC2013Science}. Anthropogenic activities starting with the industrial revolution have served to increase the amount of heat retained by the earth (radiative forcing). Similarly, the use

of CFCs has damaged the protective layer of ozone in the atmosphere \citep{o3damage}, and dangerous levels of air quality have produced an increase in respiratory distress of living organisms. Although we have guidelines and policies to determine the acceptable levels of pollutants, it is not uncommon for these to be unregulated or broken - for example, it was shown that >95% of the EU urban population were exposed to concentrations higher than the WHO regulation in \citep{eea}.

-Since it is not possible to run experiments on the entire planetary system, we are forced to rely on numerical models. Within these reactions of the atmosphere are represented in the form of a mechanism. This thesis has looked at using new data processing techniques to analyse and visualise the complexity of the chemistry within the atmosphere. It is found that when addressing complex tasks within science, relying on narrative within the visual context allows for the greatest transfer of information to a reader.

-Since a mechanism can be thought of a relational representation between reactants and products, a sociograph structure of nodes and links between them proves an effective means of visually communicating patterns in a way that is intuitive.

+To prevent further irreversible harm, both physical and environmental, we must mitigate any further damage. The problem is, however, that this is not as small of a feat in itself. Taking the production of ground-level ozone as an example, the complex interplay between emissions and production within the earth system can result in two scenarios where the same concentration of a chemical species can result in the production or the loss of ozone based on the chemical regime it is in. For example, \cite{reviewo3} discusses the role of urban vegetation, and how trees can both reduce \citep{losso3} and greatly contribute to the formation \citep{isopmcm} of ozone - all while the shading provided from tree canopies could also influence the amount of radiation and its production \citep{shady}. As we try to better represent the processes that govern the physical world, the larger and more complex our models become. This means an important balance must be struck, whereupon the selected'\footnote{In atmospheric chemistry and climate change it is common to compare the results of several different models and mechanisms when studying something new. This style ofensemble' modelling provides a good way to check that things are working whilst eliminating some of the errors presented by individual simulations.} tool must be both robust, accurate within reason and computationally efficient.

-The force-directed graph structure (a subset of the sociograph class) was found to reduce many of the problems encountered within the conventional (manual) drawing of the chemistry within a mechanism. The push-pull physics nature of these makes them easy to explain while allowing for additional information (such as the rate of reaction) to be embedded within the network shape. This allows for the automatic generation and visual comparison of graphs for different sizes of mechanism, or different points within a chemical simulation.

+With the development of construction protocols, the automatic generation of very large and comprehensive chemical mechanisms is now a possibility, \citep{gecko}. This, however, presents problems which are both cognitive and computational. This thesis has explored the use of modern techniques in visualisation, reduction and machine learning in an attempt to address the above problem.

-In addition to the visualisation of a network, it is possible to undergo a level of mathematical analysis using the graph structure. Here we can convert the relationships between nodes for a point in time (as given by the Jacobian matrix) and apply several centrality metrics. These provide information about the number of links of a node, its role within the network and where the 'flow' of information is going. It was established that although node importance as provided by the centrality metrics was useful, it did not provide any additional information to current (and tested) techniques.

+\section{Results}

-In addition to node importance, it also allows us to leverage the graph network and acquire information about its structure. Global graph metrics describe the Master Chemical Mechanism as a sparse graph with the small world (highly connected clusters) and hierarchical (structured) features - properties which are common in real-world graphs. In addition to classifying the type of network, we were also able to use the infomap graph-clustering algorithm to locate the groups within this network. It was found that this can identify modular groups of species that contain many reactions between each other for the CRI mechanism (a reduced version of the MCM). When applying this to the results of several randomly initiated simulations only a couple of small size groups were found, not making it efficient for mechanism reduction through lumping, however, more work needs to be done before this is dismissed as technique.

+In attempting to optimise the information transfer between computational models and the reader, we start by understanding human evolution and how an increase in neocortex size led to the ability (and necessity) to communicate large numbers between many people. The inial method for doing this was through the use of language, storytelling and pitograms. This sort of external information 'sharing' proved pivotal for the propagation of ideas, and ultimately the creation of technology and scientific advancement. Similarly, it is seen that even in the present-day setting, the use of narrative and selected metaphors can enrich the user's ability to navigate data, and instil a personal aspect to which they can relate to. When applied to atmospheric chemical mechanisms, the resultant output is a node-link representation of reactions, where species are analogous to people or items, and reactions (relationships) the links holding them together.

-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.

+\newpage

-\section{Future Work}

-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).

+We then encode additional information into the visualisation, first syntactically using line width and colour, then semantically by setting the node positions and line lengths based on an additional property. The latter of which was achieved by a simple physical system similar to treating nodes as like charged magnets (they repulse) and links as springs pulling them back together. This force-directed graph structure (a subset of the sociograph class) alleviates many of the traditional difficulties of manually outlining a species degradation pathways. The push-pull physics nature of force-directed graphs makes them effortless to understand while allowing for additional information (such as the rate of reaction) to be embedded within the network shape all whilst being able to juxtapose different subsets or mechanisms within the same visual space (e.g. \autoref{fig:graphc1}).

+At the limits of perceivable (visual acuity) and physical resolution, we were able to translate the graphical network structure into a purely computational one. Here we are able to perform temporal analysis of the state of a mechanism within a simulation by taking a series of static 'snapshots' or aggregating the data. This mathematical approach not only gives us information on the number of reactions of a species but also its importance within the system. Similarly in looking at where the flow of information within the network we can determine bottlenecks and controlling points, whereupon a small change to a one chemical species can have a significant effect on a large number of others. This type of analysis helps us to identify important areas to study, especially in the context of policy and air quality studies.

+The computational graph can be further leveraged to categorise the type of network a mechanism represents, For example, we see that the Master Chemical Mechanism presents a sparse structure with the many highly connected (small-world) and hierarchical features. This is a pattern commonly found in real-world graphs and other chemical mechanisms, \citep{smallworld,rscgraph}.

+The classification and ranking of species their modular structure allow us to apply several graph-based clustering techniques. Rather looking at the proximity and distribution of data within space, these techniques often navigate the links of a network, locating areas of high connectivity between species - thus forming a clique, module or cluster (depending on the field of study). This form of analysis not only highlights structural patterns from the network shape (e.g. \autoref{fig:imap2page}) but can also be used to access the suitability of combined species within mechanism lumping.

+Finally, in preparation for future research, the use of different species structure representations was run through a selection of dimensionality reduction algorithms. Here different representations were reduced to two dimensions and shown in a $x-y$ scatterplot. The analysis showed t-SNE reduced data was the most aesthetically pleasing, as it provides better separation between clustered groups. Additionally, the type of representation had a significant effect on the type of features which were outlined by each DR algorithm. This highlights the importance of careful selection regarding input data when training a computational model. Out of the methods assessed it is suggested that the molecular quantum number or tokenised SMILES stings are used in any future works.

+\newpage

+\section{General Overview}

+Although no definitive improvements over existing methods have been found, the wide reaches of the study suggest that graphs, visual representations and machine learning have their place in the field of atmospheric chemistry. Although they may not replace current 'tried and tested' solutions, they have been shown to produce similar and agreeable results and demonstrated the pattern-finding abilities of computational models (for data analysis).

+Where these methods come into their own is by demonstrating a more user-friendly approach to model diagnostics, mechanism comparison and change perturbation within a large complex system. Presenting chemistry in such a way enables us to successfully communicate what is going on in a way that policymakers and the general public can understand. This in itself can go a long way into the prevention and mitigation of the global problems described at the start of this thesis.

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

+When discussing future projects relating to this work, there are two apparent avenues which should be explored. The first lies in applying this work to better communicate issues of air quality, whilst the second focuses more on the use of graph neural networks to generate dynamic mechanisms based on user requirements. These are outlined below.

+\subsection{Policy and Communication}

+As was described previously, one of the more successful parts of this project has been the communication of atmospheric chemistry in a visually intuitive way. Building on this, it would be highly bene to create an 'immersive' user-controllable box model GUI which policymakers and students can adjust, whilst watching the chemical graph dynamically change based on the user regime the chemistry falls into at that point in time. This will go a long way into educating people about the complexities of the atmosphere and how a small change may have a large effect based on circumstances/conditions.

+\subsection{Dynamic Box Model Emulation}

+Except for long-range transport, much of the chemistry which occurs within different regions of the earth is constrained by the surrounding environment. It would be useful to develop an automatically adapting mechanism based on its position within the earth system - whereupon the number of species and calculations is adjusted in accordance to location, elevation and time of day (photolysis). This will allow global and regional models to provide higher quality results - e.g. by computing high (chemical) resolution runs within urban and surrounding areas whilst removing the same computational overhead for isolated and rural grid boxes.

+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 release of graph convolutional neural networks in 2016 - this is a neural network which takes into consideration not only its inputs but also the relationships between them. If we could get a neural network to learn the protocols for mechanism construction, and then simulate a box model output based on this, the idea of an adaptive mechanism may have potential. This would nicely tie in the visual, mathematical and ML aspects of this thesis.

+\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}. 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 is 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 as 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, this thesis aimed 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/cee49a1f044b79d5413b7a02824675 08}}

+\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 a steady state based on the user's preference and aim of the study.

+Non-observational runs are initiated 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 output of the initial condition 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}
```