Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich


CHALMERS
UNIVERSITY OF TECHNOLOGY

# Short-cut models for environmental impact metrics of chemical production

Eidgenössische Techniche Hochschule Zürich

In cooperation with
Department for Energy and Environment
Chalmers University of Technology

## Master's Thesis

Master of Science
Chemical and Bioengineering

**Paul Dieterich**
15-946-171

| | |
|---|---|
| **Supervisor:** | Prof. Dr. Alexander Wokaun |
| **Co–Supervisor:** | Dr. Stavros Papadokonstantakis |
| Start: | 06/03/2017 |
| Handed in: | 09/10/2017 |

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten Semester-, Bachelor- und Master-Arbeit oder anderen Abschlussarbeit (auch der jeweils elektronischen Version).

Die Dozentinnen und Dozenten können auch für andere bei ihnen verfasste schriftliche Arbeiten eine Eigenständigkeitserklärung verlangen.

---

Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuer und Betreuerinnen der Arbeit.

**Titel der Arbeit** (in Druckschrift):

**Verfasst von** (in Druckschrift):
*Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.*

| **Name(n):** | **Vorname(n):** |
|---|---|
| | |
| | |
| | |
| | |

Ich bestätige mit meiner Unterschrift:
- Ich habe keine im Merkblatt „Zitier-Knigge“ beschriebene Form des Plagiats begangen.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert.
- Ich habe keine Daten manipuliert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Plagiate überprüft werden kann.

| **Ort, Datum** | **Unterschrift(en)** |
|---|---|
| | |
| | |
| | |
| | |

*Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.*

# Acknowledgements

My gratitude goes to Dr. Stavros Papadokonstantakis who has welcomed me at Chalmers and Gothenburg and who has always taken his time for patient explanation and help. I would like to especially thank him for his support while I was forced to take a break from this work, and that he has continuously supported me. Prof. Alexander Wokaun deserves a great thank you for happily accepting me as a student for a master's thesis abroad, even though he barely knew me. Without his trust I could not have made the experience to conduct this work at Chalmers. Furthermore, I would like to thank Dr. Sara Badr for helping me with the EI99 - ReCiPe correlation and her help in selecting the chemicals for the modelling part. Special appreciation goes to Julia Witte, who has built the connection between Dr. Stavros Papadokonstantakis and me, as well as Dr. Merten Morales, who, together with Julia, supported me like dear friends in the most challenging of times.

# Abstract

**Keywords**: ANN regression, PLS regression, RBF, LCIA modelling, ReCiPe, EI99

This work has investigated the ability of molecular structure-based (MSB) models to predict the ReCiPe indicators for environmental impact assessment. A dataset of 189 observations and 28 molecular descriptors (MDs) has been used to predict four endpoint indicators and 18 midpoint indicators. The endpoint indicators were: Ecosystem quality (EQ), human health (HH), resource depletion (R) and the total ReCiPe score (T). Linear models in form of a partial least squares (PLS) regression and nonlinear radial basis function artificial neural networks (ANNs) have been compared. It has been found that ANNs perform significantly better than linear models. The human health (HH) indicator as well as the total (T) ReCiPe indicator could be predicted with a satisfactory precision with a coefficient of determination of 0.52 and 0.44 and model size of 15 and 13 molecular descriptors (MDs) respectively. The structure of the ANN and as well as the most important MDs has been analysed. It has been found that there is a tendency to include some oxygen related functional groups, nitrogen and the molecular weight for HH and T. The results were compared with results for the EI99 indicator from literature to investigate whether it is more useful to predict the total ReCiPe indicator directly, or to correlate it with a good prediction of the total EI99 indicator. A correlation of $r^2 = 0.92$ between EI99 and ReCiPe has been found. This correlation is useful, provided there is a good prediction of the EI99 indicator. The dataset that has been used in this work predicts the ReCiPe indicator with a higher precision than the EI99 indicator, which makes is more convenient to model the ReCiPe indicator for this particular case directly. The analysis of the results has also indicated weaknesses in the modelling procedure, suggesting improvements for future applications.

# Contents

# Nomenclature

ADC Asymmetric Dependency Coefficient

ANN Artificial neural network

CED Cumulative energy demand

EI99 Eco-Indicator 99

GA Genetic Algorithm

GWP Global Warming Potential

ITSS Information Theoretic Subset Selection

LCA Life Cycle Assessment

LCI Life Cycle Inventory

LCIA Life Cycle Impact Analysis

LOOCV Leave One Out Cross Validation

MD Molecular Descriptor

MIP Mixed Integer Programming

MLR Multi Linear Regression

MSB Molecular Structure Based

OLS Ordinary least squares

PC Principal Component

PCA Principal Component Analysis

PLS Partial least squares

RBF Radias Basis Function

# 1 Introduction

The global standard of living is deeply rooted in the chemical industry, let it be the food production, technology or dyestuff. Due to its demand for energy as well as emission of greenhouse gasses and hazardous chemicals, the chemical industry poses a threat for environmental safety. Because of its omnipresence and scale, it also bears a huge potential of improvement. Legislation therefore aims at setting environmental standards for the production of chemicals, which are supposed to help achieving climate goals. Finding room for improvement in chemical production is often a big challenge due to a lack of data about energy consumption or the environmental footprint of the used materials, the so called life cycle inventory (LCI). Life cycle assessment (LCA) is a tool that provides an insight in environmental issues concerned with chemicals, based on a holistic analysis of the chemical's LCIs, from its origins, up to the factory gate, or even waste treatment. A complete LCA is based on reliable inventory data. Such data are retrieved from an analysis of energy flows, chemical reactants, catalysts etc. This analysis is often costly and time consuming. Especially for more complex chemicals, such as pharmaceuticals, many steps of the chemical process are confidential. This results in data gaps in the LCI, which pose a hurdle in the assessment of the life cycle. In order to still be able to collect environmental impact metrics, these gaps can be closed by predictive modelling. The analysis of how the LCI affects the environment, humanity and resources is called life cycle impact analysis (LCIA) . This work aims to predict the LCIA data by using "short-cut" models alternatively to complex process-based models. These "short-cut" models are based on the molecular structure of a chemical product and are therefore called molecular structure based (MSB) models.

## 1.1 Scope

In 2008 Wernet et al. published a work in which LCIA data from molecular structure based (MSB) models were successfully modelled, comparing the modelling performance of artificial neural networks (ANN) and linear models [1]. A summary of these results is given in Table 1.1. Wernet et al. used MSB models to predict the indicator data of the so-called Eco-Indicator 99 (EI99) . He has shown that artificial neural networks perform significantly better than linear models as can be seen in Table 1.1. His results have contributed to build the so called Finechem tool [2], a tool that uses a determined set of molecular descriptors (Finechem MDs) to predict the EI99 scores. In a similar approach like that of Werent et al. (2008, 2009), this works aims at predicting the so-called ReCiPe indicator, since the ReCiPe indicator is a more up to date life cycle assessment method [3]. The main goals are to find suitable linear or nonlinear models,

**Table 1.1:** Wernet et al.'s results for MSB modelling LCIA data of the Eco-Indicator 99 [1]. EI99 Total: Total Endpoint score for the EI99, EI99 HH: Human health score, EI99 EQ: Ecosystem quality score, EI99 R: Resource score

| | EI99 Total | | EI99 HH | | EI99 EQ | | EI99 R | |
|---|---|---|---|---|---|---|---|---|
| | ANN | MLR | ANN | MLR | ANN | MLR | ANN | MLR |
| $\overline{CD}$ | 0.46 | 0.25 | 0.55 | 0.13 | 0.61 | -0.01 | 0.56 | 0.35 |
| $S_{CD}$ | 0.36 | 0.47 | 0.28 | 0.46 | 0.28 | 0.88 | 0.27 | 0.41 |

that predict the ReCiPe indicators, isolating the most important molecular descriptors (MDs) for such models, and finally comparing these results with the results of Wernet et al. for the EI99. Furthermore, there is the ambition to answer the following question. From a practical point of view, is it more convenient to model the ReCiPe indicator directly or is it better to use the results of Wernet et al. and the Finechem tool as well as a correlation between EI99 and ReCiPe? In the following a short overview over life cycle assessment, molecular descriptors as well as the basics of predictive modelling is given.

## 1.2 Life Cycle Assessment

Life cycle assessment (LCA) has become a prevalent tool to categorise a vast spectrum of products in their impact on human health, ecosystem diversity and resource availability [3]. Especially studying the environmental impact of consumer products dates back to the 1960s, where two products, fulfilling the same need had to be judged in better or worse with regard to their environmental impact [4]. A common example would be selling water in glass or PET bottles. Glass bottles are reusable while PET bottles can be recycled at best. However, the glass bottle production might require more energy for there is more material used and it has a higher melting point than PET. On the other hand, PET is indirectly produced from fossil sources, the production of which is also harmful to the environment. Obviously, it is not easy to judge which of these two is the better solution for selling the consumer product "bottled water". LCA is a method to assess such problems and to help decision makers to chose a product or production path based on the LCA results.

### 1.2.1 Methodology of LCA

Life cycle assessment can be summarised in four steps [5]:

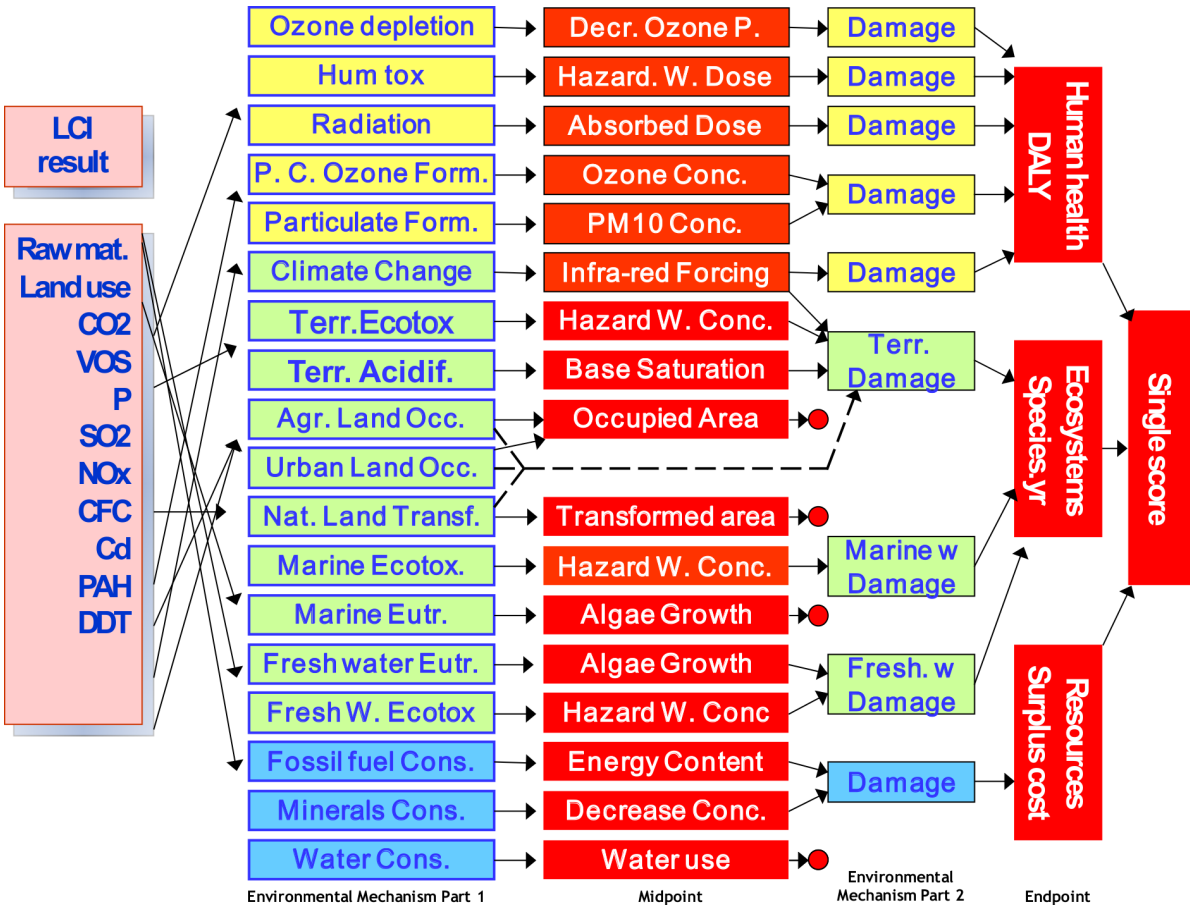
1. Define the goal and scope
2. Analyse of the life cycle inventory (LCI)
3. Life cycle impact analysis (LCIA) of midpoint and endpoint.

4. Interpret the results

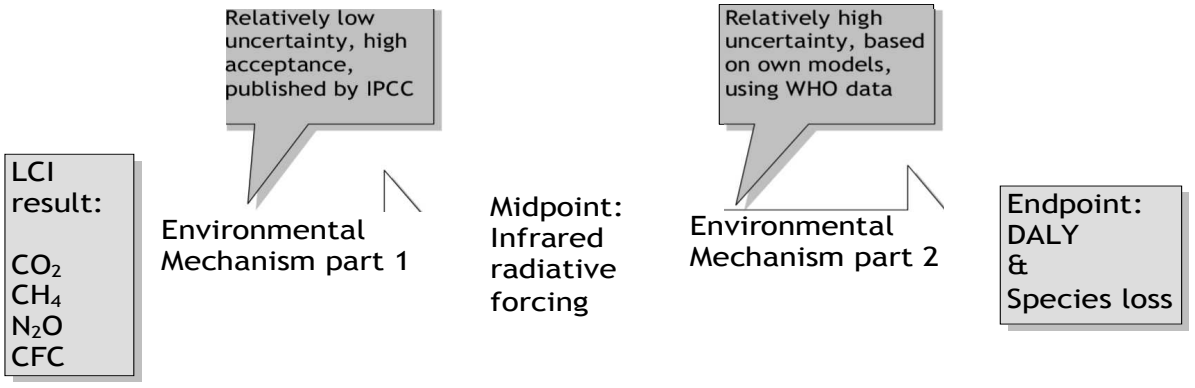
The definition of the goal and scope seems arbitrary but is cardinal to LCA. Taking the purpose of this work for instance, it is to provide a powerful tool to quickly estimate the environmental impact of chemicals when used in the production of consumer products. In the LCI (life cycle inventory) step one analyses certain impact metrics, such as the amount of carbon dioxide emitted per unit mass of the product. In the life cycle impact analysis (LCIA) these indicators are projected onto the outcome for either midpoint or endpoint indicators. For the midpoint indicators the LCI results are combined with an output, that describes a direct consequence. That is described in Figure 1.1(b), where the LCI data for carbon dioxide, methane, nitrous oxide and CFC gases are used to calculate the infrared radiative forcing, a midpoint indicator that describes the amount of solar energy absorbed by the atmosphere. Endpoint indicators use the midpoint indicators to interpret them according to their impact on human health, ecosystems, and resources, see Figure 1.1(a) and 1.1(b). Their collective values form a single score, such as the total ReCiPe score. The outcome of LCIA is influenced by the physical model behind the midpoint/endpoint calculations and therefore is different for all available methods. Wernet et al. used MSB models to predict the endpoint indicators for the EI99 and this work for the ReCiPe indicators. The results in both works will be compared in section 3.4.
The ReCiPe data were taken from the Ecoinvent 3.3 database [6]. Table 1.2 is a list of the endpoint and midpoint indicators on which a MSB modelling approach was performed.

## 1.3 Molecular Descriptors

This work focuses on "black box" models. That means that there is no scientific mechanism behind the model. It is simply fitting of data of molecular descriptors (MDs) to output data. However, depending on what output data are modeled the type of input data can have a significantly different influence. Take for instance an energy-related indicator, such as the global warming potential GWP . The global warming potential of a molecule is vastly influence by the energy used in its production process. Because often there is a thermal separation process necessary, bigger sized molecules will tend to yield a bigger GWP. Molecules containing toxic atoms will tend to have a big influence on indicators that measure toxicity, molecules that contain for example sulfur will have a big influence in indicators that measure terrestrial acidity. The mechanism behind the calculations of the indicators are not part of this work so the exact relationships between input and output cannot be investigated. Still, there are certain trends as explained above that are expected to be observed even when dealing with "black box" models. It is a paramount challenge of MSB modelling to find correlations between MDs and the LCIA indicators. Due to the black box characteristics one cannot predict quantitatively, which MDs will contribute how much to the output. Still, if there is a clear dominance of several MDs, there should be an approach to explain that correlation in order to use that information in future modelling. For example, if it is found that




(a)




(b)

**Figure 1.1:** A qualitative display of the working principle of the ReCiPe LCIA method [3]. In (a) the LCI data are displayed which are then transformed in the environmental mechanism part 1 into the midpoint indicators and from there through the environmental mechanism part 2 into the endpoint indicators. In (b) there is a more detailed example of how the LCI data of greenhouse gas emissions form the infrared radiative forcing midpoint indicator, which then influences the endpoint indicator for ecosystem quality.

**Table 1.2:** The ReCiPe indicators used in this work. All are retrieved from the ecoinvent 3.3 database [6].

| Indicator Name | Abbreviation | Unit |
|---|---|---|
| Endpoint | | |
| Ecosystem Quality | EQ | points |
| Human Health | HH | points |
| Resources | R | points |
| Total | Total | points |
| Midpoint | | |
| agricultural land occupation | ALOP | $m^2/a$ |
| climate change (also global warming potential) | GWP100 | kg $CO_2 - Eq$ |
| fossil depletion | FDP | kg $oil - Eq$ |
| freshwater ecotoxicity | FETPinf | kg $1,4 - DC.$ |
| freshwater eutrophication | FEP | kg $P - Eq$ |
| human toxicity | HTPinf | kg $1,4 - DC.$ |
| ionising radiation | IRP_ HE | kg $U235 - Eq$ |
| marine ecotoxicity | METPinf | kg $1,4 - DC.$ |
| marine eutrophication | MEP | kg $N - Eq$ |
| metal depletion | MDP | kg $Fe - Eq$ |
| natural land transformation | NLTP | $m^2$ |
| ozone depletion | ODPinf | kg $CFC - 11.$ |
| particulate matter formation | PMFP | kg $PM10 - Eq$ |
| photochemical oxidant formation | POFP | kg NMVOC |
| terrestrial acidification | TAP100 | kg $SO_2 - Eq$ |
| terrestrial ecotoxicity | TETPinf | kg $1,4 - DC.$ |
| urban land occupation | ULOP | $m^2/a$ |
| water depletion | WDP | $m^3$ |

molecular size related MDs have a high influence, a future model could be particularly rich in such MDs.

## 1.4 Modelling

Mathematical modelling aims to find a function $y = f(\beta, x)$ that uses an input variable $x$ as well as the model parameters $\beta$ and is able to calculate an output $y$. If the relationship between $x$ and $y$ is not known in details, an empirical model is used to fit the model parameters with empirical data. This process is called regression analysis. Empirical models were used in this work to predict the ReCiPe indicators using molecular descriptors of organic chemicals. In the following a short qualitative introduction into modelling is given with a more detailed discussion of the used regression techniques in section 2.4.

### 1.4.1 Introduction to Modelling

In science is often important to know the influence one or more quantities have on the physical, chemical or biological properties of a system. Either to describe a certain behaviour mathematically, to design an experiment or scale up a plant. Consequently, mathematical models can either describe a system using empirical data, or predict how variations in input values affect a system. For example, a descriptive model can describe the correlation between the height of a child and the height of its parents. A researcher could ask the question, whether tall parents will also have tall children and might come up with a simple linear model as in Equation (1.1). It describes the height of the child $H$ as a linear function of the height of the parents $x$.

$$H = \beta_1 \cdot x + \beta_0 \tag{1.1}$$

The parameter $\beta_1$ is a constant that describes how strongly the parent's height $x$ influences the child's height $H$. In order to estimate $\beta$ and $\beta_0$ empirical data can be used to perform a regression analysis. When these data are collected and $\beta_1$ and $\beta_0$ are estimated, one can judge the descriptive power by analysing how well the model fits the real data. Figure 1.2 displays how well the model in Equation (1.1) fits an imaginary data set. A commonly used metric to describe the goodness of the fit is the $r^2$. With $r^2 = 0.97$ the model fits the empirical data quite well. However, this is a pure measure of the descriptive power of the model. A more complex application of modelling is the so called predictive modelling, where available data are used to fit a model as above, which also has to be representative of new, unseen data.

The above example focuses on description of cause and effect. One could also go further and try to predict an effect with a predictive model. This is often done, when collecting the necessary data is too risky, costly or cumbersome. Consequently, predictive models are used to guide ones decisions in process development or design of experiments. In

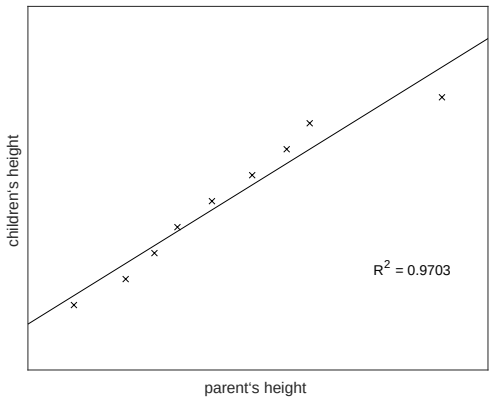



**Figure 1.2:** An example for a descriptive model, fitting the height of the children $H$ to the height of the parents $x$. One can see that the descriptive power is high with $r^2 = 0.97$. All the data are arbitrarily chosen for explanatory purposes and do not reflect real data.

1896 Svante Arrhenius tried to develop a theory to explain the development of ice ages [7]. He also came along the influence greenhouse gases have on the temperature of the earth. His research resulted in a model that correlates the amount of carbon dioxide in the earths atmosphere with the change in radiative forcing $\Delta F$. Today this model has been adjusted to Equation (1.2) [8]:

$$\Delta F(t) = \alpha \ln\left(\frac{C(t)}{C_0}\right), \tag{1.2}$$

where $C(t)$ is the $CO_2$ concentration in the atmosphere over the time $t$ and $C_0$ the concentration at time 0. The model parameter $\alpha$ can be modelled using experimental data for $\Delta F$ and $C$ using regression analysis. Arrhenius concluded that burning fossil fuels would increase the $CO_2$ content in the air and result in a global warming. Conducting such an experiment in real life would prove much too time consuming and risky as one would put the global ecosystem at risk. Going into details here is not important. It is however important, that Arrhenius was able to model that an increase of $CO_2$ in the earth's atmosphere raises the global temperature and so he would have advised against a global scale experiment. Since over 100 years later the carbon dioxide content has actually risen and the mean temperature has been monitored, we know now that Arrhenius prediction was right. His model has proven predictive enough for this purpose.

Arrhenius model is an example of a model that predicts a relationship between cause and effect and where the actual grand scale experiment (in this case increasing the amount of $CO_2$ in the atmosphere) is too risky. Predicting life cycle indicators is similar. Assuming a decision maker wants to chose between process A, B and C while the best cumulative energy demand (CED) of each process decides over which process will be implemented. Obviously, it is too costly to build all three processes and measure the CED. Instead one

could use a reliable model that predicts the CED for all processes and use it as a basis for one's decision. Arrhenius example shows one more aspect about modelling. The climate is a deeply complex system and Arrhenius was hardly able to give predictions that are as accurate as the predictions derived from modern models. However, he was essentially right about global warming. This leads to the understanding, that with a large amount of data and computational power, one can fit and predict data to an arbitrary precision. Yet, there will always remain an error at hand. This error needs to be monitored to judge the models predictive power. Basically, in predictive modelling one needs to be aware of the model's prediction error, and whether said error is small enough for the model to serve one's purpose.

"Essentially, all models are wrong, but some are useful"
*-George E.P. Box* [9]

### 1.4.2 Mathematical Basics of Modelling

A statistical model uses independent input variables $x_j$ with $j = 1, ..., p$ to predict an output variable $y$. Therefore $y$ depends on the input $x_j$ and is also called a dependent variable or target variable. The most straightforward modelling approach is to minimise the difference between a calculated or modelled output $y^*$ and the actual output $y$ taken from an empirical data set. Accordingly, the difference between $y$ and $y^*$ is called the error $\epsilon$. The calculated model output $y^*$ depends on what type of model is used and how the input variables $x_j$ are weighted, i.e. how big their influence on $y$ is. For instance, one can use a linear model with multiple input variables $x_j$:

$$y^* = \beta_0 + \sum_{j=1}^{p} (\beta_j x_j) = f(\boldsymbol{\beta}, \mathbf{x}), \tag{1.3}$$

$$\epsilon = y - y^*, \tag{1.4}$$

Where $\beta_j$ are the model parameters and $\beta_0$ the bias. For empirical models $\beta_0$ and $\beta_j$ must be estimated through a regression technique. In this work linear models have been applied in a partial least squares (PLS) regression. LCI data are based on the steps a chemical goes through in an industrial process. These are often non-linear with respect to the molecular descriptors and artificial neural network (ANN) have proven useful for this application [10],[1]. Consequently, ANNs have been used in this work to model non-linear models as they seem to be most promising. For an overview of these regressions see section 2.4.

Empirical models require a data set $\mathbf{X}$ containing data for the input variables $v_1$ to $v_p$ as well as a set representing the output variables $\mathbf{Y}$. As such, an observation (row in $\mathbf{X}$) is given the subscript $i$, while the variable (column in $\mathbf{X}$) is given $j$ as a subscript.

There are a total of $p$ variables and $n$ observations or data points.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

This work focuses on predictive modelling as described above. It is therefore necessary to assess the predictive power of the models. A common procedure is to split the data sets $\mathbf{X}$ and $\mathbf{y}$ into a training and a validation set (such as in [1] and [11]). The training sets $\mathbf{X}_{tr}$ and $\mathbf{y}_{tr}$ are used to regress the estimated model parameters $\boldsymbol{\beta}^*$, i.e. to build or "train" the model. After the training, the independent validation sets $\mathbf{X}_{val}$ and $\mathbf{y}_{val}$ are used to assess the predictive power, or "validate" the model.

1. Split the data sets $\mathbf{X}$ and $\mathbf{y}$ into a training and a validation set:
   $\mathbf{X} = \{\mathbf{X}_{tr}, \mathbf{X}_{val}\}$, $\mathbf{y} = \{\mathbf{y}_{tr}, \mathbf{y}_{val}\}$

2. Regress the training set and asses the goodness of the fit:
   $\mathbf{y}^*_{tr} = f(\boldsymbol{\beta}^*, \mathbf{X}_{tr})$, $\boldsymbol{\epsilon}_{tr} = \mathbf{y}_{tr} - \mathbf{y}^*_{tr}$

3. Use the same model as above for validation:
   $\mathbf{y}^*_{val} = f(\boldsymbol{\beta}^*, \mathbf{X}_{val})$, $\boldsymbol{\epsilon}_{val} = \mathbf{y}_{val} - \mathbf{y}^*_{val}$

It is very important for the validation set to be representative of the training set, so that every validation is an interpolation, rather than an extrapolation. Practically, there are several conditions to be fulfilled to reach a satisfactory splitting into training and validation set. Most obvious is that an observation in the validation set mustn't be an outlier with regard to the training set. Furthermore, the values of the variables in both sets need to be distributed similarly and contain about the same information content with regard to the output variable. Also, both sets need to confine a range of realistic input values that spans wide enough for a practical application. That means, chemicals of different sizes, weights and composition should be used to have a model that is representative of many chemicals that are used in industry. Section 2.3 explains how these hurdles were overcome and the splitting was performed. With an amount of data and variables that is big enough one will be able to train the model to an arbitrary precision. However, the predictive properties are assessed in the validation step. Since the parameters are fitted using the training set, it is expected to perform better than the validation set, i.e. $\epsilon_{tr} \leq \epsilon_{val}$. There can be cases, where this doesn't apply. For instance, when some points that perform really well in the training set can also be found in the validation set, or when some points in the training set perform really bad.

Up to this point it has been assumed that the model requires all the available input variables. However, in the case of black box models it is often unknown, which parameters actually contribute to the output. In other words, one often just assumes that an input variable influences the output variable. This might lead to what is known as

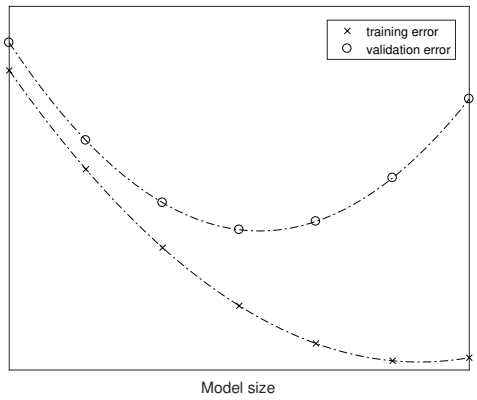



**Figure 1.3:** Qualitative display of the behaviour of the training and validation error with increasing model size. One can see that while the model becomes bigger the model is slowly over-fitted and $\epsilon_{tr}$ tends to zero. Meanwhile the over-fitted model returns a validation performance that increases with model size.

over-fitting the model. When many variables and relatively less data are available, the model parameters are adjusted so well, that the empirical output data are fitted with very high precision. This may be due to the fact, that input variables have been included in the model, that actually have nothing to do with the cause and effect relation a model aims to investigate. As a consequence the input data are so to say "learned by heart" and not used to build a predictive model. Figure 1.3 qualitatively shows how the training error approaches zero with increasing model size. Contrary, the validation error has a minimum value, at a point where the model is not at full size but reduced to its most important parameters. This minimum may well be a plateau in practice, rather than a sharp valley [12].

From Figure 1.3 one can see that is cardinal to not simply trust the model with the best training performance but to also train a number of reduced models to actually find the minimum validation error and therefore the maximum predictive performance of the model. Section 2.6 describes the methods that have been used to find the optimal molecular descriptors (MDs) given a fixed size of the model. All in all validating a model serves two purposes:

1. It makes sure, that a model can reliably predict effects and is therefore useful in the future.

2. It serves the variable selection of the multivariate model $f(\boldsymbol{\beta}, \mathbf{X})$, as over-fitting is detected through the validation performance.

# 2 Methods

In the following the methodology will be explained with the scope on the data for the molecular descriptors (MDs), an overview of the statistics and the used methods. A detailed mathematical explanation would at some points be beyond the scope of this thesis and literature for further reading is recommended. All calculations have been performed in MATLAB 2016b.

## 2.1 Data Collection

The data for the ReCiPe indicators have been taken from the ecoinvent 3.3 database. The used chemicals were restricted to all those that were confined to Europe (ReR and all its subcatogiries). A list of all the chemicals is given in the appendix. Finding a suitable set of MDs was approached in two ways: First, the finechem data, already used by Wernet et al., where calculated for the selected chemicals. For this calculation, the "finechem tool", provided by ETHZ was used. It is based on Wernet's work and calculates a set of important MDs for organic chemicals [1], [11]. Second, an independent set of MDs was searched. A complex set of MDs was found, provided by the "Milano Chemometrics and QSAR Research Group" [13], this will be refered as the "MOLE db" data. Since the performance of both sets was unsatisfactory, both sets have been combined, that is the finechem data were kept and the most promising MDs from the MOLE db data were added. A list of MDs, that were use can be found in Table 2.1. The MDs have been grouped in four categories. Size and mass related, number of atoms, functional groups, and electronics. The size and mass related MDs are expected to have an influence based on their steric behaviour. The way a molecule is polarised influences the interactions, i.e. the attractive and repulsive forces. These characteristics are influenced by some atoms, functional groups, as well as the electronic MDs. Depending on the production process behind the synthesis of groups and atomic compounds, their influence may also vary.

## 2.2 Statistical Basics and Nomenclature

Given a sample with $n$ data points $x_i$ the arithmetic mean $\bar{x}$ can be defined as:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2.1}$$

Note that $\bar{x}$ differs from the expectation value $\mu$ in the way that it is based on a finite set of empirical data. As this work is based on empirical data the statistical metrics

**Table 2.1:** Overview of the MDs used, categorised in size, atoms, groups and electronic MDs. Their abbreviations are stated in parenthesis.

| Molecular Descriptors | | | |
|---|---|---|---|
| Size/Mass | Atoms | Groups | Electronic |
| Molecular Weight (MW) | #Nitrogen Atoms (N) | #Rings (Rings) | #Double Bonds (DB) |
| Van-der-Waals Volume scaled on Carbon (VdW-V) | #Halogenes (Halogenes) | Hetero in Rings (HR) | #Number of Donor Atoms for H-bonds (DonorH) |
| #Atoms (Atoms) | #Tertiary and Quartary C-Atoms (T/Q-C) | #Functional Groups (FunctG) | #Number of Acceptor Atoms for H-bonds (AcceptorH) |
| Average Molecular Weight (AvMW) | #O in Carbonyl (OwCarb) | #Hydroxyl Groups (OH) | Sanderson Electronegativity scaled on Carbon (E-) |
| | #O without Carbonyl (Ow/oCarb) | #Carboxyl Groups (COOH) | Polarizabilites scaled on Carbon (Pol) |
| | #Oxygen Atoms (O) | #Amine and Amide (Am/Ad) | |
| | #Chlorine Atoms (Cl) | #Nitro Groups (NO) | |
| | | #Ether Groups (Ether) | |
| | | #Ester and Amide (Est/Ad) | |
| | | #Cyanide Groups (Cyanide) | |
| | | #Keto and Aldehyde (CO) | |
| | | #Other Functional Groups (OtherFun) | |

are based on the arithmetic mean rather than the expectation value. Based on $\bar{x}$ the sample variance $S^2$ can be calculated [14]:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{2.2}$$

Accordingly the sample standard deviation $S$ is defined as the square root of the sample variance.

When the data points have multiple dimension the expressions in 2.1 and 2.2 need to be adjusted. In multivariate statistics the variance of one variable with respect to other variables needs to be defined. This is called the covariance **Cov** in vector notation or $Cov_{ij}$ respectively in scalar notation. Observation vectors containing only one observation of all variables will also be in vector notation, i.e. $\mathbf{x}_i$ and variable vectors containing one variable $j$ are $\mathbf{x}_j$. The arithmetic mean $\bar{\mathbf{X}}$ of the matrix $\mathbf{X}$ is a vector containing the arithmetic mean of all columns according to (2.1):

$$\bar{\mathbf{X}} = [\bar{\mathbf{x}}_1, \dots \bar{\mathbf{x}}_j, \dots \bar{\mathbf{x}}_p] \tag{2.3}$$

Using (2.3) the covariance matrix can be defined as [14]:

$$\mathbf{Cov} = \frac{\sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{X}}\right) \left(\mathbf{x}_i - \bar{\mathbf{X}}\right)^T}{n-1} \tag{2.4}$$

with $n$ as the number of observations. The goodness of a model can be assesed by several metrics. Here the coefficient of determination $CD$ [12] is the metric of choice.

$$CD = 1 - \frac{(\mathbf{y} - \mathbf{y}^*)^T (\mathbf{y} - \mathbf{y}^*)}{(\mathbf{y} - \bar{\mathbf{y}})^T (\mathbf{y} - \bar{\mathbf{y}})} = 1 - \frac{SSE}{SST} \tag{2.5}$$

The numerator of the fraction is also called the sum square of errors (SSE) that measures the over all distance of the predicted output $\mathbf{y}^*$ and the output data $\mathbf{y}$. The denominator is the total sum of squares (SST), that calculates the distance between the output data and their arithmetic mean.

The sample correlation $r$ between a variable $x$ and $y$ is calculated as:

$$r = \frac{\mathbf{Cov}(x, y)}{S_x S_y} \tag{2.6}$$

## 2.3 Splitting

As explained in section 1.4.2 the data set is split into a training and validation set. In the following the pretreatment of the data as well as splitting criteria are explained.

### 2.3.1 Pretreatment

The data have been pretreated in order to set reasonable proportions (normalisation), to remove outliers (outlier detection) and avoid singularities (noise addition). Pretreatment of the data is of paramount importance and can make the difference between a useful model and no model at all [15].

#### Normalisation

To compensate in differences in magnitude, the data has been normalised according to equation (2.7). This is a regular transformation method that provides and easy way to compare data regardless of their physical unit or order of magnitude [16]. The data range now from zero to one.

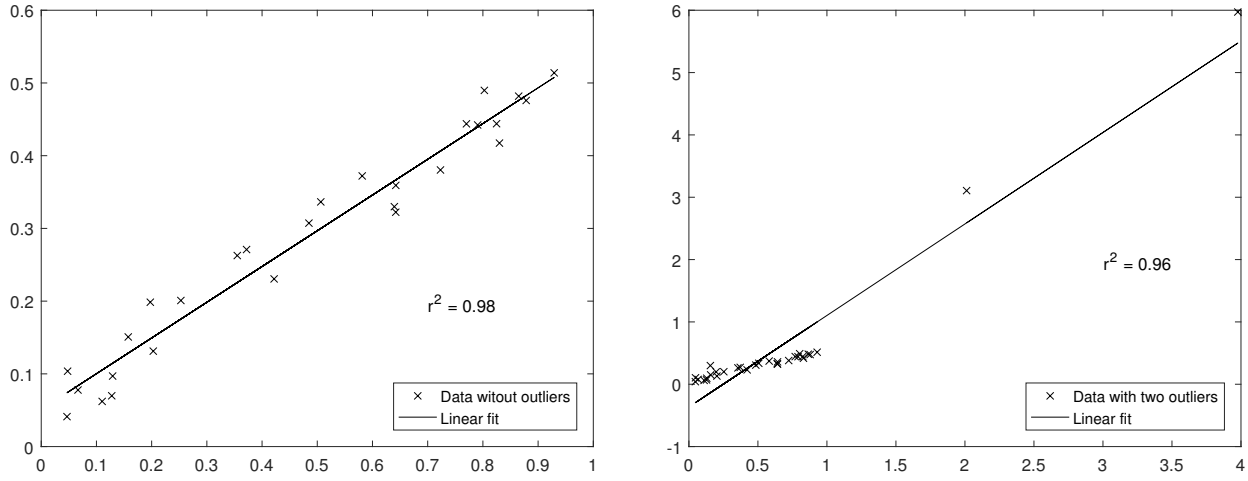
$$\mathbf{X}_{norm} = \frac{\mathbf{X} - min(\mathbf{X})}{max(\mathbf{X}) - min(\mathbf{X})} \tag{2.7}$$

Even though equation (2.7) uses the subscript $norm$, this will not be used in the further discussion, as all data can be regarded as normalised.

#### Z-test statistics

Some of the target values $y_i$ had much greater values than most of the rest. To deal with such outliers a z-test outlier detection has been performed according to [17]. Here the data in $\mathbf{y}$ are subtracted by their mean $\bar{\mathbf{y}}$ and divided by the sample standard deviation $S$. The Null-hypothesis is that the z-scores in the vector $\mathbf{z}$ have an expectation value of 0 and a standard deviation of 1.

$$\mathbf{z} = \frac{\mathbf{y} - \bar{\mathbf{y}}}{S} \tag{2.8}$$




(a) (b)

**Figure 2.1:** An example for the importance of the outlier detection. In (a) the fit is good according to the $r^2$ and the trend of the data is well represented. In (b) the outliers weight much more than the data that lay closely together. While the results for the $r^2$ seem to stand for a good fit, one can see that the actual data trend is badly represented.

For this application a z-score of 2 has been chosen as the upper limit. According to [18] the probability of the interval of $z = [0, 2]$ incorporating a measured value is 97.7%. Accordingly, all observations with $z_i \geq 2$ have been discarded. Note that for every ReCiPe indicator the sets have been split isolated from the other ReCiPe indicators. That means that if for example the EQ indicator has an outlier at observation $i$, for another indicator observation $i$ might not be an outlier. Figure 2.1 shows how important outlier detection is. One can see two data sets, one containing no outliers, and the other one containing outliers. On the $r^2$ one can see that they're fitted well by a linear model. However in Figure 2.1(b) the outliers steepen the fit and make it obviously unrepresentative of the actual data. While in a two-dimensional space this can be displayed well, in the three-dimensional space and higher, a fit may seem good according to some metric, but is actually strongly influenced by a few outliers which hides the true trend of the majority of the data.

### Noise

The data matrix in this application contains many zero elements. This is due to several integer variables, such as the number of oxygen atoms, that are zero for many chemicals. In the splitting as well as in the modelling part, it was necessary to calculate the inverse of the data matrix $\mathbf{X}$ or the covariance matrix $\mathbf{Cov}$. Noise has been added to avoid inaccurate results due to singular matrices. The added noise is randomly distributed

between $10^{-6}$ and $10^{-5}$.

### 2.3.2 Creating N Random Splits

In order to be able to validate the model it is important to split the data into a training and a validation set such that the validation is "representative" of the training set. The first approach is to chose a validation set, so that its output variables lay strictly within the range of the output variables of the training set. This way one makes sure that the validation is always an interpolation within a range of output variables rather than an extrapolation that exceeds the output variables of the training set. Furthermore, a calibration interval is an obligation for any predictive model. So for every data point $y_{i,val}$ of the output variable of the validation set the following inequality must be valid:

$$\min(\mathbf{y}_{tr}) \leq y_{i,val} \leq \max(\mathbf{y}_{tr}) \quad \forall\ i \in n_{val}$$

Additionally the training set is usually chosen to be larger than the validation set. In this case about for times. (Obviously, for a number of observation $n$, that is not divisible by 4, this factor is a little higher).

$$n_{tr} = 4n_{val},$$

With $n_{val}$ and $n_{tr}$ being the observations in the validation and training set respectively. Not only does the output variable of the validation set have to be in the same (or smaller) range than the output variable of the training set, also the input matrices for training and validation set have to be in a "close" range. For a multivariate dataset this is hard to imagine because the distances are not measured along a one dimensional axis any more.

Two multivariate distance measures are available: the euclidean and the mahalanobis distance. Both calculate the distance of a data point $\mathbf{x}_i = [x_{i,1}, \cdots x_{i,j}, \cdots x_{i,p}]$ to the mean value $\bar{\mathbf{X}}$ of the whole data set. The advantage of the mahalanobis distance is shown in Figure 2.2. The mahalanobis distance takes into account how much a variable varies with the other variables, i.e. the covariance matrix **Cov**. Figure 2.2 shows a generic data set with two variables x1 and x2 with fixed distance thresholds, indicated by circles. Assuming that the middle circle is the maximum criterion for an outlier detection, in Figure 2.2 (a) many data points would be considered outliers due to their strong variation in x1 and due to the fact, that the data points are not circular but elliptically distributed. It is clear that the distance criterion needs to be adjusted. The correlation of the data needs to be regarded to provide a credible distance measure. This is achieved in the mahalanobis distance. It adjusts to the distribution of the data in the $p$ dimensional space as well as to the magnitude of the variables [19]. In Equation (2.9) and (2.10) the general expressions for the euclidean distance $DE_i$ and mahalanobis distance $DM_i$ of every data point $i$ are given.

$$DE_i = \sqrt{\left(\mathbf{x}_i - \bar{\mathbf{X}}\right)\left(\mathbf{x}_i - \bar{\mathbf{X}}\right)^T} \tag{2.9}$$

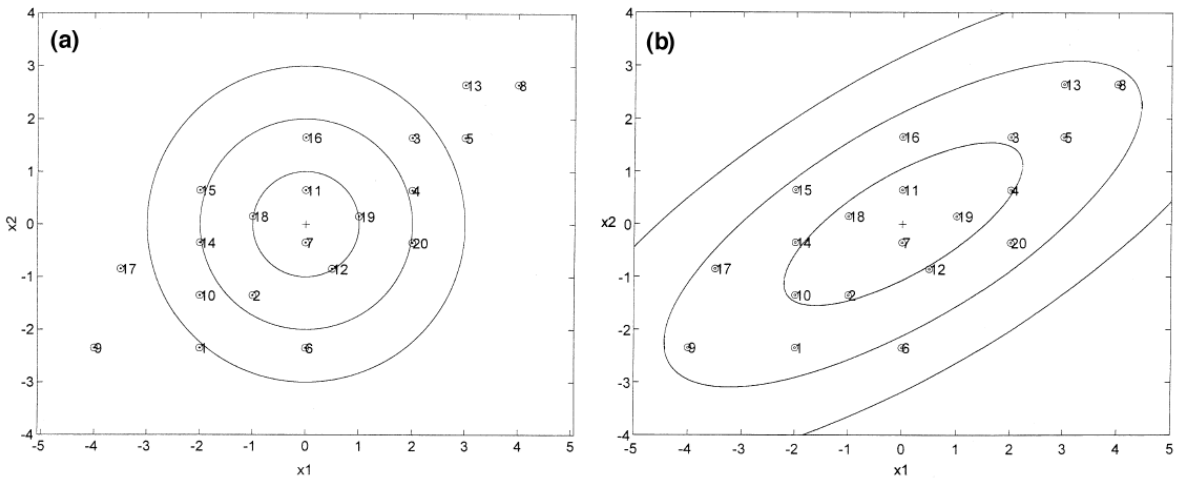



**Figure 2.2:** (a) A simulated set of data with three circles representing equal euclidean distance from the sample mean. (b) A simulated set of data with three ellipses representing equal mahalanobis distance from the sample mean[19]. One can see that the mahalanobis distance takes the covariance of the variables into account and adjusts to the shape of the data.

$$DM_i = \sqrt{\left(\mathbf{x}_i - \bar{\mathbf{X}}\right) \mathbf{Cov}^{-1} \left(\mathbf{x}_i - \bar{\mathbf{X}}\right)^T} \tag{2.10}$$

To set the validation set in an appropriate mahalanobis distance to the training set, the mahalanobis distance of every observation in the validation set $\mathbf{x}_{i,val}$ to the mean of the training set $\bar{\mathbf{X}}_{tr}$ has been taken.

$$DM_{i,val-tr} = \sqrt{\left(\mathbf{x}_{i,val} - \bar{\mathbf{X}}_{tr}\right) \mathbf{Cov}_{tr}^{-1} \left(\mathbf{x}_{i,val} - \bar{\mathbf{X}}_{tr}\right)^T} \tag{2.11}$$

In order to find validation sets that were representative of the training set in the sense of the above criteria, the following procedure has been applied:

1. Randomly create $N$ splits, such that $n_{tr} = 4n_{val}$ and $\min(\mathbf{y}_{tr}) \leq y_{i,val} \leq \max(\mathbf{y}_{tr}) \ \forall \ i \in n_{val}$
2. Calculate the mahalanobis distance $DM_{i,val-tr} \ \forall \ i \in n_{val}$. If $DM_{i,val-tr}$ is an outlier according to the z-score, replace $\mathbf{x}_{i,val}$ with a random data point of the training set. Repeat the calculation of the mahalanobis distance until $DM_{i,val-tr}$ is no outlier $\forall \ i \in n_{val}$

This way it has been made sure that $N$ random split have been created, where the validation set is representative of the training set according to the above conditions.

Besides the above mentioned conditions for a split, there are two more criteria that have been applied: the entropy and the ITSS criterion. If these criteria have been fulfilled, the splits have been kept, if not they were discarded. The goal was to have about 100 splits after applying the entropy and ITSS criterion. Accordingly, a bigger number of $N$ random splits, that fulfil the requirements above, had to be created. Since the entropy and ITSS criteria were quite strict $N$ had to be in the order of $N \approx 2000 - 4000$ to actually reach the desired 100 splits. Before those two criterions could be applied, the input as well as the output space had to be discretised. Discretisation means that the values of a variable are distributed among bins of equal size. Since the variables have already been normalised and take on values between zero and one, these bins also will be between zero and one. A discretisation has been performed, such that the number of bins, or the resolution, resulted in maximum entropy for the training set. The entropy is defined below. The resolution is the size one bin spans. For example a variable between zero and one that is distributed among five bins has a resolution of 0.2. Again it is important that the validation set is a good proxy for the training set. Ergo, the resolution of the discretisation in the training set must be equal to the one in the validation set. Otherwise the discretisation will be different for training and validation set.

### 2.3.3 Entropy Criterion

In his paper "A mathematical theory of communication", Shannon defined the statistical entropy of a variable $\mathbf{x}$ according to Equation (2.12) [20].

$$H(\mathbf{x}) \equiv E[I(\mathbf{x})] \tag{2.12}$$

Where $I[\mathbf{x}]$ is the information content in the variable $\mathbf{x}$. Explicitly this can be reformulated using the probability $p(x_i)$ that $\mathbf{x}$ takes on $n$ discrete values $x_1$ to $x_n$ and it is therefore an analogy to the thermodynamic entropy introduced by Boltzmann and Gibbs in 1870 [21].

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i)) \tag{2.13}$$

The probability $p(x_i)$ is simply the number of occurrences $N_i$ of the value $x_i$ divided by the total number of occurrences $N$. Note that this could also be applied to continuous variable, if they were rendered by a continuous function. Since empirical data are discontinuous they should be distributed among $N$ discrete bins with the occurrences $N_i$.

$$p(x_i) = \frac{N_i}{N} \tag{2.14}$$

According to the definition in Equation (2.12) the entropy represents the information content that lays within the variable $\mathbf{x}$. As an example, assume that a variable can only take on one value $x_1$, that would mean that $p(x_1) = 1$. With regard of Equation (2.13), this would result in an entropy of $H(\mathbf{x}) = 0$, since the logarithm of one is zero.

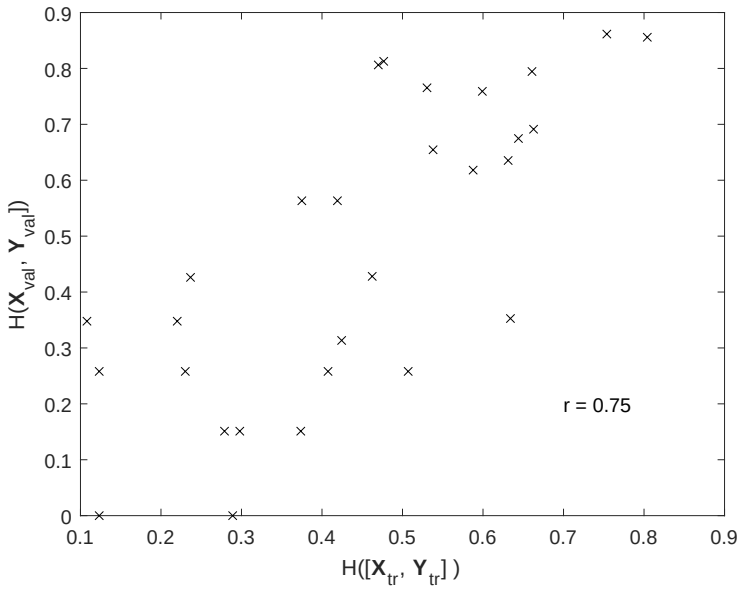



**Figure 2.3:** An example for the entropy correlation criterion for the training and validation set. On the abscissa the entropy of the training set and on the ordinate for the validation set for each column of the input matrix. The correlation coefficient is $r = 0.75$ and the split is accepted.

Now assume a variable that can take on $n$ discrete values with equal probability, i.e. $p(x_1) = 1/n, \ldots, p(x_n) = 1/n$ (for example a perfect coin or dice). The entropy of this variable would be the maximum entropy, $H(\mathbf{x}) = 1$ since it it impossible to predict which value the variable will take on.

For comparing the training and validation sets it is now cardinal that all variables have a more or less equally distributed entropy. That means that the entropy of the variables in the training and validation set should correlate. As a threshold a correlation $r$ according to Equation (3.2) of 0.7 has been set, meaning that all splits with $r < 0.7$ have been discarded. Figure 2.3 gives an example of the entropy of each variable for the training set versus the validation set. The correlation coefficient $r$ is 0.75 which is above the entropy criterion and the split will be kept. The entropy criterion is applied for a matrix consisting of all the input variables $\mathbf{X}_{tr}$ and $\mathbf{X}_{val}$ as well as the output variables $\mathbf{y}_{tr}$ and $\mathbf{y}_{val}$.

### 2.3.4 ITSS Criterion

In 1998 Sridhar et al. proposed an "information theoretic subset selection" (ITSS) to identify important variables before training an ANN [22]. By identifying the most important variables the input space could be reduced and the training accelerated while containing a satisfying performance. The ITSS criterion is also based on information

theory and the entropy according to Equation (2.13). Analogously, the combined entropy of two discrete variables $\mathbf{x}$ and $\mathbf{y}$ can be defined:

$$H(\mathbf{x},\mathbf{y}) = -\sum_i \sum_j p(x_i,y_j)\log_2(p(x_i,y_j)), \tag{2.15}$$

with $p(x_i,y_i)$ being the probability that $\mathbf{x}$ takes on $x_i$ while $\mathbf{y}$ takes on $y_i$. Note that $\mathbf{x}$ is a vector containing observations of only one variable. The input matrix $\mathbf{X}$ for the modelling has $p$ variables. The ITSS criterion is subsequently applied to all columns $\mathbf{x}_j$ of $\mathbf{X}$. For explanatory reasons the index $j$ is left away for now. Using the entropies, an asymmetric dependency coefficient (ADC) $U(\mathbf{y}|\mathbf{x})$ can be defined according to [22].

$$U(\mathbf{y}|\mathbf{x}) \equiv \frac{H(\mathbf{y}) + H(\mathbf{x}) - H(\mathbf{y},\mathbf{x})}{H(\mathbf{y})} \tag{2.16}$$

The ADC measures how much information about $\mathbf{y}$ is stored in $\mathbf{x}$. The maximum, $U(\mathbf{y}|\mathbf{x}) = 1$ means that $\mathbf{x}$ fully describes $\mathbf{y}$, while $U(\mathbf{y}|\mathbf{x}) = 0$ stands for no information at all. The ADC can be used as a selection criterion for the training/validation set splits. In this application it is not used for a single input variable $\mathbf{x}$ any more, but for each column $\mathbf{x}_j$ of the input matrix $\mathbf{X}$. For each training set the ADC has been calculated. For every split those variables, that are responsible for an ADC of over 5% were selected. For example in the following case, variables one, three and four would be selected for the split $s_1$ and all variables would be selected for the split $s_2$.

$$\mathbf{U}^{s_1}(\mathbf{y}|\mathbf{X}) = \begin{pmatrix} U(\mathbf{y}|\mathbf{X}_1) \\ U(\mathbf{y}|\mathbf{X}_2) \\ U(\mathbf{y}|\mathbf{X}_3) \\ U(\mathbf{y}|\mathbf{X}_4) \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.01 \\ 0.29 \\ 0.3 \end{pmatrix}, \ \mathbf{U}^{s_2}(\mathbf{y}|\mathbf{X}) = \begin{pmatrix} U(\mathbf{y}|\mathbf{X}_1) \\ U(\mathbf{y}|\mathbf{X}_2) \\ U(\mathbf{y}|\mathbf{X}_3) \\ U(\mathbf{y}|\mathbf{X}_4) \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.06 \\ 0.13 \\ 0.31 \end{pmatrix}$$

Each variable that is selected in one split occurs one more time. That means, that the frequency of occurrence $F(v_j)$ of each variable $v_j$ is monitored. The above example would therefore have frequencies of $F(v_2) = 1$ and $F(v_1, v_3, v_4) = 2$. Furthermore, the mean contribution $C(v_j)$ of each variable was calculated as well, using only values of ADC of over 5%. The importance of each variable is now described as a sort of weight $W(\mathbf{v})$, which is the product of frequency of occurrence and the mean contribution. Take the above example again:

$$\mathbf{F}(\mathbf{v}) = \begin{pmatrix} 2 \\ 1 \\ 2 \\ 2 \end{pmatrix}, \quad \mathbf{C}(\mathbf{v}) = \begin{pmatrix} 0.45 \\ 0.06 \\ 0.21 \\ 0.305 \end{pmatrix} \rightarrow \mathbf{W}(\mathrm{v}) = \begin{pmatrix} 0.9 \\ 0.06 \\ 0.42 \\ 0.61 \end{pmatrix}$$

Those variables, who's weight lays above the average were called important (in the above example that would be variable one and four, since the average of the weight is 0.49). For a split to be selected, i.e. for a validation set to be representative of the training set according to the ITSS criterion, the important variables in the validation set have

to contribute to 60% of the information content the important variables contribute to the training set. That way one makes sure, that about the same input variables are most descriptive of the output variable in both sets. The procedure is summarised in the following:

1. Calculate $U^s_{j,tr} = U(\mathbf{y}_{tr}|\mathbf{x}_{j,tr})\ \forall j$ and $U^s_{j,val} = U(\mathbf{y}_{val}|\mathbf{x}_{j,val})\ \forall j$ for all splits $s$.
2. Select only those splits with $\text{sum}(\mathbf{U}^s_{tr}) \wedge \text{sum}(\mathbf{U}^s_{val}) \geq 0.7$ to ensure that only sets are used where the input variables are descriptive of the output variable.
3. Count for every set all variables $v_j$ with $U^s(\mathbf{y}|\mathbf{x}_j) \geq 0.05$ to exclude variables with low information content.
4. For all training sets count the frequency of occurrence $F_{j,tr} = F_{tr}(v_j)$ of each variable $v_j$, after having discarded the ones with low information content in 3.
5. Take the average value of all variables that contribute to more than 5%: $\mathbf{C}_{tr} = mean_s\left(\left[\mathbf{U}^1_{tr}, \dots \mathbf{U}^s_{tr}, \dots \mathbf{U}^N_{tr}\right] \textit{ without all } U^s_{j,tr} < 0.05\right)$
6. Calculate the weights: $W_{j,tr} = F_{j,tr} \cdot C_{j,tr}$.
7. Define the important variables to be: $V_{important} = v_j$ if $W_{j,tr} \geq \left(\bar{\mathbf{W}}_{tr}\right)$
8. Set the contribution $c$ of the important variables to the information content in the training sets: $c = \text{sum}\left(\mathbf{C}_{tr}(v^*_j)\right) \quad \forall v^*_j \in V_{important}$
9. Keep all splits $s$ that fulfill the ITSS criterion: $\text{sum}_{j^*}\left(\mathbf{U}^s_{val}\right) \geq 0.6 \cdot c \quad \forall v^*_j \in V_{important}$

## 2.4 Linear Regressions Methods

The principle of multiple linear regression (MLR) is to calculate parameters $\boldsymbol{\beta}$ that fit the input data in $\mathbf{X}$ to the target data $\mathbf{y}$ according to a linear model [12]:

$$\mathbf{y} = \beta_0 + \beta_1 \cdot \mathbf{X}_1 + ... + \beta_p \cdot \mathbf{X}_p \ = \mathbf{X}\boldsymbol{\beta} \tag{2.17}$$

With $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^T$. Here $\beta_0$ is the so called bias and $\beta_1$ to $\beta_p$ the model parameters for variable 1 to variable $p$. The difference between the calculated target variable $\mathbf{y}^*$ and the actual target data $\mathbf{y}$ is called the error $\boldsymbol{\epsilon}$.

$$\mathbf{y}^* = \mathbf{y} + \boldsymbol{\epsilon} \tag{2.18}$$

With equation (2.17) follows equation (2.19).

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.19}$$

The minimisation of the error has been performed by OLS and PLS regression.

### 2.4.1 Ordinary Least Squares Regression

The OLS regression defines an objective function $\Phi$ according to equation (2.20), that is the squared distance of the true output $\mathbf{y}$ to the estimated output $\mathbf{y}^*$.

$$\Phi = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \tag{2.20}$$

The objective function $\Phi$ needs to be minimised now with respect to the parameters $\boldsymbol{\beta}$. The optimised parameters are then called $\boldsymbol{\beta}^*$:

$$\boldsymbol{\beta}^* = \arg\,\min_{\boldsymbol{\beta}} (\Phi) \tag{2.21}$$

One can show that Equation (2.21) can be reformulated to the analytic expression [12]:

$$\boldsymbol{\beta}^* = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2.22}$$

Of course, this is provided, that $\left(\mathbf{X}^T\mathbf{X}\right)$ is invertible, i.e. that $\mathbf{X}$ has full column rank. Since zero columns or linearly dependent columns are avoided in the calculations, this always applies. The OLS regression has been used in the regression of the weights of the ANN. For time reasons there were no data collected that show its performance in the modelling of ReCiPe indicators. All the linear modelling was performed in a PLS regression.

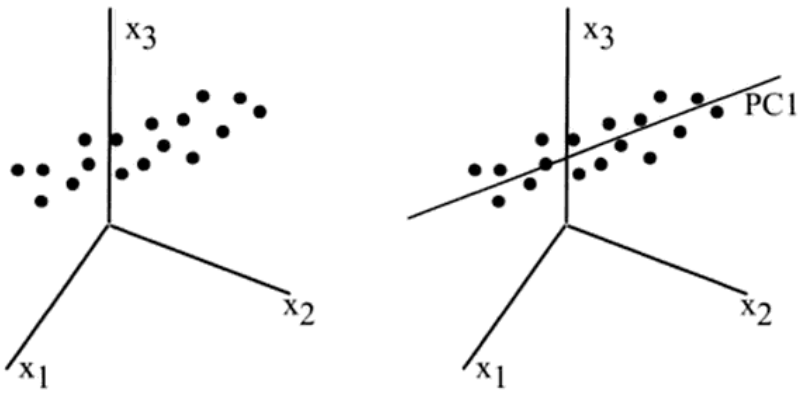
### 2.4.2 Partial Lest Squares Regression

Before introducing the partial least squares (PLS) regression it is important to quickly go over the principal component analysis (PCA). Since this methodology part cannot cover any details further reading is advised (such as: Kim. H. Esbensen et al. "Multivariate Data Analysis - in practice" [15] and Rosipal, R., and N. Kramer. "Overview and Recent Advances in Partial Least Squares."[23] ).

#### Principal Component Analysis

The purpose of multivariate data analysis is often to find the influence of one variable, or better its measurement data in column $m$ of the data matrix $\mathbf{X}$. This influence is related to the covariance $Cov_{m,j}$ of $m$ with respect to the other column $j$. Take a three dimensional input matrix with the data displayed in Figure 2.4. On the right hand side in Figure 2.4(a) one can see the data in the three dimensional input space. Figure 2.4(b) shows how a line is drawn in the direction of the biggest covariance.

The newly created axis PC1 is called the first principal component (PC) . This way one can reduce the input space to a number of PCs that display the input space with a satisfactory precision, measured by the covariance covered by the PCs. A big advantage of this procedure is, that noise, i.e. PCs, that have only very small covariance, can




(a) (b)

**Figure 2.4:** Schematic representation of the PCA in a three-dimensional space. The first PC covers most of the phenomena observed in $\{x_1, x_2, x_3\}$ as can be seen in (b)[15].

be excluded. In PCA regression and PLS regression these PCs are treated as input variables in a (linear) regression. All in all, the PCA decomposes the $p$ dimensional input space of orthogonal variables into a input space of maximal $rank(\mathbf{X})$ and minimal one orthogonal PCs. Since the PCs are spanned in the order of descending covariance, the covariance decreases with every PC, until all phenomena in $\mathbf{X}$ are covered, or only stochastic noise is left. Mathematically the PCA describes $\mathbf{X}$ as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \quad \text{or} \tag{2.23}$$

$$\mathbf{X} = \begin{pmatrix} t_{1,1} \\ \vdots \\ t_{i,1} \\ \vdots \\ t_{n,1} \end{pmatrix} \cdot \begin{pmatrix} p_{1,1} \cdots p_{1,j} \cdots p_{1,p} \end{pmatrix} + \ldots + \begin{pmatrix} t_{1,k} \\ \vdots \\ t_{i,k} \\ \vdots \\ t_{n,k} \end{pmatrix} \cdot \begin{pmatrix} p_{k,1} \cdots p_{k,j} \cdots p_{k,p} \end{pmatrix} + \ldots + \mathbf{E},$$

Where $\mathbf{T}\mathbf{P}^T$ is called the structure and $\mathbf{E}$ the noise. The structure consists of the score matrix $\mathbf{T}$ as well as the loadings matrix $\mathbf{P}$. PCA now aims to find a score vector $\mathbf{t}_i = (t_{i,1}, \ldots, t_{i,m})$ for every observation $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$ with the weight $\mathbf{p}_k = (p_{k,1}, \ldots, p_{k,p})$

$$t_{i,k} = \mathbf{x}_i \cdot \mathbf{p}_k \quad \text{for} \quad i = 1, \ldots, n \ , \quad k = 1, \ldots, m, \tag{2.24}$$

such that $t_1$ inherits the maximum possible variance from $\mathbf{X}$. Since in Equation (2.24) the loadings $\mathbf{p}_k$ are dependent on $t_{i,k}$, PCA is an iterative procedure, for which several numeric algorithms are available.

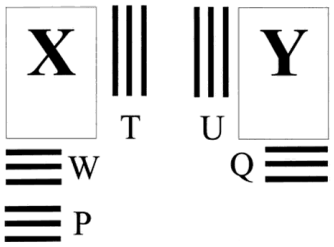



**Figure 2.5:** Conceptual decomposition of $\mathbf{X}$ and $\mathbf{y}$ for PLS regression. $\mathbf{X}$ is decomposed in the loading matrix $\mathbf{T}$ and the weights $\mathbf{P}$. The decomposition of $\mathbf{y}$ into $\mathbf{U}$ and $\mathbf{Q}$ affects the loading weights $\mathbf{W}$, which are used to calculate $\mathbf{T}$ [15].

### Partial Least Squares Regression

The above explained PCA aims to find phenomena within the input matrix $\mathbf{X}$. In multivariate regression however, the influence of each variable $\mathbf{x}_j$ on the output data $\mathbf{y}$ is of particular interest. Consider the input matrix $\mathbf{X}$ and the output vector $\mathbf{y}$, which can be decomposed according to PCA in their scores and loadings matrices $\mathbf{T}$, $\mathbf{P}$, $\mathbf{U}$ and $\mathbf{Q}$ [23]. (Note that this explanation focuses on one dimensional target variables, but the same applies for multi-dimensional target variables).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2.25}$$

$$\mathbf{y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{2.26}$$

With the residuals $\mathbf{E}$ and $\mathbf{F}$. $\mathbf{X}$ and $\mathbf{y}$ are not decomposed independently but rather in a way that the decompositions affect themselves. So is the starting point $\mathbf{t}_1$ for the PCA of $\mathbf{X}$ replaced by $\mathbf{u}_1$. The weights in $\mathbf{P}$ are then calculated, using the "new" starting values of $\mathbf{u}_1$ and then called $\mathbf{W}$. Based on the new weights $\mathbf{W}$ the scores $\mathbf{t}_1$ are calculated independently and then replace $\mathbf{u}_1$. This makes clear that both decompositions affect each other in the PLS regression. Figure 2.5 shows the structure of the PLS decomposition. $\mathbf{P}$ are the loadings of $\mathbf{X}$ and $\mathbf{W}$ the loading weights.

As in PCA the loadings $\mathbf{P}$ represent the relationship between the scores $\mathbf{T}$ and $\mathbf{X}$. However, the loading weights $\mathbf{W}$ represent the relationship between the input data $\mathbf{X}$ and the output data $\mathbf{y}$. The set of loadings $\mathbf{Q}$ describe the regression between $\mathbf{y}$ and its scores $\mathbf{U}$. The loadings and loading weights are the key parameters when it comes to regressing a linear model, such as in Equation (2.17). Equivalent to the OLS regression an expression for $\boldsymbol{\beta}^*$ can be derived [15].

$$\boldsymbol{\beta}^* = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}^T \tag{2.27}$$

Going into detailed explanation of the PLS regression and the underlying mathematics would be too much at this point. For further reading see: Kim. H. Esbensen et al.

"Multivariate Data Analysis - in practice" [15] and Rosipal, R., and N. Kramer. "Overview and Recent Advances in Partial Least Squares."[23].

The PLS regression has been performed using the MATLAB function "plsregress"[24], which uses the SIMPLS algorithm [25]. Note that, when a PLS regression is performed, the model size does not reflect on the number of MDs included in the model, but on the number of PCs that are calculated in the iterative procedure. Depending on the covariance of the different MDs with respect to $\mathbf{y}$, the PCs are composed more or less of the MDs.

## 2.5 Artificial Neural Networks

Artificial neural networks (ANN) are a fast and efficient way to model non-linear relationships [26]. They are expected to perform outstandingly better than the linear methods when non-linear processes are modeled, since they inherently are non-linear networks [27]. ANNs are rooted in the natural neural networks, such as in animal brains. An input is activated and processed through different nodes, forming different layers. Every node is constituted of a transfer function and has interconnections between the nodes of different layers. The output is then a function of a complex neural structure that has formed during the training phase. Just as a trained animal brain needs to deal with new inputs a well trained ANN is expected to perform well in the validation process. Figure 2.6 shows the basic structure of an ANN.

The boxes are called input layer and store the input to the network. This work uses the fuzzy input partitioning which is explained below. The circles are the neurons or nodes, that process the information in the hidden layer with a transfer function for which a radial basis function (RBF) is used. The hidden layer is connected and weighted to the output layer. The weights $\mathbf{w}$ are calculated by a OLS regression. The creation of an ANN consists of two steps [27]. First the network size has to be determined, i.e. the number of hidden layers and the neurons. Second all parameters, associated with the neurons and the links are regressed, so that the training error is minimised. In this work the number of hidden nodes is determined by a fuzzy partition of the input space, also see [27] and [28] and there is only one hidden layer.

### Building the Grid

Before an ANN is trained one must span the grid of inactive nodes for the $p$ dimensional input space. These nodes form the grid, compare Figure 2.6, that will be activated in the training procedure. For each variable the nodes are equidistantly distributed with an optimal resolution. That means that for different variables, i.e. different columns of the grid matrix $\mathbf{C}_{grid}$, there is a different number of nodes whose values are normalised and span between zero and one. Take for instance three variables, that have an optimal

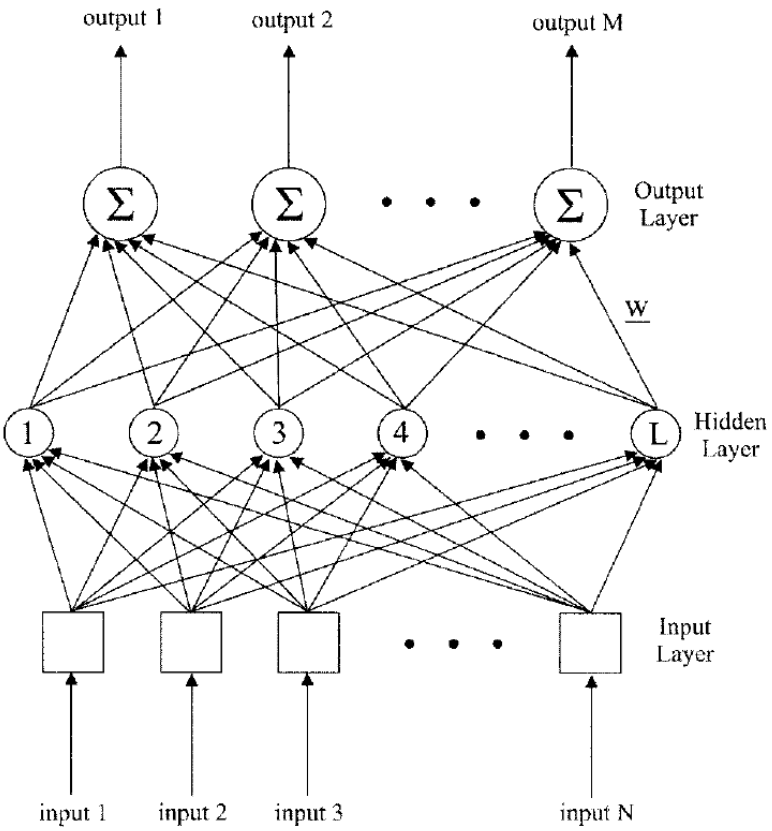



**Figure 2.6:** Basic structure of an ANN [27]. The inputs are activated in the input layer (here a fuzzy partitioning algorithm is used) and then passed to a hidden layer where the nodes transfer the input. The output layer is a linear combination of the hidden layer, with the weights **w** being calculated by OLS regression.

discrete resolution of 2,4 and 5. Their grid matrix $\mathbf{C}_{grid}$ will look as follows:

$$\mathbf{C}_{grid} = \begin{pmatrix} 0 & 0 & 0 \\ 0.5 & 0.25 & 0.2 \\ 1 & 0.5 & 0.4 \\ 0 & 0.75 & 0.6 \\ 0 & 1 & 0.8 \\ 0 & 0 & 1 \end{pmatrix}$$

Which of these nodes are activated will be decided by the actual input values of the training set. These will activate nodes according to the fuzzy partitioning of the input space.

### Fuzzy Partitioning of the Input Space

Fuzzy partitioning is used for systems with great uncertainty. Other than in a discretisation procedure, where values of a variable are distributed among discrete bins, a fuzzy partitioning describes a value with a membership function. That means that a value can "more or less" belong to a discrete point in space, depending on the value of said membership function. Take $p$ input variables $x_j$ with values $j = 1, 2, \dots p$ and partition it into $c_j$ triangular fuzzy sets $A_j^1, A_j^2, \dots, A_j^{c_j}$. The membership function $\mu_A(x)$ of the fuzzy set is described as [28]:

$$\mu_A(x) = \begin{cases} 1 - \dfrac{|x - a|}{\delta a}, & \text{if } x \in [a - \delta a,\ a + \delta a] \\ 0, & \text{otherwise} \end{cases}, \tag{2.28}$$

where $a$ is the center element to which the value $\mu_A(x) = 1$ is assigned and $\delta a$ the half of the respective width (compare Figure 2.7). The membership function is calculated for every observation in the training set and the node with the highest membership function will be activated by said observation.

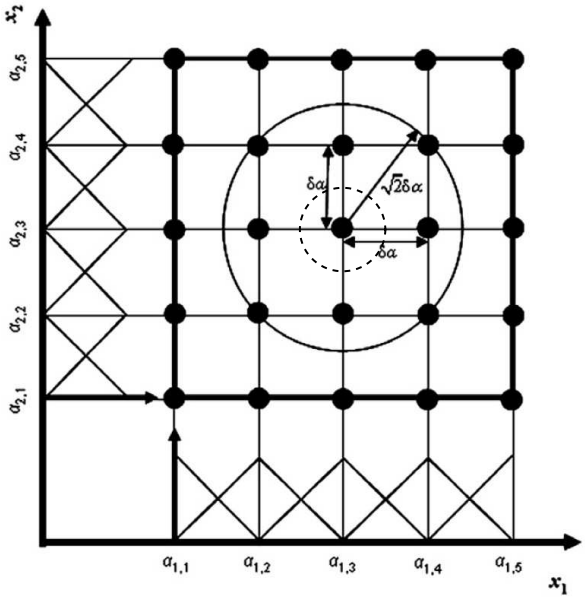



**Figure 2.7:** Display of a fuzzy partition between values of a two dimensional variable **x** [28]. On can see that the the selection of the membership function influences the number of fuzzy sets, that are activated. The dashed circle indicates the reach of the membership function of the fuzzy partitioning and the full circle the reach of the membership function if the euclidean distance had been used.

Figure 2.7 shows two important aspects about the fuzzy input space. Firstly the grid. Each variable $x_j$ spans a grid of $c_j$ nodes, in Figure 2.7 that's two variables, with a

grid size of 5. The size of the grid is therefore determined by how many nodes each variable should be partitioned over. After the grid is spanned, the actual $n$ observations of $x_j$ decide over which nodes are activated and how strongly they activate the center according to the membership function. The dashed circle in Figure 2.7 shows the range of values a variable must have in order to activate the grid point $(a_{1,3}, a_{2,3})$. After the grid and the activated grid points have been calculated the weights between the hidden and the output layers are regressed by determining the respective width of each activated node using the P-nearest neighbour statistics.

### P-nearest Neighbour Statistics and Weight Regression

The width $\sigma_l$ of each hidden node is calculated, so that it covers all input values that activate it. This procedure allows a smooth fit of the desired outputs [29].

$$\sigma_l = \left( \frac{1}{P} \sum_{p=1}^{P} \parallel \hat{\mathrm{x}}_l - \hat{\mathrm{x}}_p \parallel^2 \right)^{1/2} \tag{2.29}$$

Where $\hat{x}_l$ is the $l$th activated center point and $\hat{x}_p$ its P-nearest neighbour. That means that if for variable $x$ the $l$th center has been activated by the fuzzy partitioning method, there are $P$ more activated centers left for that variable. After the determination of the width of the hidden layer nodes $\sigma_l$ the weights of each output node $i$ with respect to the hidden node $l$ can be regressed. The linear function for the estimated output values $y_i^*$ is described by:

$$y_i^* = \sum_{l=1}^{L} w_{i,l} f_l(\nu) + \beta_0 \tag{2.30}$$

With $\beta_0$ being the bias, $w_{i,l}$ the weight of the output $i$ and the hidden node $l$ and $f(\nu)$ the radial basis function (RBF). In this work the Gaussian function has been chosen, as suggested by [27]:

$$f(\nu) = \exp\left( -\frac{\nu}{\sigma_l^2} \right) \tag{2.31}$$

Here the parameter $\nu$ is the euclidean distance from the input value $x_i$ to the centers it activated $\hat{x}_{l,i}$.

$$\nu^2 = \sum_{i=1}^{n} (x_i - \hat{x}_{l,i}) \tag{2.32}$$

To regress the weights $w_{i,l}$ the OLS regression is used as in Equation (2.22). The validation happens by using the same activated nodes $\hat{x}_l$ and widths $\sigma_l$ but calculating $\nu^2$ with the observations in the validation set.

There are several key points that need to be understood when dealing with ANNs. The first one being the training. Based on the resolution of the input variables, a grid of inactive nodes is built. The input data in the training set will then activate certain grid points in the multidimensional space. Imagine a four dimensional grid with 10 nodes

for each variable i.e. 10 intervals (compared to the 28 dimensional input space in this work this is quite small). Such a multidimensional grip consists of 10,000 nodes. Now take a training input of 100 observations. The performance of the validation is vastly influenced by how many nodes will be activated by the training set. Imagine every observation in the training set will activate a single node. This means that any data point, that is similar to the training set (e.g. data points in the validation set) would most likely also activate a new node, since there are 9,900 inactive nodes left. However, observations in the validation set do not activate new nodes but are given a value for $\sigma_l$ according to Equation (2.29). Chances that the observations in the validation set will lay far apart from the activated nodes are therefore relatively high. This means that any point in the validation set has a low distance value from the activated nodes and will perform badly when it comes to predicting the target variable of the validation set. The ANN has been over-fitted. It has only learned the training set by heart and was not able to generalise the input in the training to the validation input and therefore is not practically applicable. If the 100 observations in the training set activate significantly less then 100 nodes, the training input has been grouped and generalised. If now the validation set is representative of the training set, chances are high that the input of the validation set lays close to the activated nodes, which will result in a good validation performance. The number of activated nodes is therefore an important characteristic of the ANN and should be monitored in order to understand the behaviour of the training and validation. The distance of the grid points affect the number of activated nodes. This distance is a function of the resolution of the training input. Therefore, the resolution of the input variables has been an optimisation parameter together with the weights and the selection of molecular descriptors.

## 2.6 Mixed Integer Programming

Assume that one wants to build a model that only contains a fixed number of MDs in order to avoid over-fitting as explained in section 1.4.2 or to find the most important MDs according to their contribution to the output. This process is called mixed integer programming (MIP) . It can be displaced easily by using the example of a linear model:

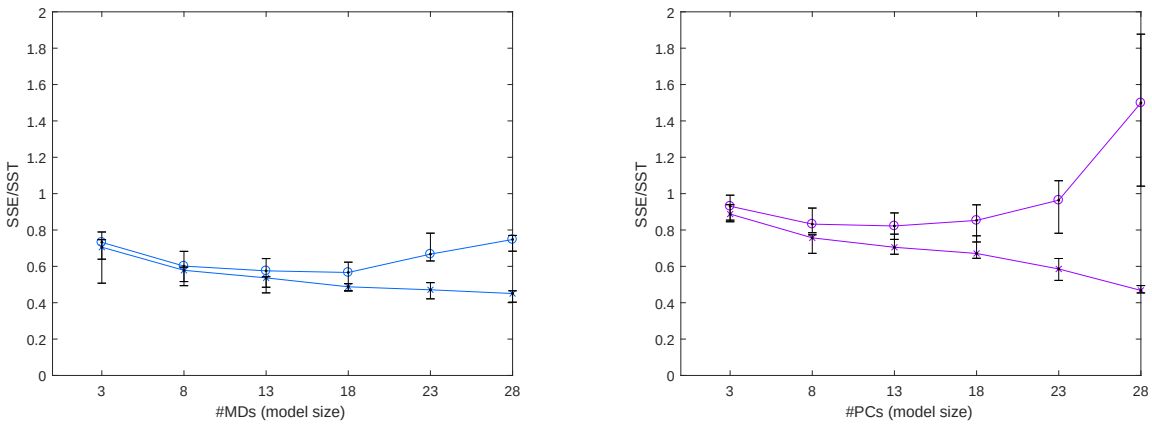
$$y^* = \beta_0 + z_1 \cdot \beta_1 x_1 + \ldots + z_p \cdot \beta_p x_p \quad \text{s.t.} \quad z_{min} \leq \sum_i^p z_i \leq z_{max} \quad z_i \in \{0, 1\}, \tag{2.33}$$

where $z_i$ to $z_p$ are decision variables, taking on discrete values between zero and one, ergo they decide whether a MD is included in the model or not. The overall amount of MDs in the model is restricted by $z_{max}$ and $z_{min}$. Note that for the PLS regression $z_i$ does not correspond to a selected MD but to a principal component PC. For the implementation of MIP the genetic algorithm (GA) in MATLAB has been used. The exact functionality is explained in the MathWorks® documentation [30]. The GA has been used to do the parameter selection as well as fit the optimal parameters $\boldsymbol{\beta}^*$ at the same time and find the optimal resolution for the ANN regression.

# 3 Results and Discussion

In the following the results for the modelling are presented. There will be a comparison between the PLS and the ANN regression, an analysis of the different indicators as well as a comparison between the results of Wernet et al. Furthermore, they will be an assessment of the most important molecular descriptors as well as the ANN structure.

## 3.1 Comparing PLS and ANN Regression




(a) ANN (b) PLS

**Figure 3.1:** Screening of only 10 splits over the whole model dimension in steps of 5 MDs for the total ReCiPe score. One can see that the error behaves as expected for such a case, with a decreasing training error (over-fitting) and a minimum validation error at the ideal model size. (a) shows the model behaviour for the ANN regression, (b) for the PLS regression.

Before rigorous training, a first, more shallow screening has been performed for two reasons. The first one being selecting the ideal regression method, i.e. analysing which method is most promising for a deeper look. Second, the screening aimed to determine the ideal model size for further analysis. The model size has been increased from $MD = 3, 8, \ldots, 28$. These data have been used to find the area of best validation performance for each indicator separately and to look in said area more deeply. In Figure 3.1 the validation and training performance over the number of MDs is shown for the total ReCiPe score. One can see nicely, that the training and validation error behave

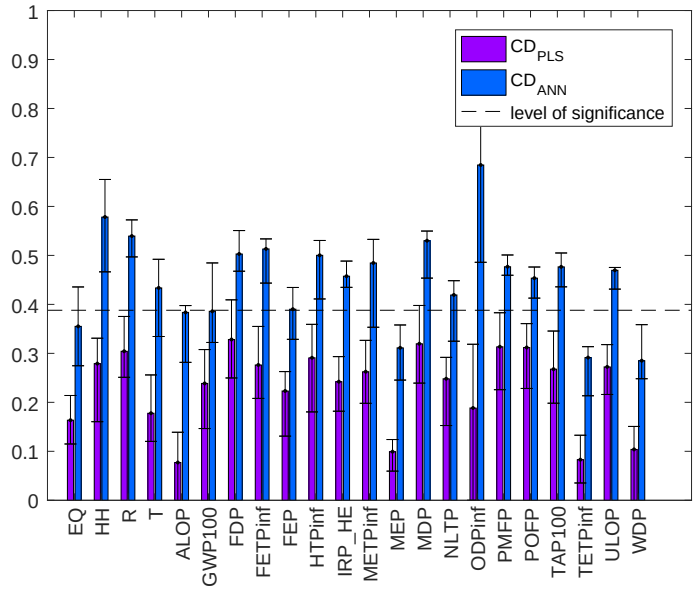



**Figure 3.2:** A comparison between the performance of the best CDs for a screening using the ANN ans PLS regression. The CDs are the best of the median of ten splits in a screening through $MD = 3, 8, ..., 28$, the errorbars indicate the 25% best and worst values. The ANNs perform significantly better, with CDs vastly over the level of significance for $\alpha = 0.05$.

as expected, see section 1.4.2. The training error decreases with increasing model size, while the validation error goes through a minimum. Figure 3.1(a) shows the ANN performance and Figure 3.1(b) the PLS performance. Figure 3.1 is an example to see the model behaviour over increasing model size. A summary of this screening is given in Figure 3.2. The median of the 10 splits has been taken and the maximum $CD$ is plotted in Figure 3.2. The error bars are 25% upper and 25% lower bound respectively. As mentioned in section 2.5, ANNs are expected to perform better than linear models, which is confirmed by the data in Figure 3.2. The $CD$ for the ANN is not only higher but also in most cases above the level of significance, other than the $CD$ for the PLS regression.

Because the ANN models perform significantly better, they have been investigated more thoroughly. In the wider area around the minimum validation error the reduced model has been calculated for 30 different training/validation splits over a range of MDs, that lay around the minimum validation error. Figure 3.3 shows the validation performance (blue bars) as well as the maximum training $CD$ (red squares). The bars have been grouped by decreasing validation error for the endpoint indicators R, HH, EQ (dark blue). The midpoint indicators (lighter blue) are put in groups that form the basis for the calculation of the endpoint indicators. Note that since the GWP100 influences the human health as well as the ecosystem quality, it appears twice in Figure 3.3. Be-

cause the water depletion as well as the marine eutrophication don't influence any of the endpoint indicators, they're grouped together at the end of the plot. To put the endpoint indicators and their respective midpoint indicators face-to-face as in Figure 3.3 is interesting from a practical point of view, regardless whether the results are good or not. Imagine someone wants to model the total ReCiPe score but doesn't succeed. If it was possible to model the other endpoint indicators, R, HH, EQ, one could compose a total ReCiPe score out of these other endpoint indicators according to the weighting approaches of the ReCiPe method. Similar thinking is applied below, when the results of Wernet et al. and the results of this work are put in contrast.

The optimisation of the parameters as well as the ideal model size has been performed by minimising the validation error. Originally it was intended to minimise the training error, however this procedure was unsuccessful. The major drawback is now that it is more likely that the training performance is worse than the validation performance. Obviously, when a model performs better in the validation than in the training it does not behave according to the expected performance as discussed in section 1.4.2 . It is hard to imagine that one would trust a model which cannot be trained well but can be validated well. If this happens a good or better validation performance seems somewhat coincidental since the model is predictive but not descriptive.

Figure 3.3 shows that the resource indicator (R) performs well in the validation ($CD_{val} = 0.53$), however has a particularly large error bar as well as a very bad training performance. Human health (HH) as well as the total (T) score performs satisfactorily ($CD_{val} = 0.52$ and $CD_{val} = 0.44$) concerning the validation performance and also good when comparing the validation with the training performance. The ecosystem quality (EQ) indicator is below the significance level ($CD_{val} = 0.36$). Among the midpoint indicators of the HH indicator the ODPinf, HTPinf, PMFP and IRP_ HE perform well ($CD_{val} = 0.54$, 0.52, 0.47 and 0.46). The GWP100 has a good validation performance but is trained badly. For the EQ indicator only the TAP100 midpoint indicator has a good performance ($CD_{val} = 0.47$). The others are either performing badly in the validation or the training. The two singled out midpoint indicators, MEP and WDP, as well as the midpoint indicators of the R endpoint indicator, MDP and FDP, don't perform well.

The resource indicator has a quite peculiar behaviour. When one looks at the MDs used in this work, compare Table 2.1, one would not expect a good performance when it comes to resource depletion. There are no MDs concerning metals or minerals such as iron or copper which is why a correlation between the MDs and the metal depletion (MDP) indicator would be unexpected. It seems convenient to apply similar thinking to the fossil depletion (FDP) indicator: Even though organic chemicals are often derived from fossil fuels, the amount of carbon, the molecular weight etc. does not give a insight of how much comes from fossil fuels and how much from biological feed stock. However, the fact, that the vast majority of molecules is derived from fossil feed stock de-validates that point. But, the amount of carbon stored in the molecule could be

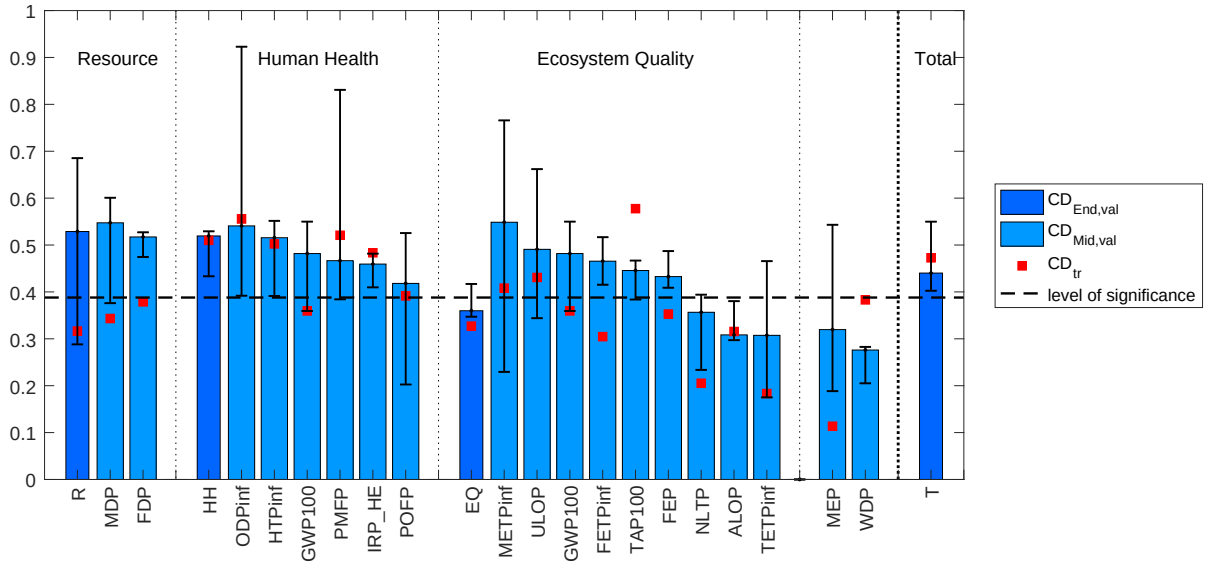



**Figure 3.3:** The models have been validated in the area of ideal model size for 30 splits. The median $CD$ of these 30 splits have been taken for all model sizes and the maximum value, at ideal model size, is displayed in the blue bars. The $CD_{End,val}$ is in a darker blue and the $CD_{Mid,val}$ in a lighter blue. The bars have been grouped to have an overview on how the endpoint indicators are composed. The red squares are the training performance at the same model size as the optimum validation performance. The error bars are 25% upper and lower bound. The dashed line shows the level of significance. For the endpoint indicators the HH and T indicators perform satisfactorily. The best and worst 10% have been left out in the calculation of the median to avoid extreme values.

small compared to the amount of carbon used in the production process which could disguise the relationship between the MDs and the FDP indicator. Still the model has succeeded to predict the data of the validation set. The low causality between the MDs in the resource depletion indicator confirms that even though the validation error is low the training error is much worse. Even though the R indicator could not be modeled well, it shows nicely, that R is composed of MDP and FTP. All three seem to have similar validation and training performance. It is unfortunate, that MDP and FDP don't perform better, since R could be calculated from reliable results of MDP and FDP.

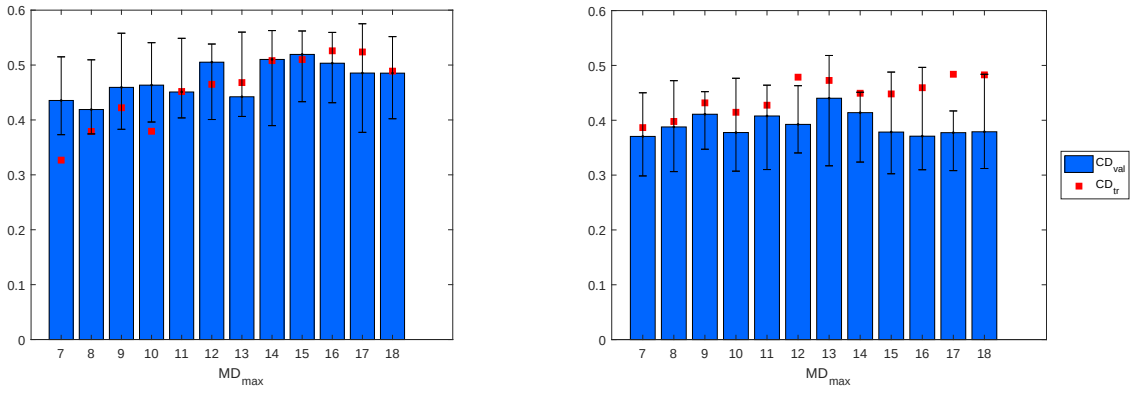
The human health (HH) indicator shows a good validation and training performance. The respective midpoint indicators appear to lay on average around on the same level as the HH indicator. However, some of them have much larger error bars, and the GWP100 performs significantly worse in the training set than in the validation set. This shows one major advantage of the ReCiPe method, which is a midpoint as well as an endpoint

oriented method. Predicting the endpoint scores may not be dependent on predicting well each one of the respective midpoint scores. A future modelling approach could try to improve the performance of the GWP100 and try to fuse the results of the midpoint indicators to return values for the HH indicator. It would be interesting to see, whether the directly modelled HH indicator performs better or worse, than the one that has been calculated from its midpoint indicators.

For the ecosystem quality (EQ) indicator a trend is observable as well. EQ behaves similarly to an average of its midpoint indicators. A successful modelling of the EQ indicator in this case is however not possible, because most of the midpoint indicators perform badly. If they performed well, the results for the midpoint indicator's prediction could be summed up to yield a reasonable EQ indicator.

The total indicator behaves as expected with a performance that lies between the other three endpoint indicators. This is a good sign because the total indicator can still perform well even though the resource and ecosystem quantity indicator perform badly. Again if the other three and point indicators performed good enough and the total indicator not, the R, HH and EQ could be summed up to result in the total indicator.

Figure 3.4 shows the coefficient of determination for the HH and T indicator over a model size of $MD = 7$ to 18. The best validation performance for HH is reached at $MD = 15$ and for the T indicator $MD = 13$. Consequently, this is the model size which will be referred to in the following as ideal model size and which will be used for the further analysis.



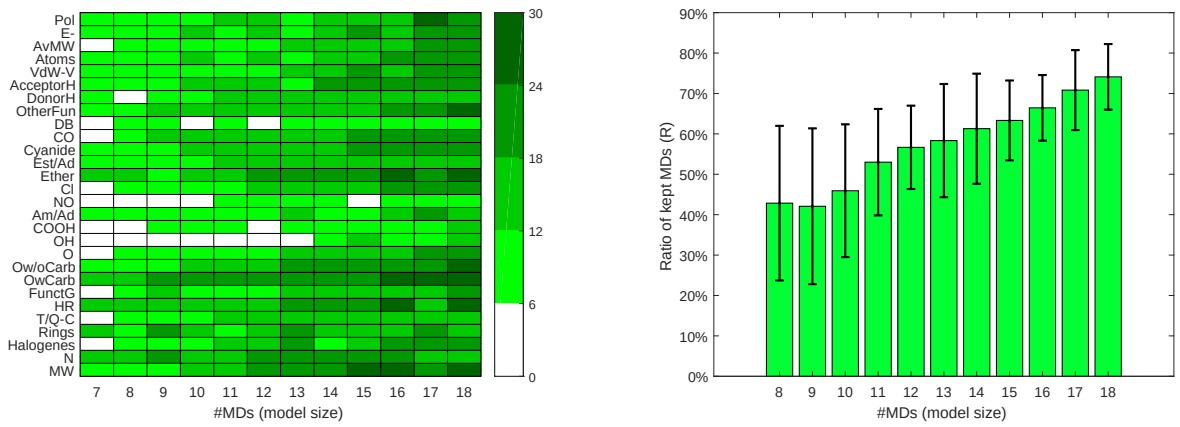

(a) HH (b) T

**Figure 3.4:** Coefficient of determination over a range of model sizes between $MD = 7$ to $MD = 18$. The ideal size of HH (a) is $MD = 15$ and for T (b) $MD = 13$.

## 3.2 Model Structure and Stability

The reliability of a model can be assessed through different means. One of them being looking for MDs that were included in the model with increasing model size and then analysing whether these MDs have been kept or discarded with increased model size. If a MD is included and then kept that is an indicator that it has a significant influence on the model output and that said influence is not coincidental. If a MD is included and in the next bigger model, i.e. a model grown by one dimension, is again discarded, then its addition in the previous steps seems somewhat arbitrary. A MD can be discarded for two reasons. First, there could be MDs that have a similar contribution to the output which would mean that several MDs are exchangeable. An example would be the number of atoms and the molecular weight, which are both size related MDs and will therefore have a similar contribution. The other reason could be that there is only an apparent contribution of this particular MD and it is just randomly included to be able to fit the model. However, an increase in validation performance would be a contradiction to this point. An increase in validation performance is always an indicator that a MD has an actual contribution to the output and has not only been included to for the purpose of over-fitting. Additionally, some MDs might contribute only a little to the output. Then the MDs with high contribution would be included in the model and then some with lower contribution. The selection of the lower contributing MDs may well depend on the split or noise which is why they may differ from split to split. Consider the total ReCiPe score. Figure 3.5(a) shows a map that colours the frequency of occurrence for every MD in different green shades from light to dark. When a MD is selected the respective rectangle is coloured green. Therefore, for a model size of seven, the ideal case would be that the same seven MDs are selected throughout the whole set of 30 splits. Stability of the model can be assessed by analysing how close one gets to that and how often MDs are discarded after being selected. In Figure 3.5(a) this means that when one row, that is one MD, is coloured green, i.e. selected in the reduced model, its colour shade should ideally become darker and darker from left to right. Note that this is an average and is not giving insight in the different splits. One can see that the colour is trending to become darker from left to right. However, there are many spots where for a bigger model less of the same MD selected, i.e. where the colour shade becomes lighter when going further to the right. This is an indicator, that MDs are discarded again after being selected and shows that the model stability is not ideal. A more detailed and quantitative overview is given in Figure 3.5(b) again for the total ReCiPe indicator. One can see how many of the included MDs are kept while the model increases from a model dimension of $MD = 7$ to 18. If the model size increases from $MD = 7$ to 8, there is a maximum of 7 MDs that can be kept for each split. The ratio of kept MDs on the y-axis in Figure 3.5(b) shows the average ratio over 30 splits of the amount of kept MDs ($R$) with respect to the model size, before it was increased:

$$R = \frac{\#MDs\ kept}{Model\ Size - 1} \tag{3.1}$$




(a) (b)

**Figure 3.5:** This figure shows trends in keeping or discarding MDs after they have been included in the model for the total ReCiPe indicator. With every increasing step, the MDs are newly selected. A perfectly stable model would always select the same MDs as before, while including one new MD. (a) shows the colour map over increasing model size $MD$. (b) shows an average percentage over 30 splits of how many MDs are kept in the next bigger model. The error bars are one standard deviation.

Figure 3.5(a) shows that there is a general trend to keep MDs which have been selected. But in Figure 3.5(b) it is clearly visible, that the amount of MDs kept is on average lower than the ideal case (100%) which shows that it is not safe to say that MDs are generally kept after being selected. That means, that even though there may be general preferences for MDs through the 30 splits, a single model will not generally value MDs higher than others. So, even though the total ReCiPe score has a good performance the future focus should be on increasing the stability in the MD selection process.

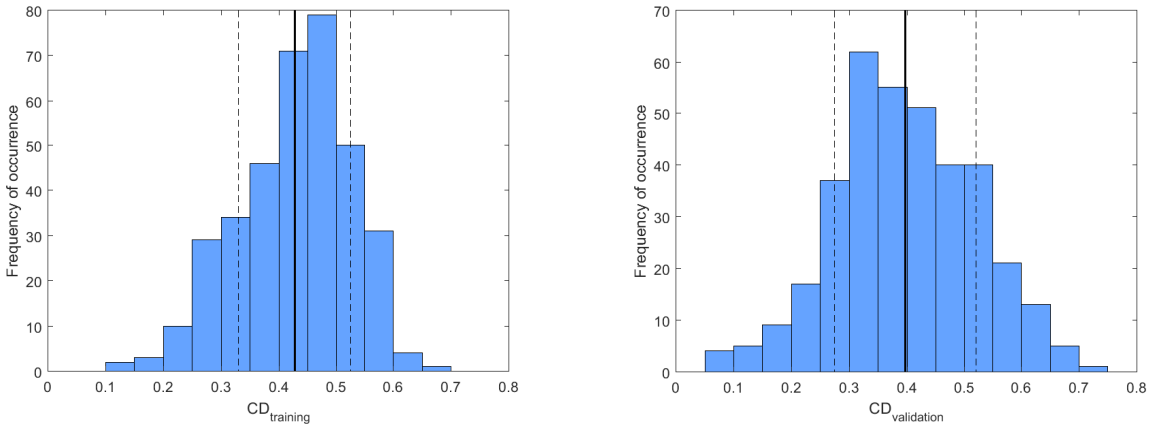
Another key indicator to understand the behaviour of an ANN model is the number of activated grid points. As discussed in section 2.5 an ANN spans a grid of center points, which are activated by the training input. If the validation set is similar to the training set concerning the input and output data, the validation input data will be close (see membership function in Equation 2.28) to the centers that have been activated by the training set. If a model is over-fitted, i.e. it has a good training and a bad validation performance, there will be many activated centers, close to the number of observations in the training set. Subsequently, the training set has been learned by heart and no general or predictive model has been built. Table 3.1 shows the mean activated grid size as well as the standard deviation for the endpoint indicators. Since the training set has 141 observations, the maximum grid size would be 141 (total over fitting). Interestingly, the mean grid size as well as the standard deviation is very close for all

endpoint indicators. Also, the mean grid size is quite small, compared to the maximum grid size of 141. Definitely there is no over fitting happening, as one can also see in the training CD in Figure 3.3. This is due to the fact that not the training error but the validation error has been optimised. The mean grid sizes are so similar and do not show any correlation with the model behaviour in Figure 3.3 so that the different behaviour of the different endpoint indicators cannot be explained by the size of the activated grid.

**Table 3.1:** The mean number of activated grid points for the four endpoint indicators ecosystem quality (EQ), human health (HH), resource depletion (R) and total score (T). Maximum grid size: 141.

| | EQ | HH | R | T |
|---|---|---|---|---|
| mean grid size | 24 | 28 | 26 | 25 |
| standard deviation | 5.28 | 4.93 | 4.12 | 4.8 |

One observation that indicates a rather stable behaviour of the model for the total ReCiPe score is the reproducibility of the results as well as the relatively small error bar. If one performs the calculations for the same 30 splits again, the model performance with always lie in the same area as in Figure 3.3. That has been successfully tested. As mentioned above the optimal validation performance has been searched in a wide valley of maximum MDs, i.e. maximum model size. The distribution of the training and the validation CD in this valley over 30 splits for the total ReCiPe score is plotted in Figure 3.6. The training CD tends to have more occurrences in the higher values as well as a higher mean $CD$, indicated by the bold vertical line. The dashed lines are one standard deviation. The training performance is on average better and also has a similar standard deviation. Still for both training and validation set the standard deviation is not large, which makes the model seem more reliable.



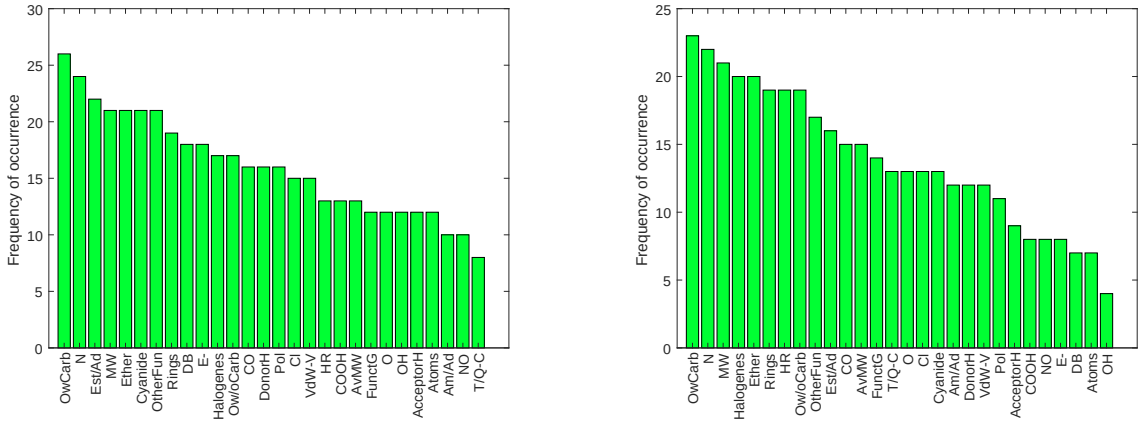

(a) (b)

**Figure 3.6:** The CD for the training (a) and validation (b) set for 30 splits and modeled sizes of $MD = 7, \ldots, 18$ for the total ReCiPe score. Bold line: $\overline{CD}$, dashed line: $\overline{CD} \pm s$

## 3.3 Analysis of the MDs

One very important aspect about the modelling is to investigate which MDs contribute most to the information stored in the input about the output. As such it is very important to be aware of the descriptors that are selected for the reduced model, i.e. the model with the lowest validation error. As displayed in Figure 3.4 the optimal model size lays at $MD = 15$ and 13 for the human health and total indicator. For these reduced models, the frequency of occurence of the MDs has been counted over the 30 splits and ranked. Figure 3.7(a) shows the frequency of occurrences for the HH indicator and Figure 3.7(b) for the T indicator.

When looking at Figure 3.10, there is a descent in the frequency of occurrences over the different MDs, that leads to the conclusion, that some MDs are more important than others. For the HH and T indicator the oxygen in carbonyl groups is the most frequently used MD. It is hard to say whether this influence is based on the oxidation process (Hydro formulation or others) or based on the alternated structure of the molecule concerning sterical and polarisability issues. It seems unrealistic, that the Ow-Carb ranks so high due to its production process, because carbonyls can be a feedstock for the production of carboxylic acids groups, which rank for both indicators among the lowest. If the production was responsible for the high contribution, the COOH would be expected to occur more often. For both indicators Nitrogen occurres the second most often as well. Nitrogen can be introduced in the model over different functional groups, such as amines, amides and nitro groups. Interestingly, the nitro groups rank low for both indicators. Also the amines and amides don't occur particularly often. If a reactant, catalysts, or energy in the production of amines, amides or nitro compounds

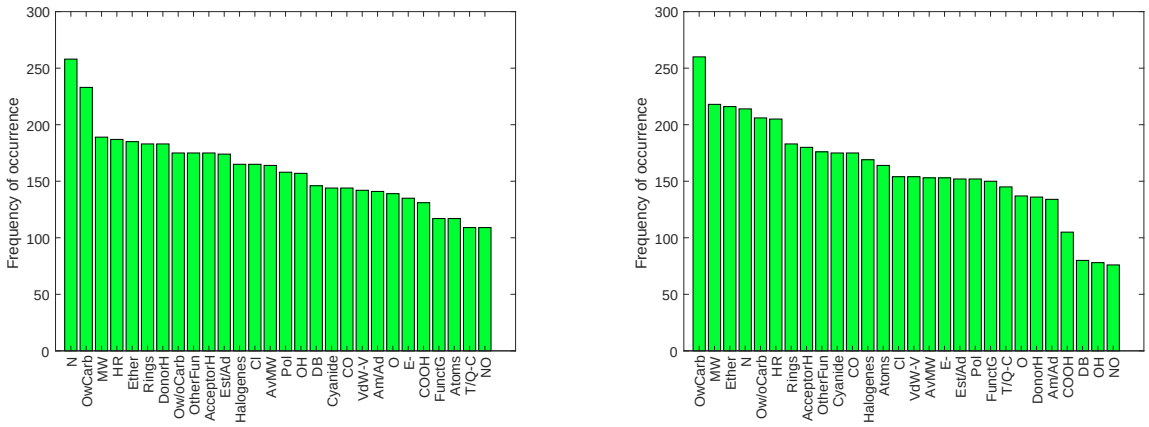



(a) HH (b) T

**Figure 3.7:** Frequency of occurrences for the MDs of the HH and T indicator.

were causing this high occurrence of the nitrogen MD, any of them would be expected to rank higher than the others. Additionally, there is a vast spectrum of production paths to synthesise any sort of nitrogen groups. Another explanation could be the sterical and polarizability behaviour of nitrogen. This however is regarded in the acceptor for hydrogen atoms (AcceptorH) as well as the polarisability and Van-der-Waals Volume which rank medium or low. Consequently, it is hard to find a causality between the indicator and the nitrogen MD even though there is a clear correlation. The molecular weight ranks fourth or third respectively. The heavier the molecule gets the more synthesis steps are needed in its production path and the more energy is needed to thermally treat such a molecule. Because in most chemical processes a thermal separation is part of the process the molecular weight is expected to have a high influence on energy-related indicators, which influence both the HH and the T indicator. The fact that the molecular weight ranks high makes it seem logical, that bigger molecules yield bigger scores for the HH and T indicator. Contradictive to this however is, that the VdW-V MD as well as the number of atoms don't appear as frequently for both indicators. Summarising, there is neither a clear tendency towards size related MDs nor to polarisability related MDs. This makes it seem more logical, that MDs related to specific molecules such as the amount of nitrogen atoms, the amount of oxygen atoms and carbonyl groups, or also the amount of ether groups contribute so much based on their production process rather than the way they shape or polarise the molecule. Additionally, when considering section 3.2 the selection of important MDs doesn't seem very reliable. It has been found that the MDs tend to be exchanged with other MDs when the model size is increased and therefore MDs that seem selected frequently in Figure 3.10 could be replaced when the model size is increased. This would mean that they only seem to be important. If the frequency of occurrence of the MDs is analysed over a whole range of model sizes, the ones that occur frequently have a higher credibility, since they have not been discarded that often. Of course this range of model sizes should also be a range of good validation

performance. Figure 3.8 shows the frequency of occurrence of the MDs over a range of model sizes of $MD = 7$ to 18. This range is the range that has been determined as area of maximum validation performance and therefore any MDs selected frequently in this area appear to be important.



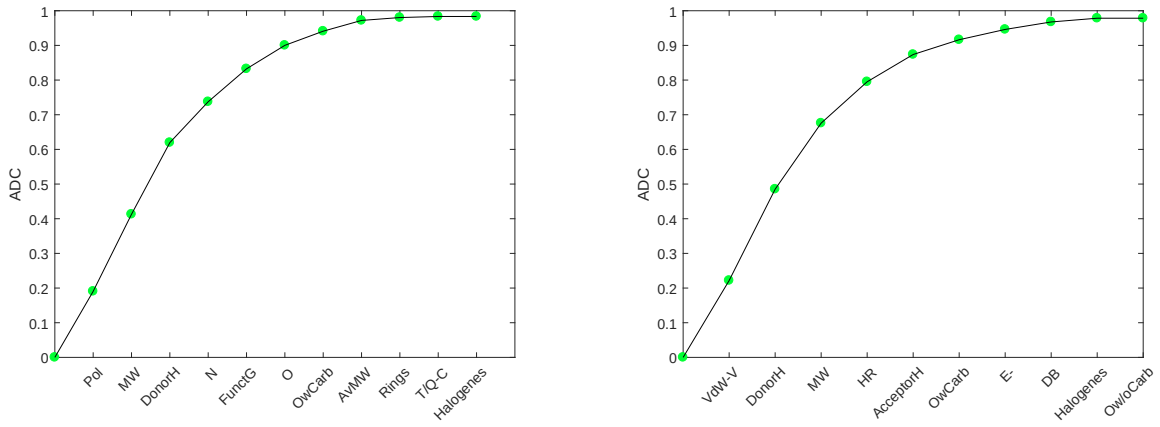

(a) HH (b) T

**Figure 3.8:** Frequency of occurrences for the MDs of HH and T for $MD = 7$ to 18. Other than in Figure 3.10 the distinction between the N and OwCarb in (a) and the OwCarb in (b) MDs is sharper now.

The frequency of occurrence is now less steep which has of course to do with the replacement of MDs when the model size is increased. However, for the HH indicator the MDs N and OwCarb separate themselves more clearly from the others. The same applies for the OwCarb MD for the T indicator. This leads to the conclusion that these MDs actually are highly important. The ones that are selected less frequently now will have experienced exchange with other while the model size grew. Ergo, they are less important.

Despite the conclusions drawn before, it is hard to derive a clear tendency about which MDs are most important and why. Another insight in the the importance of MDs is given by the ADC, as mentioned in section 2.3. The ADC gives the information content of one variable stored in one variable with respect to an output variable. It depends on the resolution of the discrete variables. To see how the ADC selects the most important MDs compared to the neural network, the optimised resolution from the neural network has been averaged and rounded to the nearest integer. The MD selection according to the ADC can be seen in Figure 3.9. There are two interesting observations. The first one is that size and polarisability related MDs contribute the most to the target variable (polarizability, molecular weight, donor hydrogen in Figure 3.9(a) and Van-der-Waal's volume, donor hydrogen and molecular weight in Figure 3.9(b)). Second, again the nitrogen and oxygen in carbonyl groups has been selected for the HH indicator and also

the oxygen in carbonyl again for the T indicator. The OwCarb MD has appeared in all three figures now: The analysis of the most frequently selected MDs for the ideal model size and the area of the ideal model size, as well as the analysis of the most important MDs according to the ADC. The same applies for the number of nitrogen atoms (here only for the HH indicator, because the N MD has not been selected for the ADC of the T indicator). The importance of OwCarb and N can therefore not be ignored and would be a most interesting part of future investigation and modelling. The tendency of the ADC to include mass and size related MDs is only confirmed for the MW, which is selected frequently. The fact, that it doesn't appear as frequent as the OwCarb and N MD could be explained that it is replaced with other, similar MDs (such as the number of atoms or the average molecular weight). It is definitely a factor when it comes to modelling ReCiPe indicators and should always be considered in future applications.

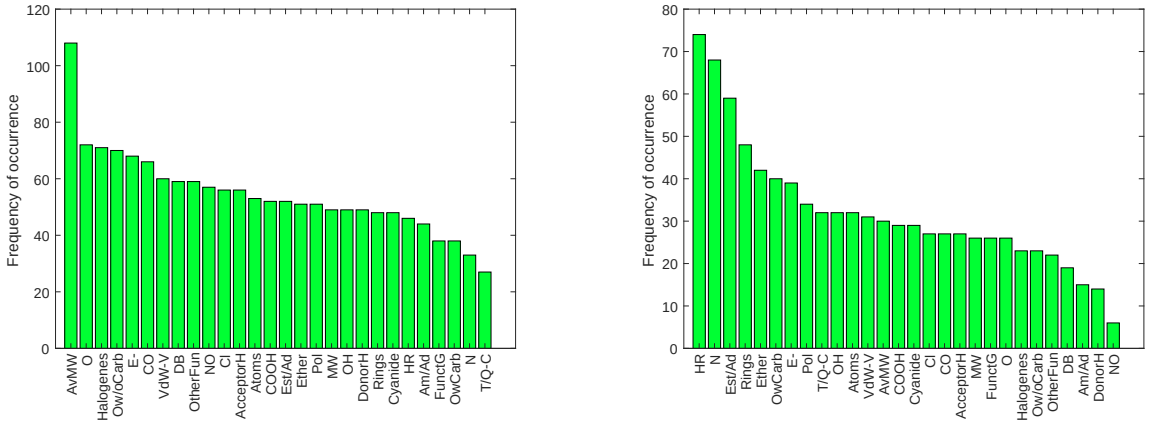



(a) HH (b) T

**Figure 3.9:** The ADC determines, which input variables contribute how much to the target variable in terms of entropy.

When looking at the results for the validation and training performance in Figure 3.3, it could be explained to some extend, why several indicators behave better than others. The analysis of the MDs gives another opportunity to approach an explanation. One of the best performing midpoint indicators was the ODPinf, which describes the ozone depletion. Figure 3.10(a) shows that the average molecular weight is distinctly more frequently selected than other MDs. A reasonable question would be, if such a distinction applies mostly to well predicted indicators. I.e: when MDs are clearly selected, can the indicator be modelled well? This is contradicted by Figure 3.10(b). It shows the frequency of occurrence for the MEP indicator, which performed very badly in terms of training and validation. Still there is a clear tendency to include the MDs: HR, N, Est/Ad and Rings in the model. Subsequently, if the selection of MDs is distinct (such as in Figure 3.10), it doesn't automatically mean, that the respective indicators have been predicted well or vice versa. Accordingly, there must be another reason, why some

indicators performed better than others. A detailed analysis could be performed by regarding the rigorous calculation methods that the ReCiPe indicators are based on. Maybe keeping these mechanisms in mind while performing the black box modelling will help future works to find a reasonable explanation for the different validation and training performances.




(a) ODPinf (b) MEP

**Figure 3.10:** Frequency of occurrences for the MDs of ODPinf and MEP. There is a distinctly higher selection of some MDs for both indicators, even though ODPinf performs well and MEP badly.

## 3.4 Comparing EI99 with ReCiPe

For a practical application, modelling the ReCiPe score directly might not be the ideal way to go. Since Wernet et al. have already successfully modelled the EI99 indicator, it could be more convenient to make use of his model, which has already been established. In order to do that it is possible to predict the EI99 indicator, and then correlate it with the ReCiPe indicator. Using the finechem tool, a set of 627 molecules has been used to build a correlation between the EI99 and the ReCiPe indicator. It was possible to use this bigger set because here there was no local restriction. That means that data from all over the world has been used and not just data from Europe. Figure 3.11 shows that the EI99 and ReCiPe indicator correlate very nicely for this set of chemicals. The function for the linear correlation is given in Equation (3.2). The coefficient of determination for this correlation is 0.92.

$$t_{ReCiPe} = 1.2862 \cdot t_{EI99} + 0.0285 \tag{3.2}$$

After this correlation has been established, the finechem tool has been used to predict

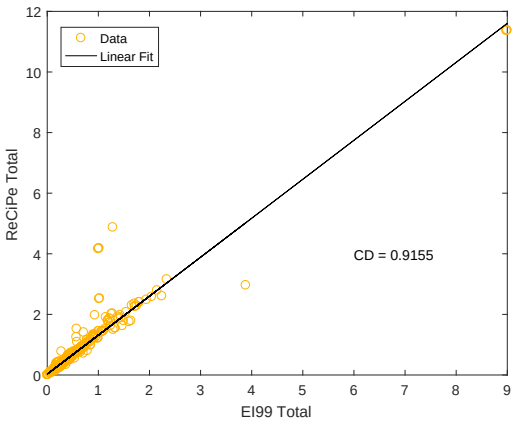



**Figure 3.11:** Correlation between the total score of the EI99 and ReCiPe indicator, with help from Dr. Sara Badr. Apparently, the two indicators correlate very well and a predictive model or the EI99 indicator could also be predictive for the ReCiPe indicator.

**Table 3.2:** Comparison of Wernet et al.'s results for the EI99 indicator [1] and the results for this work for the ReCiPe indicator. As the EQ and R indicators did not perform well, they are left out. Wernet's results are based on the the average of a LOOCV and the results of this work on the median of a validation, using a validation set.

| | Total | Human Health |
|---|---|---|
| $\bar{CD}_{LOOCV}$ (EI99) | 0.46 | 0.55 |
| $\hat{CD}$ (ReCiPe) | 0.44 | 0.52 |

the EI99 indicator. These predictions have been plugged in the above correlation to predict the ReCiPe indicator. Even though the finechem tool is based on Wernet et al.'s successful work, the data used in this work could only predict the EI99 with a performance of $CD = 0.0593$. Using the predicted EI99 scores yields a $CD$ for the ReCiPe prediction of 0.0697. In this case directly modelling the ReCiPe indicator is therefore a much better idea since it has been done successfully with $CD = 0.44$.

These results show that the used data set can significantly influence the model performance. Assuming that the above data set could be well predicted by the model of Wernet et al., what would yield better results: modelling the ReCiPe indicator directly or correlating it with the EI99 indicator predicted by Wernet et al.? In order to assess this question the two results can be compared. Still, there remains one hurdle. Wernet et al.'s results are based on a leave one out cross validation (LOOCV) [1]. This procedure tends to return better results. In order to have comparable data a LOOCV has been performed using the models that have been optimised above. In order to do that, a

fixed resolution of the variables has been selected, that is the average of the resolution of the 30 splits. However, this did not yield good performances for the validation or training. Next, the number of sets has become a optimisation parameter again, this time for the LOOCV. It has been tried to optimise for the training error (as it is the default procedure) as well as the validation error. As before, the optimisation for the training error did not result in anything near a good validation performance. Optimising the validation error yielded training performances that were again worse than the validation performances throughout the whole set of indicators. Consequently it is not possible to compare Wernet et al.'s results with these results quantitatively. Table 3.2 shows the results for Wernet et al.'s LOOCV $\bar{CD}_{LOOCV}$ and the median of the validation performances $\hat{CD}$ for the HH and T indicator. Even though the methods are different, the two results seem similar. Remember, that the LOOCV tends to have better results for the validation. As a conclusion, it has been successful to model the total and the human health indicator for the ReCiPe method directly. The results are in a comparable range to Wernet et al.'s results and therefore it is unlikely, that correlating the EI99 prediction with the ReCiPe will not render a better validation performance than predicting the ReCiPe indicator directly.

# 4 Conclusion and Outlook

This work was aiming to investigate the behaviour of linear and nonlinear models. Using a data set of 189 observations and 28 molecular descriptors it has been found that nonlinear artificial neural networks perform significantly better than linear models whose parameters have been regressed using PLS regression. Particularly the human health and total ReCiPe score where modelled keeping a good coefficient of determination for the validation set as well as for the training set. The optimisation has been done with regard to the validation error which has the advantage of yielding a good validation performance. However, it has been found that it may happen that the training performance is worse in this procedure which makes the model impractical. It would be more straightforward to train the model by optimising the training error but this resulted in largely over fitted models that had no predictive power. For any future modelling it is interesting to know that optimising the validation error can yield good predictive performance but may result in useless models with a training error that is worse than the validation error. Consequently, even though it seems convenient to optimise the validation error the default should stay in optimisation with respect to the training error. Maybe in the future it will still be possible to avoid extreme over fitting as it has happened here and end up with models that have a coefficient of determination for the training set which is strictly better than for the validation set. The artificial neural networks have been trained by optimising the weights between the hidden and output layer. It has been tried to keep the resolution of the discrete input variables constant at the optimal resolution with respect to the entropy. However, similar as above this did not result in good models. As a result, the resolution has become a part of the optimisation. Even though this resulted in a better performance, it means that for every training validation split a different resolution will be selected. Meaning that there is no clear rule for setting the resolution but rather an empirical mean out of the 30 optimised resolution sets. Also here it is good to know that this approach has worked, but it would be more convenient to have an ANN regression done under constant resolution. Furthermore, this resulted in problems when a LOOCV has been performed. Also here the resolution had to be an optimisation parameter. For future works finding a fixed resolution could be a major task. It would make things easier because otherwise for every split there needs to be an optimisation with respect to the resolution which poses a contradiction to the objective of a general model.

Analysing the molecular descriptors it has been found that nitrogen, oxygen in carbonyl groups, and the molecular weight have a high influence. It is hard to say why nitrogen and oxygen in carbonyl groups have such influence while other atoms don't. However,

the selection of the molecular weight and also the fact that the ADC mainly selects size, mass and polarisability MDs, shows that these MDs do have a tendency to contribute a lot. For any future modelling they should definitely be included in a screening for optimal molecular descriptors. Generally the screening for optimal molecular descriptors has been challenged by the fact that many of them have been exchanged with other MDs when the model size has been changed. This makes it questionable if the MDs that seem important actually do contribute as much as it seems. Still, the above mentioned MDs have been big players in all three screenings that have been performed (The frequency of occurrence for the ideal model size, the frequency of occurrence for a range of model sizes of good validation performance and the ADC). The identification of important MDs would however be more reliable, if it was a possible to build models, that are more stable, when it comes to keeping MDs when the model size is increased. Together with a constant resolution, this stability could be a major scope for approaches that follow this work.

It has been a big success, that the total ReCiPe indicator could be predicted with a performance $CD = 0.44$, that is comparable to the performance of the prediction of the EI99 indicator $CD = 0.46$. The ReCiPe is a more up to data indicator and could be more prevalent for decision makers and legislation. It could be shown that it is better to predict the ReCiPe indicator directly, rather than through a correlation with the EI99 indicator. Still, it remains questionable if the predictive performance is good enough for a model that can be used in practise.

Concluding, two of the major goals of this work have been reached. Firstly, it has been found that it is possible to predict the total ReCiPe score with a satisfactory performance using artificial neural networks. Second, the structure of the ANNs as well as the selected molecular descriptors could be analysed. It was possible to find MDs that are most likely relevant and most important, structural weaknesses and room for improvement has been assessed.

# List of Figures

# List of Tables

# Bibliography


[1] Wernet, Gregor ; Hellweg, Stefanie ; Fischer, Ulrich ; Papadokonstantakis, Stavros ; Hungerbühler, Konrad: Molecular-structure-based models of chemical inventories using neural networks. In: *Environmental science & technology* 42 (2008), Nr. 17, S. 6717–6722

[2] `https://www.ethz.ch/content/specialinterest/chab/chemical-n-bioengineering/set-group/en/research/downloads/software---tools/fine-chem.html`

[3] Goedkoop, Mark ; Heijungs, Reinout ; Huijbregts, Mark ; De Schryver, An ; Struijs, Jaap ; Van Zelm, Rosalie: ReCiPe 2008. In: *A life cycle impact assessment method which comprises harmonised category indicators at the midpoint and the endpoint level* 1 (2009)

[4] Guinee, Jeroen B. ; Heijungs, Reinout ; Huppes, Gjalt ; Zamagni, Alessandra ; Masoni, Paolo ; Buonamici, Roberto ; Ekvall, Tomas ; Rydberg, Tomas: *Life cycle assessment: past, present, and future*. 2010

[5] Deru, M: US Life Cycle Inventory Database Roadmap (Brochure) / National Renewable Energy Laboratory (NREL), Golden, CO. 2009. – Forschungsbericht

[6] `http://www.ecoinvent.org/support/documents-and-files/information-on-ecoinvent-3/information-on-ecoinvent-3.html#2196`

[7] Arrhenius, Svante: XXXI. On the influence of carbonic acid in the air upon the temperature of the ground. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 41 (1896), Nr. 251, S. 237–276

[8] `https://www.esrl.noaa.gov/gmd/aggi/aggi.html`

[9] Box, George E. ; Draper, Norman R. u. a.: *Empirical model-building and response surfaces*. Bd. 424. Wiley New York, 1987

[10] Nahas, EP ; Henson, MA ; Seborg, DE: Nonlinear internal model control strategy for neural network models. In: *Computers & Chemical Engineering* 16 (1992), Nr. 12, S. 1039–1057

[11] Wernet, Gregor ; Papadokonstantakis, Stavros ; Hellweg, Stefanie ; Hungerbühler, Konrad: Bridging data gaps in environmental assessments: Modeling

impacts of fine and basic chemical production. In: *Green Chemistry* 11 (2009), Nr. 11, S. 1826–1831

[12] Yan, Xin ; Su, Xiaogang: *Linear regression analysis: theory and computing*. World Scientific, 2009

[13] Ballabio, Davide ; Manganaro, Alberto ; Consonni, Viviana ; Mauri, Andrea ; Todeschini, Roberto: Introduction to MOLE DB-on-line molecular descriptors database. In: *MATCH Commun Math Comput Chem* 62 (2009), S. 199–207

[14] Goos, Peter ; Meintrup, David: *Statistics with JMP: graphs, descriptive statistics and probability*. John Wiley & Sons, 2015

[15] Esbensen, Kim H. ; Guyot, Dominique ; Westad, Frank ; Houmoller, Lars P.: *Multivariate data analysis: in practice: an introduction to multivariate data analysis and experimental design*. Multivariate Data Analysis, 2002

[16] Mohamad, Ismail B. ; Usman, Dauda: Standardization and its effects on K-means clustering algorithm. In: *Research Journal of Applied Sciences, Engineering and Technology* 6 (2013), Nr. 17, S. 3299–3303

[17] Sprinthall, Richard C.: *Basic statistical analysis*. 2011

[18] `http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm`

[19] De Maesschalck, Roy ; Jouan-Rimbaud, Delphine ; Massart, Désiré L: The mahalanobis distance. In: *Chemometrics and intelligent laboratory systems* 50 (2000), Nr. 1, S. 1–18

[20] Shannon, Claude E.: A mathematical theory of communication. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5 (2001), Nr. 1, S. 3–55

[21] Borda, Monica: *Fundamentals in information theory and coding*. Springer Science & Business Media, 2011

[22] Sridhar, Dasaratha V. ; Bartlett, Eric B. ; Seagrave, Richard C.: Information theoretic subset selection for neural network models. In: *Computers & Chemical Engineering* 22 (1998), Nr. 4-5, S. 613–626

[23] Rosipal, Roman ; Krämer, Nicole: Overview and recent advances in partial least squares. In: *Lecture notes in computer science* 3940 (2006), S. 34

[24] `https://se.mathworks.com/help/stats/plsregress.html`

[25] De Jong, Sijmen: SIMPLS: an alternative approach to partial least squares regression. In: *Chemometrics and intelligent laboratory systems* 18 (1993), Nr. 3, S. 251–263

[26] Bhat, Naveen ; McAvoy, Thomas J.: Use of neural nets for dynamic modeling and control of chemical process systems. In: *Computers & Chemical Engineering* 14 (1990), Nr. 4-5, S. 573–582

[27] Sarimveis, Haralambos ; Alexandridis, Alex ; Tsekouras, George ; Bafas, George: A fast and efficient algorithm for training radial basis function neural networks based on a fuzzy partition of the input space. In: *Industrial & engineering chemistry research* 41 (2002), Nr. 4, S. 751–759

[28] Alexandridis, Alex ; Sarimveis, Haralambos ; Ninos, Konstantinos: A Radial Basis Function network training algorithm using a non-symmetric partition of the input space–Application to a Model Predictive Control configuration. In: *Advances in Engineering Software* 42 (2011), Nr. 10, S. 830–837

[29] Leonard, James A. ; Kramer, Mark A.: Radial basis function networks for classifying process faults. In: *IEEE Control Systems* 11 (1991), Nr. 3, S. 31–38

[30] `https://se.mathworks.com/help/gads/how-the-genetic-algorithm-works.html`

# 5 Appendix

**Table 5.1:** List of the chemicals used as well as their molecular descriptors taken from the finechem tool [2] as well as the mole DB database [13]

| Chemical Name | Molecular Descriptors | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MW | N | X | R | C | HR | FG | OwC | Ow/oC | O | OH | COOH | Am/Ad | NO | Cl | Eth | Est/Ad | Cy | CO | DB | OF | DH | AH | VdW | A | $\overline{MW}$ | E- | Pol |
| 'glyphosate' | 169 | 1 | 0 | 0 | 0 | 0 | 6 | 0 | 5 | 5 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 6 | 9.83 | 18 | 9.39 | 1.93 | 1.10 |
| 'chloronitrobenzene' | 157 | 1 | 1 | 1 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 9.61 | 13 | 1.21 | 1.39 | 9.92 |
| 'benzyl chloride' | 126 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.78 | 14 | 9.00 | 1.39 | 1.05 |
| 'benzal chloride' | 161 | 0 | 2 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.08 | 15 | 1.07 | 1.52 | 1.18 |
| 'dimethyl sulfoxide' | 78 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 5.39 | 10 | 7.80 | 1.01 | 6.39 |
| 'ethylene glycol diethyl ether' | 118 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1.12 | 22 | 5.36 | 2.18 | 1.22 |
| 'trichloropropane' | 147 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7.47 | 11 | 1.34 | 1.15 | 8.65 |
| '2,4-dichlorotoluene' | 161 | 0 | 2 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.08 | 15 | 1.07 | 1.52 | 1.18 |
| 'o-dichlorobenzene' | 147 | 0 | 2 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.18 | 12 | 1.23 | 1.23 | 1.00 |
| 'p-dichlorobenzene' | 147 | 0 | 2 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.18 | 12 | 1.23 | 1.23 | 1.00 |
| 'ethylene glycol dimethyl ether' | 90 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 8.02 | 16 | 5.63 | 1.61 | 8.72 |
| 'melamine' | 126 | 6 | 0 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 6 | 8.97 | 15 | 8.40 | 1.56 | 9.03 |
| 'butane-1,4-diol' | 90 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 8.02 | 16 | 5.63 | 1.61 | 8.72 |
| 'dioxane' | 88 | 0 | 0 | 1 | 0 | 2 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 7.42 | 14 | 6.29 | 1.42 | 7.95 |
| 'ethylene bromide' | 108 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 4.58 | 7 | 1.54 | 6.94 | 5.26 |
| '1-butanol' | 74 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 7.50 | 15 | 4.93 | 1.47 | 8.26 |
| 'o-chlorotoluene' | 126 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.78 | 14 | 9.00 | 1.39 | 1.05 |
| 'diethyl ether, without water, in 99.95% solution state' | 74 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7.50 | 15 | 4.93 | 1.47 | 8.26 |
| 'imidazole' | 68 | 2 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5.59 | 9 | 7.56 | 9.09 | 5.77 |
| 'phthalimide' | 147 | 1 | 0 | 2 | 2 | 1 | 5 | 0 | 2 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 3 | 1.12 | 16 | 9.19 | 1.65 | 1.14 |
| 'pyrazole' | 68 | 2 | 0 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5.59 | 9 | 7.56 | 9.09 | 5.77 |
| 'methylene diphenyl diisocyanate' | 250 | 2 | 0 | 2 | 2 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 2.04 | 29 | 8.62 | 2.94 | 2.10 |
| 'N-methyl-2-pyrrolidone' | 99 | 1 | 0 | 1 | 0 | 1 | 2 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 8.90 | 16 | 6.19 | 1.60 | 9.51 |
| 'alpha-naphthol' | 144 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1.29 | 19 | 7.58 | 1.89 | 1.35 |
| '1-pentanol' | 88 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.10 | 18 | 4.89 | 1.76 | 1.00 |
| 'propyl amine' | 59 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 6.39 | 13 | 4.54 | 1.26 | 7.06 |
| '1-propanol' | 60 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5.90 | 12 | 5.00 | 1.19 | 6.50 |
| '4-tert-butyltoluene' | 148 | 0 | 0 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.58 | 27 | 5.48 | 2.61 | 1.71 |
| 'captan' | 300 | 1 | 3 | 2 | 2 | 1 | 11 | 0 | 2 | 2 | 0 | 0 | 2 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 1 | 0 | 3 | 1.69 | 23 | 1.30 | 2.43 | 1.86 |
| 'triethylene glycol' | 150 | 0 | 0 | 0 | 0 | 0 | 4 | 2 | 2 | 4 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 1.22 | 24 | 6.25 | 2.45 | 1.31 |
| 'epichlorohydrin' | 92 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5.70 | 9 | 1.02 | 9.35 | 6.23 |
| 'o-nitrophenol' | 139 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 3 | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 9.73 | 15 | 9.27 | 1.59 | 9.88 |
| '2,4-dichlorophenol' | 163 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.69 | 13 | 1.25 | 1.36 | 1.05 |
| 'toluene diisocyanate' | 174 | 2 | 0 | 1 | 1 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 1.32 | 19 | 9.16 | 1.96 | 1.34 |
| 'maleic anhydride' | 98 | 0 | 0 | 1 | 0 | 1 | 3 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3 | 6.14 | 9 | 1.09 | 9.89 | 6.12 |
| 'anthranilic acid' | 137 | 1 | 0 | 1 | 1 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 1.08 | 17 | 8.06 | 1.74 | 1.12 |
| 'monoethanolamine' | 61 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 5.30 | 11 | 5.55 | 1.11 | 5.74 |
| 'o-aminophenol' | 109 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 9.30 | 15 | 7.27 | 1.51 | 9.74 |
| 'phthalic anhydride' | 148 | 0 | 0 | 2 | 2 | 1 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1.07 | 15 | 9.87 | 1.58 | 1.09 |
| '2-butanol' | 74 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 7.50 | 15 | 4.93 | 1.47 | 8.26 |
| 'o-chlorobenzaldehyde' | 140 | 0 | 1 | 1 | 1 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 9.70 | 13 | 1.08 | 1.34 | 1.02 |
| 'p-chlorophenol' | 128 | 0 | 1 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8.70 | 12 | 1.07 | 1.24 | 9.23 |
| 'chloropropionic acid' | 108 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 6.21 | 10 | 1.08 | 1.07 | 6.68 |
| '2-cyclopentone' | 82 | 0 | 0 | 1 | 0 | 0 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 | 7.31 | 12 | 6.83 | 1.20 | 7.74 |
| 'acetone cyanohydrin' | 85 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 7.30 | 13 | 6.54 | 1.31 | 7.74 |
| '2-methyl-1-butanol' | 88 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.10 | 18 | 4.89 | 1.76 | 1.00 |
| 'isobutanol' | 74 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 7.50 | 15 | 4.93 | 1.47 | 8.26 |
| '2-methyl-2-butanol' | 88 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.10 | 18 | 4.89 | 1.76 | 1.00 |
| 'methacrylic acid' | 86 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 2 | 6.82 | 12 | 7.17 | 1.23 | 7.19 |

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 'isohexane' | 86 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.02 | 20 | 4.30 | 1.92 | 1.13 |
| 'alpha-picoline' | 93 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 8.79 | 14 | 6.64 | 1.38 | 9.29 |
| 'naphthalene sulfonic acid' | 208 | 0 | 0 | 2 | 2 | 0 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 1.50 | 22 | 9.45 | 2.26 | 1.61 |
| 'isopropylamine' | 59 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 6.39 | 13 | 4.54 | 1.26 | 7.06 | |
| 'isopropanol' | 60 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5.90 | 12 | 5.00 | 1.19 | 6.50 | | |
| 'aminopyridine' | 94 | 2 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 8.19 | 13 | 7.23 | 1.30 | 8.54 | |
| '2-pyridinol' | 95 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 7.70 | 12 | 7.92 | 1.22 | 7.98 | | |
| '3-methyl-1-butanol' | 88 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.10 | 18 | 4.89 | 1.76 | 1.00 | | |
| 'p-nitrophenol' | 139 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 3 | 3 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 9.73 | 15 | 9.27 | 1.59 | 9.88 | |
| '4-methyl-2-pentanone' | 100 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1.01 | 19 | 5.26 | 1.86 | 1.10 | | |
| 'atrazine' | 215 | 5 | 1 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 1.64 | 27 | 7.96 | 2.73 | 1.73 | |
| 'acetaldehyde' | 44 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 3.71 | 7 | 6.29 | 7.09 | 3.98 | |
| 'acetic acid, without water, in 98% solution state' | 60 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 4.22 | 8 | 7.50 | 8.43 | 4.43 | | |
| 'acetone, liquid' | 58 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5.31 | 10 | 5.80 | 9.97 | 5.74 | |
| 'acetonitrile' | 41 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 3.59 | 6 | 6.83 | 5.99 | 3.77 | |
| 'acrolein' | 56 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 1 | 4.71 | 8 | 7.00 | 8.10 | 4.98 | |
| 'acrylic acid' | 72 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 2 | 5.22 | 9 | 8.00 | 9.43 | 5.43 | |
| 'acrylonitrile' | 53 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 1 | 4.59 | 7 | 7.57 | 6.99 | 4.77 | |
| 'aniline' | 93 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 8.79 | 14 | 6.64 | 1.38 | 9.29 | | |
| 'anthraquinone' | 208 | 0 | 0 | 3 | 4 | 0 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1.74 | 24 | 8.67 | 2.42 | 1.80 | | |
| 'benzaldehyde' | 106 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 9.31 | 14 | 7.57 | 1.40 | 9.74 | |
| 'benzene' | 78 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7.79 | 12 | 6.50 | 1.16 | 8.29 | |
| 'benzoic acid' | 122 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 9.82 | 15 | 8.13 | 1.53 | 1.02 | |
| 'bisphenol A, powder' | 228 | 0 | 0 | 2 | 3 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 2.08 | 33 | 6.91 | 3.27 | 2.20 | | |
| 'butyl acetate' | 116 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 1.06 | 20 | 5.80 | 1.99 | 1.15 | |
| 'butyl acrylate' | 128 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 | 1.16 | 21 | 6.10 | 2.10 | 1.25 | |
| 'chloromethyl methyl ether' | 80 | 0 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 4.70 | 8 | 10.00 | 8.35 | 5.23 | | |
| 'chloroacetic acid' | 94 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 4.92 | 8 | 1.18 | 8.74 | 5.30 | | |
| 'monochlorobenzene' | 112 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8.19 | 11 | 1.02 | 1.10 | 8.77 | | |
| 'trichloromethane' | 119 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4.27 | 5 | 2.38 | 5.69 | 5.13 | | |
| 'cyanoacetic acid' | 85 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 3 | 5.62 | 9 | 9.44 | 9.65 | 5.67 | | |
| 'cyclohexane' | 84 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.59 | 18 | 4.67 | 1.73 | 1.06 | | |
| 'cyclohexanol' | 100 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1.01 | 19 | 5.26 | 1.86 | 1.10 | | |
| 'cyclohexanone' | 98 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 9.50 | 17 | 5.76 | 1.67 | 1.03 | | |
| 'dichloromethane' | 84 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3.28 | 4 | 21.00 | 4.44 | 3.88 | | |
| 'dimethyl malonate' | 132 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 4 | 9.45 | 17 | 7.76 | 1.79 | 9.86 | | |
| 'dimethyl sulfate' | 126 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 4 | 6.93 | 13 | 9.69 | 1.40 | 7.74 | | |
| 'N,N-dimethylformamide' | 73 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 6.30 | 12 | 6.08 | 1.21 | 6.74 | | |
| 'DTPA, diethylenetriaminepentaacetic acid' | 393 | 3 | 0 | 0 | 0 | 0 | 8 | 0 | 10 | 10 | 0 | 5 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 13 | 2.81 | 50 | 7.86 | 5.25 | 2.92 | | |
| 'ethanol, without water, in 99.7% solution state, from ethylene' | 46 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4.31 | 9 | 5.11 | 8.97 | 4.74 | | |
| 'ethyl acetate' | 88 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 7.42 | 14 | 6.29 | 1.42 | 7.95 | | |
| 'ethyl benzene' | 106 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.10 | 18 | 5.89 | 1.74 | 1.18 | | |
| 'ethylene glycol' | 62 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 4.82 | 10 | 6.20 | 1.03 | 5.19 | | |
| 'formaldehyde' | 30 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2.11 | 4 | 7.50 | 4.21 | 2.21 | | |
| 'formic acid' | 46 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 2.63 | 5 | 9.20 | 5.55 | 2.67 | | |
| 'glycerine' | 92 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 3 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 6.93 | 14 | 6.57 | 1.45 | 7.40 | | |
| 'adipic acid' | 146 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 4 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 1.10 | 20 | 7.30 | 2.07 | 1.16 | | |
| 'hydroquinone' | 110 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 8.82 | 14 | 7.86 | 1.43 | 9.19 | | |
| 'p-nitrotoluene' | 137 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1.08 | 17 | 8.06 | 1.74 | 1.12 | | |
| 'nitrobenzene' | 123 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 9.22 | 14 | 8.79 | 1.45 | 9.43 | | |
| 'hydroxylamine' | 33 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 2 | 2.11 | 5 | 6.60 | 5.31 | 2.22 | | |
| 'methyl iodide' | 141 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 3.32 | 4 | 3.53 | 3.90 | 4.80 | | |
| 'isobutyl acetate' | 116 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 1.06 | 20 | 5.80 | 1.99 | 1.15 | | |
| 'phenyl isocyanate' | 119 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 9.70 | 14 | 8.50 | 1.42 | 9.98 | | |
| 'isopropyl acetate' | 102 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 9.02 | 17 | 6.00 | 1.71 | 9.72 | | |
| 'cumene' | 120 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.26 | 21 | 5.71 | 2.03 | 1.36 | | |
| 'methylamine' | 31 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 3.19 | 7 | 4.43 | 6.86 | 3.53 | | |
| 'methane sulfonic acid' | 96 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 4.82 | 9 | 1.07 | 9.84 | 5.53 | | |
| 'methyl methacrylate' | 100 | 0 | 0 | 0 | 1 | 0 | 3 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 | 8.42 | 15 | 6.67 | 1.52 | 8.95 | |
| 'methyl acetate' | 74 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 5.82 | 11 | 6.73 | 1.13 | 6.19 | | |
| 'methylcyclohexane' | 98 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.12 | 21 | 4.67 | 2.02 | 1.23 | | |
| 'triethyl amine' | 101 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1.12 | 22 | 4.59 | 2.13 | 1.23 | | |

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 'dimethylamine' | 45 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 4.79 | 10 | 4.50 | 9.74 | 5.29 |
| 'trimethylamine' | 59 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6.39 | 13 | 4.54 | 1.26 | 7.06 |
| 'dipropyl amine' | 101 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1.12 | 22 | 4.59 | 2.13 | 1.23 |
| 'diethanolamine' | 105 | 1 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 9.01 | 18 | 5.83 | 1.82 | 9.72 |
| 'diethylene glycol' | 106 | 0 | 0 | 0 | 0 | 0 | 3 | 1 | 2 | 3 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 8.53 | 17 | 6.24 | 1.74 | 9.17 |
| 'triethanolamine' | 149 | 1 | 0 | 0 | 0 | 0 | 4 | 0 | 3 | 3 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 4 | 1.27 | 25 | 5.96 | 2.53 | 1.37 | |
| 'acetanilide' | 135 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 1.19 | 19 | 7.11 | 1.90 | 1.25 | |
| 'o-cresol' | 108 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.90 | 16 | 6.75 | 1.59 | 1.05 | | |
| 'glyoxal' | 58 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 2 | 3.63 | 6 | 9.67 | 6.55 | 3.67 | |
| 'xylene' | 106 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.10 | 18 | 5.89 | 1.74 | 1.18 | | |
| 'pentaerythritol' | 136 | 0 | 0 | 0 | 1 | 0 | 4 | 0 | 4 | 4 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 4 | 1.06 | 21 | 6.48 | 2.16 | 1.14 | | |
| 'pentane' | 72 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8.59 | 17 | 4.24 | 1.63 | 9.58 | | |
| 'phenol' | 94 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8.31 | 13 | 7.23 | 1.30 | 8.74 | | |
| 'phenyl acetic acid' | 136 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 1.14 | 18 | 7.56 | 1.82 | 1.20 | | |
| 'benzyl alcohol' | 108 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 9.90 | 16 | 6.75 | 1.59 | 1.05 | | |
| 'piperidine' | 85 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 8.98 | 17 | 5.00 | 1.65 | 9.82 | | |
| 'propanal' | 58 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 5.31 | 10 | 5.80 | 9.97 | 5.74 | |
| 'propionic acid' | 74 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 5.82 | 11 | 6.73 | 1.13 | 6.19 | | |
| 'propylene glycol, liquid' | 76 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 6.42 | 13 | 5.85 | 1.32 | 6.95 | | |
| 'pyridine' | 79 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7.19 | 11 | 7.18 | 1.09 | 7.53 | | |
| 'purified terephthalic acid' | 166 | 0 | 0 | 1 | 2 | 0 | 2 | 0 | 4 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 1.18 | 18 | 9.22 | 1.90 | 1.21 | | |
| 'ethyl tert-butyl ether' | 102 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1.07 | 21 | 4.86 | 2.05 | 1.18 | | |
| 'tert-butyl amine' | 73 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 7.98 | 16 | 4.56 | 1.55 | 8.82 | | |
| 'tetrahydrofuran' | 72 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 6.90 | 13 | 5.54 | 1.29 | 7.50 | | |
| 'toluene, liquid' | 92 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.39 | 15 | 6.13 | 1.45 | 1.01 | | |
| 'trichloroacetic acid' | 163 | 0 | 3 | 0 | 0 | 0 | 4 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 6.30 | 8 | 2.04 | 9.36 | 7.03 | | |
| 'trifluoroacetic acid' | 114 | 0 | 3 | 0 | 0 | 0 | 4 | 0 | 2 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 4.55 | 8 | 1.43 | 9.96 | 4.24 | | |
| 'urea, as N' | 60 | 2 | 0 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4 | 3 | 4.10 | 8 | 7.50 | 8.42 | 4.22 | | |
| 'vinyl acetate' | 86 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 2 | 6.82 | 12 | 7.17 | 1.23 | 7.19 | |
| 'styrene' | 104 | 0 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1.04 | 16 | 6.50 | 1.55 | 1.11 | |
| '2-nitroaniline' | 138 | 2 | 0 | 1 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 1.02 | 16 | 8.63 | 1.66 | 1.04 | | |
| '4-tert-butylbenzaldehyde' | 162 | 0 | 0 | 1 | 3 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1.57 | 26 | 6.23 | 2.55 | 1.68 | |
| 'acetic anhydride' | 102 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 7.33 | 13 | 7.85 | 1.36 | 7.64 | |
| 'acetoacetic acid' | 102 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 2 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 7.33 | 13 | 7.85 | 1.36 | 7.64 | |
| 'acetyl chloride' | 78 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4.40 | 7 | 1.11 | 7.41 | 4.84 | |
| 'aclonifen' | 264 | 2 | 1 | 2 | 0 | 0 | 4 | 1 | 2 | 3 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 4 | 1.34 | 5 | 5.28 | 7.52 | 1.21 | |
| 'bromopropane' | 122 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 6.18 | 10 | 1.22 | 9.82 | 7.02 | |
| 'butadiene' | 54 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 5.79 | 10 | 5.40 | 9.64 | 6.29 | |
| 'butane' | 58 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6.99 | 14 | 4.14 | 1.34 | 7.81 | |
| 'butene, mixed' | 56 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 6.39 | 12 | 4.67 | 1.15 | 7.05 | |
| 'chloroacetyl chloride' | 112 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 4.80 | 6 | 1.87 | 6.78 | 5.33 | | |
| 'chlorothalonil' | 265 | 2 | 4 | 1 | 2 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 1.31 | 13 | 2.04 | 1.44 | 1.39 | |
| 'chlorotoluron' | 212 | 2 | 1 | 1 | 1 | 0 | 5 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 1.65 | 26 | 8.15 | 2.62 | 1.75 | | |
| 'cyanogen chloride' | 61 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 2.69 | 3 | 2.03 | 3.42 | 2.87 | | |
| 'dimethyl ether' | 46 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 4.31 | 9 | 5.11 | 8.97 | 4.74 | | |
| 'dimethyl sulfide' | 62 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 4.88 | 9 | 6.89 | 8.72 | 5.94 | | |
| 'dipropylene glycol monomethyl ether' | 148 | 0 | 0 | 0 | 0 | 0 | 3 | 2 | 1 | 3 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 4 | 1.33 | 26 | 5.69 | 2.60 | 1.45 | | |
| 'EDTA, ethylenediaminetetraacetic acid' | 292 | 2 | 0 | 0 | 0 | 0 | 6 | 0 | 8 | 8 | 0 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 10 | 2.03 | 36 | 8.11 | 3.80 | 2.10 | | |
| 'ethylamine' | 45 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 4.79 | 10 | 4.50 | 9.74 | 5.29 | | |
| 'ethylene dichloride' | 98 | 0 | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4.88 | 7 | 14.00 | 7.32 | 5.64 | | |
| 'ethylene oxide' | 44 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 3.71 | 7 | 6.29 | 7.09 | 3.98 | | |
| 'ethylene, average' | 28 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 3.20 | 6 | 4.67 | 5.76 | 3.53 | | |
| 'ethylenediamine' | 60 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 2 | 5.78 | 12 | 5.00 | 1.18 | 6.30 | | |
| 'glycine' | 75 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 | 5.22 | 10 | 7.50 | 1.05 | 5.43 | | |
| 'isoproturon' | 206 | 2 | 0 | 1 | 2 | 0 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 1.93 | 33 | 6.24 | 3.26 | 2.06 | | |
| 'lactic acid' | 90 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 3 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 | 6.33 | 12 | 7.50 | 1.26 | 6.64 | | |
| 'metaldehyde' | 176 | 0 | 0 | 1 | 0 | 4 | 4 | 4 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 1.48 | 28 | 6.29 | 2.84 | 1.59 | | |
| 'metamitron' | 202 | 4 | 0 | 2 | 1 | 3 | 9 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 5 | 1.11 | 3 | 6.73 | 4.86 | 9.61 | | |
| 'methyl ethyl ketone' | 72 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6.90 | 13 | 5.54 | 1.29 | 7.50 | | |
| 'methyl formate' | 60 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 4.22 | 8 | 7.50 | 8.43 | 4.43 | | |
| 'methyl tert-butyl ether' | 88 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 9.10 | 18 | 4.89 | 1.76 | 1.00 | | |
| 'metolachlor' | 283 | 1 | 1 | 1 | 2 | 0 | 4 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 3 | 2.40 | 40 | 7.08 | 3.98 | 2.58 | | |
| 'napropamide' | 271 | 1 | 0 | 2 | 2 | 0 | 3 | 1 | 1 | 2 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 3 | 2.50 | 41 | 6.61 | 4.06 | 2.65 | | |
| 'pendimethalin' | 281 | 3 | 0 | 1 | 2 | 0 | 3 | 0 | 4 | 4 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5 | 2.28 | 39 | 7.21 | 3.97 | 2.39 | | |
| 'phosgene, liquid' | 98 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3.20 | 3 | 3.27 | 3.90 | 3.57 | | |

| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 'propylene' | 42 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 4.79 | 9 | 4.67 | 8.64 | 5.29 |
| 'propylene oxide, liquid' | 58 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 5.31 | 10 | 5.80 | 9.97 | 5.74 |
| 'prosulfocarb' | 251 | 1 | 0 | 1 | 1 | 0 | 3 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 2.26 | 38 | 6.61 | 3.73 | 2.47 |
| 'sodium formate' | 68 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 2 | 5.71 | 15 | 4.53 | 1.42 | 6.34 |
| 'thionyl chloride' | 118 | 0 | 2 | 0 | 0 | 0 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 1 | 3.28 | 3 | 3.93 | 3.97 | 4.22 |
| 'vinyl chloride' | 62 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 3.89 | 6 | 1.03 | 6.07 | 4.39 |
| 'chlorodifluoromethane' | 86 | 0 | 3 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3.11 | 5 | 1.72 | 6.09 | 3.27 |
| 'methanol, from biomass' | 32 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 2.71 | 6 | 5.33 | 6.09 | 2.98 |
| 'acetylene' | 26 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.60 | 4 | 6.50 | 3.88 | 2.76 |
| 'methane, 96% by volume' | 16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.20 | 5 | 3.20 | 4.76 | 2.53 |
| 'tetrachloroethylene' | 165 | 0 | 4 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5.67 | 5 | 33.00 | 6.07 | 6.61 |
| 'methylchloride' | 50 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2.89 | 5 | 10.00 | 5.07 | 3.39 |
| 'trichloroethylene' | 131 | 0 | 3 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5.27 | 6 | 2.18 | 6.70 | 6.13 |

**Table 5.2:** Corresponding data to Figure 3.2 with the median of 10 split for the validation coefficient of determination $\hat{CD}_{val}$ for the PLS and ANN regression.

| Indicators | PLS | | | ANN | | |
|---|---|---|---|---|---|---|
| | $\hat{CD}_{val}$ | 25% ub | 75% lb | $\hat{CD}_{val}$ | 25% ub | 75% lb |
| EQ | 0.16 | 0.21 | 0.11 | 0.36 | 0.44 | 0.27 |
| HH | 0.28 | 0.33 | 0.16 | 0.58 | 0.66 | 0.47 |
| R | 0.30 | 0.38 | 0.25 | 0.54 | 0.57 | 0.50 |
| Total | 0.18 | 0.26 | 0.12 | 0.43 | 0.49 | 0.33 |
| ALOP | 0.08 | 0.14 | 0.00 | 0.38 | 0.40 | 0.28 |
| GWP100 | 0.24 | 0.31 | 0.15 | 0.39 | 0.48 | 0.32 |
| FDP | 0.33 | 0.41 | 0.25 | 0.50 | 0.55 | 0.47 |
| FETPinf | 0.28 | 0.35 | 0.21 | 0.51 | 0.53 | 0.44 |
| FEP | 0.22 | 0.26 | 0.13 | 0.39 | 0.43 | 0.33 |
| HTPinf | 0.29 | 0.36 | 0.18 | 0.50 | 0.53 | 0.41 |
| IRP_ HE | 0.24 | 0.29 | 0.18 | 0.46 | 0.49 | 0.43 |
| METPinf | 0.26 | 0.33 | 0.20 | 0.48 | 0.53 | 0.35 |
| MEP | 0.10 | 0.12 | 0.06 | 0.31 | 0.36 | 0.25 |
| MDP | 0.32 | 0.40 | 0.24 | 0.53 | 0.55 | 0.45 |
| NLTP | 0.25 | 0.29 | 0.15 | 0.42 | 0.45 | 0.32 |
| ODPinf | 0.19 | 0.32 | 0.00 | 0.68 | 0.95 | 0.49 |
| PMFP | 0.31 | 0.38 | 0.23 | 0.48 | 0.50 | 0.46 |
| POFP | 0.31 | 0.36 | 0.23 | 0.45 | 0.48 | 0.41 |
| TAP100 | 0.27 | 0.35 | 0.20 | 0.48 | 0.50 | 0.44 |
| TETPinf | 0.08 | 0.13 | 0.04 | 0.29 | 0.31 | 0.21 |
| ULOP | 0.27 | 0.32 | 0.22 | 0.47 | 0.48 | 0.43 |
| WDP | 0.10 | 0.15 | -0.04 | 0.29 | 0.36 | 0.25 |

**Table 5.3:** The results of the ANN regression for each indicator. $\hat{CD}$ is the median $CD$, which has been calculated, leaving the best and worst 10% out.

| | $\hat{CD}_{val}$ | $\hat{CD}_{tr}$ | 25% lb | 25% ub |
|---|---|---|---|---|
| Ecosystem Quality | 0.3599 | 0.3270 | 0.2881 | 0.6852 |
| Human Health | 0.5194 | 0.5100 | 0.4332 | 0.5095 |
| Resources | 0.5288 | 0.3166 | 0.3471 | 0.4170 |
| Total | 0.4402 | 0.4727 | 0.4778 | 0.5545 |
| agricultural land occupation | 0.2848 | 0.3677 | 0.2221 | 0.3166 |
| climate change (also global warming potential) | 0.5004 | 0.3762 | 0.4036 | 0.5305 |
| fossil depletion | 0.5301 | 0.3734 | 0.4815 | 0.5349 |
| freshwater ecotoxicity | 0.4583 | 0.3861 | 0.3733 | 0.4715 |
| freshwater eutrophication | 0.4445 | 0.3803 | 0.3354 | 0.4451 |
| human toxicity | 0.5029 | 0.4470 | 0.4257 | 0.5228 |
| ionising radiation | 0.4803 | 0.4926 | 0.4085 | 0.5036 |
| marine ecotoxicity | 0.5138 | 0.3697 | 0.4485 | 0.5374 |
| marine eutrophication | 0.3136 | 0.1076 | 0.2150 | 0.3032 |
| metal depletion | 0.5396 | 0.3397 | 0.4351 | 0.5340 |
| natural land transformation | 0.3894 | 0.2230 | 0.3015 | 0.2767 |
| ozone depletion | 0.5715 | 0.6016 | 0.4400 | 0.3356 |
| particulate matter formation | 0.4825 | 0.4467 | 0.4144 | 0.0978 |
| photochemical oxidant formation | 0.4269 | 0.3878 | 0.3948 | 0.2967 |
| terrestrial acidification | 0.4456 | 0.5776 | 0.3837 | 0.4669 |
| terrestrial ecotoxicity | 0.3074 | 0.1832 | 0.2026 | 0.3108 |
| urban land occupation | 0.4908 | 0.4308 | 0.4397 | 0.4656 |
| water depletion | 0.2761 | 0.3830 | 0.2050 | 0.2828 |

**Table 5.4:** Ideal Model size for the Total indicator: The number of sets over which the input variables are distributed for the ANN input. This is an optimisation parameter and the ANNs can be reproduced using these values.

| Input Variable | number of sets for each split | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 'MW' | 19 | 11 | 7 | 8 | 16 | 20 | 18 | 16 | 9 | 18 | 20 | 6 | 17 | 19 | 16 | 14 | 20 | 10 | 7 | 13 | 7 | 12 | 11 | 6 | 8 | 15 | 14 | 18 | 18 | 17 |
| 'N' | 6 | 4 | 6 | 6 | 4 | 5 | 7 | 8 | 8 | 5 | 8 | 8 | 10 | 5 | 4 | 9 | 4 | 8 | 10 | 4 | 8 | 7 | 6 | 5 | 7 | 9 | 8 | 8 | 9 | 9 |
| 'Halogenes' | 5 | 8 | 7 | 4 | 4 | 9 | 7 | 6 | 6 | 8 | 7 | 7 | 10 | 9 | 10 | 5 | 8 | 8 | 6 | 6 | 10 | 6 | 7 | 4 | 7 | 7 | 6 | 10 | 9 | 5 |
| 'Rings' | 6 | 10 | 6 | 8 | 10 | 6 | 8 | 5 | 9 | 8 | 8 | 9 | 4 | 8 | 6 | 6 | 10 | 6 | 10 | 8 | 5 | 6 | 6 | 6 | 8 | 10 | 5 | 4 | 9 | 7 |
| 'T/Q-C' | 10 | 10 | 8 | 7 | 4 | 4 | 5 | 9 | 8 | 5 | 7 | 6 | 7 | 7 | 9 | 6 | 10 | 5 | 9 | 6 | 5 | 7 | 10 | 6 | 7 | 4 | 9 | 6 | 5 | 7 |
| 'HR' | 8 | 4 | 6 | 6 | 6 | 6 | 7 | 9 | 9 | 9 | 8 | 6 | 6 | 7 | 5 | 4 | 9 | 8 | 5 | 7 | 8 | 7 | 6 | 9 | 9 | 5 | 9 | 6 | 8 | 9 |
| 'FunctG' | 5 | 9 | 6 | 7 | 7 | 5 | 8 | 9 | 7 | 5 | 10 | 8 | 10 | 9 | 7 | 6 | 6 | 6 | 10 | 5 | 8 | 6 | 9 | 4 | 4 | 6 | 6 | 7 | 9 | 8 |
| 'OwCarb' | 6 | 8 | 10 | 7 | 10 | 9 | 8 | 5 | 7 | 7 | 8 | 6 | 8 | 5 | 8 | 7 | 9 | 4 | 4 | 5 | 9 | 9 | 8 | 10 | 7 | 8 | 5 | 5 | 8 | 5 |
| 'Ow/oCarb' | 10 | 5 | 5 | 4 | 9 | 6 | 5 | 6 | 5 | 8 | 7 | 5 | 7 | 5 | 4 | 10 | 4 | 8 | 5 | 6 | 9 | 5 | 8 | 8 | 6 | 10 | 8 | 6 | 6 | 6 |
| 'O' | 9 | 6 | 5 | 4 | 9 | 5 | 9 | 5 | 7 | 6 | 9 | 10 | 10 | 4 | 4 | 7 | 7 | 9 | 10 | 7 | 7 | 8 | 5 | 7 | 5 | 10 | 7 | 9 | 5 | 7 |
| 'OH' | 8 | 7 | 4 | 5 | 10 | 9 | 6 | 6 | 6 | 9 | 10 | 10 | 6 | 4 | 10 | 10 | 7 | 6 | 7 | 6 | 6 | 7 | 10 | 10 | 6 | 6 | 9 | 9 | 6 | 5 |
| 'COOH' | 4 | 7 | 4 | 10 | 8 | 8 | 8 | 5 | 9 | 9 | 6 | 6 | 8 | 7 | 6 | 7 | 4 | 9 | 5 | 5 | 9 | 7 | 10 | 4 | 9 | 7 | 9 | 7 | 5 | 7 |
| 'Am/Ad' | 9 | 7 | 8 | 7 | 5 | 9 | 6 | 6 | 7 | 10 | 5 | 7 | 6 | 8 | 4 | 7 | 9 | 4 | 4 | 8 | 7 | 9 | 5 | 9 | 7 | 8 | 4 | 4 | 6 | 6 |
| 'NO' | 4 | 4 | 8 | 5 | 8 | 6 | 7 | 4 | 5 | 6 | 4 | 6 | 7 | 4 | 7 | 4 | 4 | 7 | 10 | 5 | 8 | 7 | 8 | 4 | 6 | 8 | 8 | 9 | 4 | 6 |
| 'Cl' | 6 | 10 | 4 | 6 | 10 | 9 | 6 | 6 | 6 | 7 | 6 | 8 | 8 | 4 | 6 | 10 | 7 | 6 | 10 | 5 | 5 | 7 | 8 | 5 | 7 | 7 | 8 | 8 | 5 | 7 |
| 'Ether' | 7 | 8 | 7 | 7 | 5 | 6 | 4 | 4 | 7 | 6 | 4 | 6 | 6 | 7 | 10 | 7 | 7 | 7 | 5 | 4 | 5 | 9 | 9 | 6 | 4 | 7 | 5 | 8 | 9 | 7 |
| 'Est/Ad' | 7 | 4 | 4 | 10 | 9 | 4 | 6 | 7 | 6 | 6 | 9 | 4 | 8 | 10 | 8 | 6 | 10 | 8 | 4 | 6 | 5 | 9 | 4 | 10 | 5 | 8 | 6 | 4 | 9 | 6 |
| 'Cyanide' | 7 | 6 | 7 | 7 | 9 | 7 | 7 | 5 | 9 | 4 | 4 | 6 | 6 | 4 | 7 | 8 | 6 | 7 | 4 | 9 | 8 | 9 | 5 | 7 | 9 | 6 | 8 | 4 | 8 | 7 |
| 'CO' | 7 | 5 | 5 | 9 | 8 | 7 | 5 | 9 | 9 | 6 | 10 | 10 | 5 | 7 | 8 | 7 | 4 | 5 | 4 | 8 | 7 | 7 | 8 | 9 | 6 | 7 | 6 | 5 | 4 | 6 |
| 'DB' | 7 | 6 | 5 | 6 | 5 | 9 | 7 | 4 | 6 | 9 | 5 | 6 | 7 | 4 | 8 | 8 | 4 | 6 | 9 | 9 | 9 | 5 | 7 | 6 | 6 | 6 | 9 | 8 | 7 | 5 |
| 'OtherFun' | 8 | 10 | 8 | 5 | 4 | 10 | 4 | 9 | 8 | 6 | 4 | 9 | 8 | 9 | 10 | 8 | 5 | 6 | 10 | 6 | 10 | 6 | 6 | 9 | 6 | 4 | 5 | 5 | 8 | 8 |
| 'DonorH' | 8 | 9 | 8 | 8 | 10 | 9 | 9 | 7 | 8 | 8 | 5 | 10 | 5 | 4 | 7 | 6 | 10 | 7 | 10 | 8 | 5 | 8 | 10 | 4 | 9 | 10 | 6 | 10 | 10 | 7 |
| 'AcceptorH' | 8 | 8 | 9 | 8 | 6 | 5 | 6 | 7 | 7 | 7 | 7 | 5 | 7 | 6 | 10 | 7 | 4 | 9 | 10 | 9 | 6 | 4 | 6 | 7 | 6 | 6 | 9 | 10 | 10 | 7 |
| 'VdW-V' | 5 | 15 | 6 | 14 | 16 | 19 | 10 | 16 | 10 | 10 | 14 | 12 | 19 | 6 | 12 | 6 | 9 | 18 | 5 | 15 | 11 | 14 | 17 | 10 | 9 | 11 | 14 | 13 | 10 | 7 |
| 'Atoms' | 6 | 10 | 5 | 8 | 4 | 4 | 4 | 10 | 4 | 6 | 7 | 4 | 9 | 7 | 10 | 8 | 5 | 8 | 8 | 10 | 8 | 8 | 10 | 5 | 5 | 4 | 7 | 9 | 8 | 9 |
| 'AvMW' | 19 | 11 | 17 | 15 | 6 | 4 | 18 | 12 | 14 | 7 | 10 | 16 | 5 | 8 | 15 | 15 | 19 | 13 | 10 | 7 | 8 | 8 | 9 | 5 | 10 | 8 | 11 | 12 | 7 | 11 |
| 'E-' | 6 | 5 | 7 | 17 | 15 | 16 | 6 | 16 | 12 | 5 | 5 | 19 | 11 | 7 | 8 | 11 | 18 | 19 | 15 | 10 | 19 | 19 | 5 | 11 | 18 | 14 | 14 | 11 | 7 | 11 |
| Pol' | 18 | 4 | 14 | 14 | 6 | 18 | 9 | 16 | 8 | 9 | 15 | 9 | 18 | 5 | 10 | 7 | 18 | 12 | 14 | 8 | 11 | 14 | 10 | 5 | 12 | 16 | 10 | 14 | 16 | 16 |

**Table 5.5:** Ideal Model size for the Total indicator: The number of sets over which the input variables are distributed for the ANN input. This is an optimisation parameter and the ANNs can be reproduced using these values.

| Input Variable | number of sets for each split | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 'MW' | 15 | 19 | 15 | 12 | 7 | 16 | 13 | 18 | 11 | 18 | 5 | 9 | 11 | 14 | 15 | 14 | 10 | 4 | 5 | 19 | 8 | 20 | 12 | 18 | 6 | 19 | 5 | 7 | 9 | 4 | |
| 'N' | 10 | 6 | 10 | 8 | 8 | 5 | 9 | 5 | 9 | 5 | 10 | 10 | 9 | 5 | 6 | 9 | 8 | 10 | 4 | 4 | 9 | 7 | 8 | 9 | 9 | 8 | 4 | 10 | 10 | 7 | |
| 'Halogenes' | 6 | 4 | 7 | 5 | 8 | 5 | 5 | 9 | 4 | 7 | 5 | 9 | 8 | 9 | 8 | 10 | 7 | 4 | 4 | 6 | 7 | 4 | 6 | 6 | 4 | 10 | 9 | 7 | 8 | 4 | |
| 'Rings' | 7 | 6 | 9 | 8 | 8 | 8 | 7 | 5 | 8 | 8 | 10 | 8 | 9 | 7 | 10 | 5 | 9 | 4 | 4 | 7 | 6 | 10 | 4 | 8 | 10 | 7 | 6 | 4 | 8 | 10 | |
| 'T/Q-C' | 9 | 8 | 7 | 8 | 5 | 8 | 5 | 6 | 8 | 7 | 10 | 7 | 9 | 7 | 10 | 9 | 10 | 10 | 4 | 6 | 9 | 6 | 10 | 5 | 10 | 9 | 5 | 8 | 9 | 6 | |
| 'HR' | 4 | 6 | 10 | 8 | 8 | 5 | 8 | 7 | 9 | 6 | 9 | 5 | 5 | 7 | 9 | 5 | 5 | 7 | 4 | 9 | 10 | 8 | 10 | 9 | 7 | 10 | 6 | 10 | 10 | 8 | |
| 'FunctG' | 9 | 9 | 4 | 7 | 4 | 4 | 4 | 9 | 8 | 7 | 7 | 5 | 6 | 10 | 6 | 5 | 8 | 4 | 9 | 4 | 6 | 7 | 9 | 5 | 5 | 9 | 4 | 9 | 4 | 5 | |
| 'OwCarb' | 10 | 9 | 10 | 8 | 9 | 6 | 8 | 4 | 8 | 7 | 9 | 6 | 8 | 7 | 7 | 10 | 10 | 10 | 5 | 7 | 10 | 9 | 5 | 4 | 10 | 9 | 9 | 9 | 9 | 4 | |
| 'Ow/oCarb' | 10 | 9 | 10 | 7 | 5 | 4 | 10 | 8 | 6 | 9 | 7 | 6 | 8 | 5 | 10 | 6 | 7 | 4 | 9 | 4 | 5 | 8 | 9 | 9 | 4 | 5 | 6 | 5 | 8 | 9 | |
| 'O' | 9 | 5 | 5 | 8 | 7 | 6 | 6 | 6 | 7 | 6 | 8 | 6 | 6 | 7 | 10 | 6 | 10 | 5 | 7 | 9 | 7 | 8 | 7 | 4 | 6 | 5 | 6 | 5 | 8 | 4 | |
| 'OH' | 6 | 7 | 8 | 8 | 8 | 4 | 6 | 7 | 10 | 8 | 10 | 5 | 10 | 4 | 8 | 7 | 6 | 7 | 6 | 5 | 6 | 8 | 9 | 6 | 7 | 8 | 7 | 7 | 7 | 9 | |
| 'COOH' | 10 | 4 | 4 | 7 | 8 | 4 | 8 | 7 | 7 | 4 | 8 | 7 | 5 | 8 | 8 | 7 | 8 | 4 | 10 | 10 | 7 | 5 | 4 | 10 | 6 | 5 | 9 | 10 | 7 | 6 | |
| 'Am/Ad' | 5 | 5 | 5 | 5 | 5 | 5 | 9 | 10 | 7 | 4 | 7 | 6 | 8 | 7 | 9 | 6 | 9 | 10 | 10 | 4 | 4 | 8 | 4 | 5 | 10 | 8 | 7 | 10 | 8 | 6 | |
| 'NO' | 6 | 6 | 6 | 7 | 10 | 6 | 4 | 7 | 5 | 7 | 8 | 6 | 8 | 7 | 4 | 8 | 9 | 5 | 4 | 10 | 6 | 8 | 10 | 10 | 4 | 9 | 6 | 8 | 8 | 7 | |
| 'Cl' | 10 | 6 | 5 | 6 | 9 | 8 | 7 | 7 | 9 | 8 | 8 | 8 | 6 | 6 | 5 | 9 | 6 | 5 | 4 | 7 | 6 | 6 | 9 | 6 | 5 | 6 | 7 | 6 | 6 | 8 | |
| 'Ether' | 7 | 7 | 10 | 7 | 10 | 7 | 7 | 9 | 7 | 8 | 8 | 7 | 4 | 7 | 4 | 9 | 8 | 10 | 4 | 4 | 8 | 9 | 8 | 5 | 8 | 6 | 10 | 8 | 10 | 10 | |
| 'Est/Ad' | 9 | 4 | 4 | 8 | 7 | 6 | 4 | 6 | 10 | 5 | 10 | 7 | 8 | 10 | 4 | 9 | 8 | 7 | 7 | 6 | 10 | 8 | 7 | 9 | 6 | 6 | 5 | 8 | 5 | 4 | |
| 'Cyanide' | 9 | 9 | 6 | 5 | 5 | 9 | 4 | 8 | 5 | 6 | 8 | 6 | 6 | 8 | 5 | 7 | 5 | 7 | 6 | 8 | 5 | 5 | 8 | 4 | 5 | 5 | 9 | 8 | 6 | 8 | |
| 'CO' | 4 | 8 | 4 | 5 | 9 | 7 | 6 | 6 | 5 | 5 | 7 | 10 | 7 | 8 | 6 | 8 | 6 | 9 | 7 | 9 | 6 | 6 | 10 | 8 | 5 | 7 | 9 | 9 | 6 | 5 | |
| 'DB' | 10 | 9 | 6 | 6 | 8 | 8 | 5 | 8 | 8 | 4 | 8 | 8 | 9 | 10 | 6 | 10 | 4 | 9 | 8 | 7 | 4 | 8 | 6 | 6 | 8 | 5 | 8 | 7 | 9 | 9 | |
| 'OtherFun' | 4 | 7 | 9 | 8 | 7 | 8 | 8 | 7 | 9 | 6 | 4 | 6 | 8 | 9 | 9 | 6 | 7 | 4 | 5 | 7 | 6 | 8 | 8 | 6 | 10 | 6 | 5 | 4 | 5 | 4 | |
| 'DonorH' | 10 | 8 | 10 | 8 | 10 | 6 | 8 | 9 | 5 | 8 | 6 | 6 | 6 | 9 | 8 | 5 | 7 | 7 | 8 | 6 | 9 | 7 | 9 | 6 | 7 | 10 | 5 | 6 | 5 | 4 | |
| 'AcceptorH' | 5 | 9 | 5 | 8 | 6 | 7 | 5 | 9 | 7 | 6 | 8 | 7 | 7 | 9 | 8 | 5 | 10 | 5 | 4 | 6 | 8 | 4 | 4 | 7 | 5 | 9 | 8 | 4 | 9 | 7 | |
| 'VdW-V' | 9 | 5 | 9 | 12 | 8 | 8 | 13 | 7 | 10 | 14 | 18 | 7 | 14 | 20 | 5 | 16 | 4 | 18 | 4 | 5 | 17 | 9 | 15 | 17 | 20 | 10 | 12 | 17 | 9 | 9 | |
| 'Atoms' | 9 | 7 | 9 | 10 | 5 | 5 | 6 | 4 | 6 | 8 | 8 | 7 | 7 | 6 | 10 | 6 | 5 | 7 | 5 | 5 | 4 | 5 | 6 | 9 | 7 | 9 | 5 | 4 | 9 | 10 | |
| 'AvMW' | 11 | 18 | 14 | 7 | 13 | 6 | 14 | 6 | 16 | 12 | 6 | 14 | 12 | 17 | 16 | 13 | 10 | 11 | 6 | 10 | 7 | 6 | 12 | 14 | 7 | 6 | 13 | 5 | 7 | 9 | |
| 'E-' | 9 | 11 | 13 | 7 | 6 | 15 | 15 | 8 | 9 | 6 | 11 | 15 | 15 | 5 | 13 | 7 | 13 | 15 | 15 | 12 | 9 | 12 | 20 | 17 | 15 | 18 | 12 | 10 | 12 | 16 | |
| Pol' | 4 | 17 | 12 | 16 | 10 | 16 | 8 | 8 | 6 | 14 | 13 | 11 | 11 | 12 | 8 | 17 | 10 | 7 | 11 | 19 | 10 | 15 | 10 | 10 | 20 | 17 | 20 | 20 | 16 | 10 | |

**Table 5.6:** Frequency of occurrence of the MDs for all indicators.

| | EQ | HH | R | Total | ALOP | GWP100 | FDP | FETPinf | FEP | HTPinf | IRP_HE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 'MW' | 155 | 189 | 184 | 218 | 57 | 67 | 99 | 60 | 64 | 44 | 56 |
| 'N' | 262 | 258 | 170 | 214 | 42 | 106 | 83 | 56 | 53 | 63 | 36 |
| 'Halogenes' | 158 | 165 | 121 | 169 | 54 | 54 | 83 | 76 | 74 | 51 | 77 |
| 'Rings' | 114 | 183 | 75 | 183 | 71 | 66 | 57 | 79 | 56 | 72 | 30 |
| 'T/Q-C' | 167 | 109 | 33 | 145 | 25 | 53 | 62 | 56 | 29 | 38 | 17 |
| 'HR' | 161 | 187 | 159 | 205 | 59 | 54 | 88 | 48 | 27 | 50 | 55 |
| 'FunctG' | 140 | 117 | 91 | 150 | 75 | 58 | 78 | 66 | 88 | 59 | 73 |
| 'OwCarb' | 254 | 233 | 97 | 260 | 88 | 88 | 97 | 51 | 79 | 81 | 58 |
| 'Ow/oCarb' | 178 | 175 | 119 | 206 | 49 | 57 | 86 | 51 | 74 | 60 | 78 |
| 'O' | 148 | 139 | 107 | 137 | 53 | 47 | 80 | 48 | 49 | 48 | 77 |
| 'OH' | 186 | 157 | 134 | 78 | 68 | 47 | 82 | 60 | 45 | 50 | 63 |
| 'COOH' | 123 | 131 | 41 | 105 | 56 | 42 | 51 | 43 | 32 | 37 | 32 |
| 'Am/Ad' | 145 | 141 | 88 | 134 | 32 | 35 | 57 | 45 | 34 | 48 | 52 |
| 'NO' | 114 | 109 | 17 | 76 | 34 | 24 | 38 | 27 | 34 | 26 | 50 |
| 'Cl' | 177 | 165 | 193 | 154 | 57 | 56 | 98 | 81 | 80 | 44 | 71 |
| 'Ether' | 152 | 185 | 70 | 216 | 58 | 56 | 89 | 51 | 59 | 54 | 53 |
| 'Est/Ad' | 133 | 174 | 83 | 152 | 47 | 48 | 65 | 44 | 41 | 53 | 40 |
| 'Cyanide' | 165 | 144 | 88 | 175 | 50 | 50 | 73 | 51 | 56 | 49 | 42 |
| 'CO' | 129 | 144 | 44 | 175 | 39 | 36 | 73 | 41 | 42 | 59 | 46 |
| 'DB' | 145 | 146 | 44 | 80 | 71 | 45 | 57 | 66 | 63 | 83 | 78 |
| 'OtherFun' | 150 | 175 | 115 | 176 | 67 | 48 | 90 | 49 | 53 | 89 | 47 |
| 'DonorH' | 207 | 183 | 60 | 136 | 57 | 39 | 68 | 35 | 59 | 26 | 48 |
| 'AcceptorH' | 179 | 175 | 148 | 180 | 54 | 70 | 85 | 44 | 68 | 46 | 81 |
| 'VdW-V' | 152 | 142 | 129 | 154 | 54 | 47 | 86 | 50 | 57 | 48 | 49 |
| 'Atoms' | 163 | 117 | 61 | 164 | 38 | 55 | 69 | 58 | 49 | 58 | 43 |
| 'AvMW' | 141 | 164 | 63 | 153 | 47 | 45 | 69 | 50 | 39 | 68 | 42 |
| 'E-' | 151 | 135 | 73 | 153 | 51 | 65 | 69 | 57 | 49 | 44 | 47 |
| Pol' | 151 | 158 | 93 | 152 | 47 | 42 | 68 | 57 | 47 | 52 | 59 |

| | METPinf | MEP | MDP | NLTP | ODPinf | PMFP | POFP | TAP100 | TETPinf | ULOP | WDP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 'MW' | 47 | 26 | 44 | 20 | 49 | 76 | 45 | 65 | 56 | 73 | 64 |
| 'N' | 47 | 68 | 56 | 61 | 33 | 83 | 67 | 84 | 75 | 57 | 56 |
| 'Halogenes' | 65 | 23 | 100 | 47 | 71 | 76 | 25 | 65 | 59 | 77 | 77 |
| 'Rings' | 43 | 48 | 55 | 33 | 48 | 83 | 63 | 52 | 50 | 68 | 55 |
| 'T/Q-C' | 36 | 32 | 17 | 11 | 27 | 39 | 50 | 31 | 52 | 35 | 45 |
| 'HR' | 71 | 74 | 57 | 27 | 46 | 45 | 28 | 48 | 54 | 32 | 48 |
| 'FunctG' | 80 | 26 | 60 | 18 | 38 | 73 | 34 | 58 | 51 | 67 | 63 |
| 'OwCarb' | 50 | 40 | 49 | 31 | 38 | 74 | 40 | 73 | 61 | 74 | 88 |
| 'Ow/oCarb' | 69 | 23 | 68 | 45 | 70 | 59 | 27 | 77 | 64 | 60 | 61 |
| 'O' | 57 | 26 | 63 | 39 | 72 | 61 | 28 | 47 | 65 | 65 | 53 |
| 'OH' | 56 | 32 | 72 | 25 | 49 | 38 | 33 | 45 | 43 | 62 | 29 |
| 'COOH' | 40 | 29 | 45 | 25 | 52 | 41 | 21 | 43 | 54 | 45 | 41 |
| 'Am/Ad' | 30 | 15 | 56 | 46 | 44 | 53 | 30 | 52 | 41 | 58 | 39 |
| 'NO' | 24 | 6 | 16 | 21 | 57 | 40 | 17 | 49 | 52 | 37 | 37 |
| 'Cl' | 84 | 27 | 90 | 38 | 56 | 63 | 25 | 70 | 42 | 90 | 80 |
| 'Ether' | 49 | 42 | 53 | 27 | 51 | 52 | 33 | 65 | 61 | 49 | 63 |
| 'Est/Ad' | 39 | 59 | 33 | 39 | 52 | 36 | 22 | 47 | 57 | 33 | 37 |
| 'Cyanide' | 53 | 29 | 32 | 23 | 48 | 41 | 21 | 52 | 28 | 39 | 28 |
| 'CO' | 60 | 27 | 46 | 21 | 66 | 35 | 26 | 48 | 45 | 41 | 37 |
| 'DB' | 80 | 19 | 59 | 27 | 59 | 43 | 11 | 39 | 58 | 67 | 69 |
| 'OtherFun' | 85 | 22 | 61 | 53 | 59 | 47 | 38 | 54 | 62 | 44 | 76 |
| 'DonorH' | 32 | 14 | 42 | 39 | 49 | 33 | 16 | 41 | 40 | 56 | 32 |
| 'AcceptorH' | 64 | 27 | 76 | 55 | 56 | 88 | 35 | 56 | 62 | 81 | 62 |
| 'VdW-V' | 50 | 31 | 47 | 20 | 60 | 45 | 38 | 50 | 54 | 43 | 61 |
| 'Atoms' | 44 | 32 | 48 | 31 | 53 | 44 | 32 | 53 | 54 | 37 | 52 |
| 'AvMW' | 56 | 30 | 55 | 24 | 108 | 63 | 26 | 56 | 71 | 36 | 49 |
| 'E-' | 54 | 39 | 53 | 23 | 68 | 49 | 31 | 35 | 48 | 38 | 48 |
| Pol' | 35 | 34 | 47 | 31 | 51 | 50 | 38 | 45 | 41 | 36 | 50 |