Imperial College

Doctoral Thesis

Modelling the free energy of solvation: from data-driven to statistical mechanical approaches

---

Author:

Nur Redzuan Nur Jazlan

Supervisor:

Professor Amparo Galindo

Professor Claire Adjiman

A thesis submitted for the degree of Doctor of Philosophy degree and diploma of Imperial College London

in the

Molecular Systems Engineering Group

Department of Chemical Engineering

South Kensington Campus

Imperial College London

London SW7 2AZ

United Kingdom

Declaration of Authorship

I, Nur Redzuan Nur Jazlan, declare that this thesis titled, "Modelling the free energy of solvation: from data-driven to statistical mechanical approaches" and the work presented in it are my own. I confirm that any ideas or quotations from the work of others, published or otherwise, are fully acknowledged.

Abstract

The Gibbs free energy of solvation ΔG_s° for a given solute in a solvent, usually considered at infinite dilution, provides a simple thermodynamic description of the solution and is related to numerous solvation properties. In the context of solution chemistry, it provides a route to understanding the effect of solvents on equilibrium constants and reaction rates. In the discovery of new drugs, the effectiveness of a drug depends in part on solubility and permeability, leading to the prediction of ΔG_s° values to be used frequently in quantitative drug design. Given the importance of the Gibbs free energy of solvation, many predictive tools were developed, spanning quantum mechanical (QM) methods, empirical methods, and classical methods. Of note, empirical methods are data-driven approaches through statistical learning.

In this work, we assembled a database of experimental Gibbs free energies of solvation and a corresponding set of 9 quantum mechanical (QM) solute descriptors and 12 bulk solvent descriptors. We also partitioned the ΔG_s° into an electrostatic term, ΔE^{el}, and a nonelectrostatic term G^{CDS} such that $G^{CDS} = \Delta G_s^\circ - \Delta E^{el}$. The electrostatic term ΔE^{el} is the difference between the electronic energies of a solute in a vacuum and solvent obtained though using the X3LYP/6-31 G(d,p) electronic structure method and the Polarizable Continuum Model (PCM). We then obtain a separate database of derived G^{CDS} energies alongside the ΔG_s° database which are used to develop models using statistical and regression methodologies such as partial least squares (PLS), quadratic partial least squares (QPLS) and automatic learning of algebraic models for optimisation (ALAMO).

We then carry out a systematic comparison of various activity coefficients, data-driven models, an equation of state, and a hybrid QM/activity coefficient model. Notable models include the Dortmund version of UNIFAC model (modUNIFAC (Do)), the statistical associating fluid theory (SAFT-γ Mie), and the conductor-like screening model segmented activity coefficients (COSMO-SAC). We carry out calculations for the free energy of solvation on a common data set of 404 solute/solvent pairs with examples such as alcohols, alkanes, and aromatic molecules as solutes and alkanes, alcohols and water as solvents. We also assess the strengths and weaknesses of each method based on the overall data set and for specific subsets of solute/solvent pairs (e.g., aqueous/nonaqueous pairs.)

Acknowledgements

While I have declared this thesis and the work presented within it to be my own (*which it is*), I attribute the completion of my PhD to the many people who supported me along the way. Firstly, I would like to offer my thanks and congratulations to my supervisors, Claire Adjiman and Amparo Galindo, for toughing it out while trying to get me to this point. It has been an uphill battle for both them and me. I will always be grateful for their teachings, their patience, and their understanding. You are my role models, and I only wish I could have made the process easier for you.

My ten years at Imperial consisted of both the best and the worst times of my life. I would like to thank Dr Colin P. Hale, my mentor and friend, who has dispensed invaluable advice to me and always listened for a knock on the window. The MSE group will forever be my perfect office, and I am so blessed to have worked there. I thank everyone in the group for connecting with me, laughing with me, and never telling me to stop talking. To Eliana and Mohamad — thank you for always being there.

To my friends around the world, thank you for being my home. To Elsie, thank you for everything. To my brother, where is the grape? To my sister, it's warm outside. Finally, to my father and mother — my heroes — I owe everything I am to you.

Finally, to myself in the past, present and future: It has, is, and will forever be a real ride.

Contents

List of Figures

List of Tables

Chapter 1

Foreword

1.1 Predictive tools for the Gibbs free energy of solvation

The Gibbs free energy of solvation, $\Delta G_{i,j}^{solv}$, is a fundamental thermodynamic property defined as the change in Gibbs free energy when a solute i is transferred from the ideal gas phase to a solvent j at a specified temperature and pressure (Cramer, 2004). It is related to thermodynamic quantities such as the solubility, Henry's constant, infinite dilution activity coefficient and partition coefficient of the relevant species. As such, it is relevant in a broad range of applications. For example, in the pharmaceuticals industry, the effectiveness of a drug is dependent on its solubility and permeability; therefore, the Gibbs free energy of solvation can be used for quantitative drug design (Lipinski et al., 1997). Moreover, in the context of solution chemistry, the Gibbs free energy of solvation influences equilibrium constants which in turn relate to reaction rates.

Experimentally, the Gibbs free energy of solvation is determined using the thermodynamic quantities mentioned above. Partition coefficients and Henry's constants can be directly converted by taking their logarithms and multiplying the universal gas constant a desired temperature into the solvation free energy while solubilities can be used in conjunction with solute vapour pressures to obtain the solvation free energy (Abraham et al., 1987c; Abraham et al., 1990; Ben-Naim, 2006; Marenich et al., 2012). Thus, several databases (Abraham et al., 1990; Plyasunov and Shock, 2000; Marenich et al., 2012; Moine et al., 2017; Duarte Ramos Matos et al., 2017) collect experimental partition coefficients, Henry's constants, solubilities and vapour pressures for large sets of solutes and solvents at dilute conditions; most commonly at 298 K and 1 bar, but often also at other thermodynamic conditions.

Despite the amount of experimental data present in Gibbs free energy of solvation databases,

the data of interest can be unavailable because the corresponding system may not have been measured yet or because often the solute of interest is not stable (i.e., reaction intermediates, transition states), or dangerous to handle. Furthermore, in the case of mixed solvents, measuring a potentially infinite number of concentrations and solvent combinations is infeasible. In this context, predictive computational tools with precision comparable to that of experimental measurements (1 kcal mol^{-1}) are of interest (Deublein et al., 2011; Glass et al., 2014). Numerical approaches are not subject to the same limitations as experimental measurement and can be used to predict solvation energies for unstable compounds at any condition, including continuous solvent concentrations in mixtures. This has seen their implementation in solvent and molecule design problems such as the screening of the solubility of potential drug molecules during the early stages of drug discovery (Lipinski et al., 1997) and the screening of promising solvent candidates for the acceleration of reaction kinetics (Struebing et al., 2013).

A range of predictive methods that span from empirically-based methods (usually quantitative structure-property relationships (QSPRs)) to ab initio quantum mechanical (QM) ones have been applied to calculate the free energy of solvation. Data-driven empirical methods relate molecular properties and the Gibbs free energy of solvation through a linear or non-linear expression (Abraham et al., 1987a; Abraham et al., 1987b; Famini and Wilson, 1999). QSPRs are simple to use but are wholly reliant on the initial set of experimental data used for training. Semi-empirical methods such as molecular simulations, equations of state, or activity coefficient models are based on physical theory and parameterised against experimental data. These models are less reliant on experimental data and with the right set of parameters, semi-empirical methods can be quite predictive tools (Borhani et al., 2019). Ab initio QM models consider solutes at a detailed electronic level surrounded by an explicit or implicit treatment of the solvent (Tomasi, Mennucci, and Cammi, 2005a). Explicit solvent models also consider the solvents at a detailed electronic level which gives a high level of accuracy but is computationally intensive. In contrast, implicit solvent models maintain the detailed treatment of the solute and account for any polarization or conformational changes induced by a field induced by the solvent dielectric via a bulk electrostatics term. Implicit solvation models (or continuum solvation models) are less computationally intensive while maintaining a high level of accuracy.

In recent work (Borhani et al., 2019), the most commonly used models have been reviewed, with the predictive capability of each model assessed using different performance metrics or

a specific set of experimental data. Thus, for each model, while metrics are describing their errors, a potential user cannot infer whether one model is genuinely superior to another due to a lack of a fair comparison. As such, systematic assessments of predictive tools using the same set of experimental data are required. Currently, there are systematic assessments between different tools (Klamt and Diedenhofen, 2015). However, there is only a handful that compares different models (Voutsas and Tassios, 1996; Fingerhut et al., 2017; Borhani et al., 2019; Nait Saidi, Mielczarek, and Paricaud, 2020) and none that compare models that span a broader range of predictive methods.

1.2 Scope

The first objective of this thesis is to develop a robust data-driven predictive tool for the free energy of solvation. To achieve this, several requirements need to be fulfilled. For data-driven models, there is no framework that aids in the explanation of physical phenomena as no physical theory is supplied. Instead, a mathematical framework that solely relates the target variable, the free energy of solvation, to a set of descriptor variables such as the boiling temperature, van der Waal's volume or dipole moments is employed. The choice of mathematical framework heavily influences model performance as it can support linear or nonlinear behaviours. Therefore, choosing an appropriate framework is a crucial part of the development of data-driven models. A database of experimental observations that is both reliable and sufficiently large is required to "teach" the framework about the relationships between the target and descriptor variables. Further, methods such as cross-validation prevent the overfitting of data-driven models are required. Therefore, the first part of this thesis utilises these three key aspects to develop a series of data-driven models for the free energy of solvation. After examining the performance of the data-driven models, a quantum mechanical description of the solute molecule is incorporated into the methodology to improve the performance of the models.

The second objective is to collate popular predictive tools that can predict the free energy of solvation from different classes of solvation models and benchmark their predictive performance. These include data-driven models, activity coefficient models, equations of state or *ab initio* quantum mechanical models. This benchmarking requires finding a common subset of binary solute/solvent systems that fits all models where the predictions of each model need

to be calculated and extracted. This study is carried out to assess the performance of the data-driven models from the first objective and expand the library of systematic assessments that compare a broader range of predictive tools. Therefore, a user can refer to the previous systematic assessments and use them as a guide for deciding the optimal model for their purpose.

1.3 Outline

After a review of the range of predictive tools found in Chapter 1, the scope of the thesis is to develop a set of generalised data-driven models for the prediction of the free energy of solvation and systematically assess its performance against popular models.

The focus of Chapter 2 is to highlight key systematic studies, introduce currently available experimental data and how to select an appropriate subset. Further, topics related to experimental data are discussed, such as standard states and scales, the origins of experimental data, and the classification of molecules. Critical aspects of predictive tools are also presented.

In Chapter 3, the development and testing of data-driven solvation models are presented. Three methodologies serve as the basis of these data-driven models, the partial least squares (PLS) (Wold, 1973), quadratic partial least squares (QPLS) (Wold, Kettaneh-Wold, and Skagerberg, 1989; García-Muñoz, 2020) and the automatic learning of algebraic models for optimisation (ALAMO) (Cozad, Sahinidis, and Miller, 2014; Cozad, Sahinidis, and Miller, 2015; Wilson and Sahinidis, 2017). The derivation of training and testing experimental data sets used for the development of the data-driven models are also presented.

In the Chapter 4, a combined quantum mechanical and data-driven approach is proposed. The state of hybrid solvation models is discussed with a focus on deriving the quantum mechanical aspects of the proposed hybrid model. The hybrid solvation model is also tested and optimised using the PLS, QPLS and ALAMO methodologies as a basis.

In Chapter 5, a systematic assessment of the chosen predictive tools are presented. This assessment involves benchmarking a range of predictive tools using a set of performance metrics. The methodology for selecting a nonaqueous subset and a subset containing aqueous experimental data is discussed. Computational details for the predictive and the testing metrics are presented. The results of the systematic assessment are shown for both subsets of

experimental data. The study includes the overall performance and case-specific performance of the predictive tools.

The thesis is concluded in Chapter 6 with a summary, the main contributions and recommendations for future work to expand on the scope of this thesis.

Chapter 2

Experimental database and selected predictive tools for the free energy of sovlation

2.1 Current state of systematic assessments of predictive tools for the free energy of solvation

Several works have presented an assessment of different methods for the prediction of the solvation free energy; however, there are few reported systematic comparisons for different methods assessed against the same experimental data set. Such comparisons are essential because they serve as benchmarks across the range of predictive tools. In terms of comparative studies, it is worth highlighting five studies, in particular: i) Voutsas and Tassios (1996) compared versions of the universal quasi-chemical functional group activity coefficients (UNIFAC) model and the Pierotti Deal and Derr (PDD) correlation; ii) Klamt and Diedenhofen (2015) compared versions of the conductor-like screening continuum solvation model for realistic solvation (COSMO-RS); (iii) Fingerhut et al. (2017) compared two UNIFAC-type models and two versions of the conductor-like screening segmented activity coefficients (COSMO-SAC) hybrid quantum mechanical/activity coefficient models; (iv) Borhani et al. (2019) compared two QSPRs and two ab initio quantum mechanical (QM) models and; (v) Nait Saidi et al. (2020) re-optimised the parameters for several models from the COSMO-SAC framework and assessed their performances against their original counterparts.

In the work of Voutsas and Tassios (1996), six versions of the UNIFAC model and the PDD correlation were assessed for the prediction of infinite dilution activity coefficients compared

against 600 experimental data points at different temperatures. The PDD correlation (1959) is a logarithmic correlation useful in predicting infinite-dilution activity coefficients in aqueous mixtures. The UNIFAC model is a group-contribution (GC) approach that partitions the molecular size and shape effects into a combinatorial term and the interactions between groups into a residual term (Fredenslund, Jones, and Prausnitz, 1975; Fredenslund et al., 1977). They studied the UNIFAC model of Hansen (1991), who developed temperature-dependent parameters for the original UNIFAC model, the model of Magnussen et al. (1981) that used liquid-liquid equilibria (LLE) experimental data to determine the parameters of the model, UNIFAC-LLE, the modified UNIFAC models (Lyngby (modUNIFAC (Lyngby))(Larsen, Rasmussen, and Fredenslund, 1987) and Dortmund (modUNIFAC (Do)) (Wittig, Lohmann, and Gmehling, 2003) which have include changes to the combinatorial term to better account for alkane and alcohol experimental data and interaction parameters to improve general performance, the model of Bastos et al. (1988), who developed a model specifically for the prediction of infinite dilution activity coefficient data, by offering a new parameter table (UNIFAC-γ^{∞}) and the model of Hooper (1988), who developed a model specifically for water/hydrocarbon LLE mixtures with a new combinatorial term and interaction parameters specific for these mixtures (modUNIFAC (Hooper)).

Voutsas and Tassios (1996) compared the models in terms of the prediction of infinite dilution activity coefficient, which is directly related to the free energy of solvation. They considered 600 experimental data points of binary mixtures at various temperatures, including alkane/alkane, nonaqueous polar, and aqueous solute/solvent pairs. The best overall performance was obtained with the modUNIFAC (Do) model, which delivered predictive calculations of the infinite dilution activity coefficient across all the solute classes with an average absolute relative error (AARE%) of 3-26% for the 600 data points considered. Furthermore, the modUNIFAC (Do) model was found to have an AARE% of 10-15% for nonaqueous mixtures except for the case of strongly associating acid/solvent mixtures. The UNIFAC-γ^{∞} model resulted in an overall AARE% of 9.8-78.6% but it is interesting to point out that in the comparisons carried out in the study of Voutsas and Tassios (1996), it delivered the smallest error for the case of non-polar alkane-alkane systems. The PDD correlation was found to deliver the most accurate predictions for the case of highly-nonideal aqueous systems; however, the PDD correlation was developed for the prediction of various solute series with water as a solvent.

The conductor-like screening model (COSMO) was developed by Klamt and Schuurmann (1993) as a modification to the general class of apparent surface charge dielectric continuum solvation models (ASMs) (Tomasi, Mennucci, and Cammi, 2005b). The main feature of ASMs is the embedding of a detailed quantum-mechanical electronic solute structure into dielectric continuum. The solute electron density and polarisation charges are iterated until the solute becomes self-consistent with the dielectric medium, resulting in changes to the conformation of the solute. The difference between COSMO and the general class of ASMs is that the scaled conductor boundary condition is employed instead of an exact dielectric boundary condition, as it is easier to implement in quantum chemical codes (Klamt, 2018). However, a drawback is that it cannot be used to distinguish between two solvents with identical dielectric constants. Moreover, the majority of dielectric constant data are at room temperature, limiting the range of application.

Instead of treating the solvent as a dielectric field, COSMO-RS was developed to treat the solute and solvent on equal, quantum-mechanical and statistical thermodynamics footing which allows for the prediction of mixture thermodynamics at varying temperatures. COSMO-RS models have also been shown to have excellent predictive capability in the prediction of Gibbs free energies of solvation across broad ranges of solutes in solvents (Bannan et al., 2016; Klamt, 2016; Zhang, Tuguldur, and Van Der Spoel, 2015; Zhang, Tuguldur, and Van Der Spoel, 2016; Pye et al., 2009). However, a limitation of COSMO-RS is that the QM calculations are performed only in the reference state of a conductor and not for real solvent polarity. This limitation prevented COSMO-RS from predicting properties such as solvatochromic effects, solvent shifts of spectra, or electronic responses. As an improvement, the direct COSMO-RS (DCOSMO-RS) (Sinnecker et al., 2006) incorporates a solvent response function, the sigma potential, to overcome this limitation.

Klamt and Diedenhofen (2015) carried out a comparison between the original COSMO, COSMO-RS, and DCOSMO-RS models using a vast experimental data set of Gibbs free energy of solvation data for 2346 solute/solvent pairs. The data set comprised a broad range of molecule classes, including alcohols, aromatics, and esters as solutes in a range of polar and apolar molecules, as well as aqueous solvents. Klamt and Diedenhofen found the COSMO-RS model achieved a mean unsigned error (MUE) of 0.42 kcal mol^{-1} and a R^2 value of 0.914, while the DCOSMO-RS model and COSMO models which had MUE values of 0.66 kcal mol^{-1} and 2.14 kcal mol^{-1} and R^2 values of 0.871 and 0.243, respectively.

The promising results obtained with the COSMO-RS family of models have also prompted the development of related methods that extend and correct its capability. Lin and Sandler (2002) presented the COSMO-SAC model from the group-contribution solvation (GCS) model (Lin and Sandler, 1999) into an activity coefficient model in the COSMO-RS framework that corrects the Gibbs-Duhem inconsistency of the original COSMO-RS model. Over the years, further improvements were made to the COSMO-SAC model, resulting in newer models such as the COSMO-SAC10 which sought to introduce temperature-dependent electrostatics and differentiated hydrogen bonding with respect to hydroxyl groups and the COSMO-SAC-dsp model which explicitly accounted for dispersion via a term derived from molecular dynamics simulations.

Fingerhut et al. (2017) benchmarked two COSMO-SAC models, COSMO-SAC10 (Hsieh, Sandler, and Lin, 2010) and COSMO-SAC-dsp (Hsieh, Lin, and Vrabec, 2014), against two UNIFAC models, the original UNIFAC (Fredenslund et al., 1977) and modUNIFAC (Do) (Gmehling, Li, and Schiller, 1993), on an experimental data set of 29173 infinite dilution activity coefficients in 10897 solute/solvent pairs for temperatures from 213.3 K to 576.15 K (Fingerhut et al., 2017). They found a definite improvement from the COSMO-SAC10 model to the COSMO-SAC-dsp model (mean absolute deviation (MAD) of 95% to 86%) and from the original UNIFAC model to the modUNIFAC (Do) model (MAD of 73% to 58%).

In the work of Borhani et al. (2019), the PLS and MLR models are compared against two ab initio QM models. A set of 33 experimental free energy of solvation data points, adapted from the work of Zanith and Pliego (2015), were used as a benchmark. The data set included a selection of solutes in acetonitrile, methanol and DMSO. Further, the ab inition QM models were SMD and SM8 methods calculated using the X3LYP/6-31G(d) and B3LYP/6-31G(d) levels of theory, respectively. The RMSE values of the PLS, MLR, SMD and SM8 models were 0.59, 0.71, 1.11, and 1.08 kcal mol^{-1}, whereas the MUE values of the models were 0,46, 0.59, 0.83, 0.79 kcal mol^{-1}, respectively. From these results, it is clear the PLS and MLR models significantly outperforms the SMD and SM8 models.

The work of Nait Saidi et al. (2020) involved several comparisons of predictive models for the free energy of solvation. These models include the Abraham solvation model, some from the COSMO-SAC family of models, a group contribution model from Moine et al. (2017) and the QSPR model of Borhani et al. (2019). The COSMO-SAC frameworks found in this work include the original COSMO-SAC model (Lin and Sandler, 2002), the COSMO-SAC

2006 model (Mullins et al., 2006), the COSMO-SAC dsp model (Hsieh et al., 2011; Hsieh, Lin, and Vrabec, 2014) and the COSMO-SAC 2010 (Hsieh, Sandler, and Lin, 2010). The COSMO-SAC framework is an *ab initio* quantum mechanical methodology that utilises sets of universal parameters that describe atomic properties. Examples of these atomic properties include the size of a molecular segment in terms of area and volume, the hydrogen bonding coefficient or effective hydrogen bonding distance. To improve the predictive performance, Nait Saidi et al. re-optimised the universal parameters of the COSMO-SAC 2002, COSMO-SAC 2006 and COSMO-SAC dsp models. Therefore, their work includes the unoptimised and re-optimised versions of the aforementioned COSMO-SAC models.

The first comparison was between the Abraham solvation model with the unoptimised original COSMO-SAC 2002 (Lin and Sandler, 2002), and the latest COSMO-SAC dsp Hsieh, Lin, and Vrabec, 2014 models using reference data from the CompSol database (Moine et al., 2017). The Abraham solvation model achieved an average absolute deviation (AAD) of 0.81 kcal mol^{-1} compared to AAD values of 0.59 and 0.72 kcal mol^{-1} for the COSMO-SAC 2002 and COSMO-SAC dsp models, respectively. The next comparison was between the group-contribution model of Moine et al. (2017), and the unoptimised and re-optimised versions of the COSMO-SAC 2002, COSMO-SAC 2006 and COSMO-SAC dsp models (using DMOL3 cavities). The group-contribution model had an AAD of 0.36 kcal mol^{-1} whereas the unoptimised COSMO-SAC 2002, COSMO-SAC 2006 and COSMO-SAC dsp models had AAD values of 0.443, 0.625 and 0.37 kcal mol^{-1}, respectively. The re-optimised versions of the latter three models achieved AAD values of 0.347, 0.321 and 0.367 kcal mol^{-1}, respectively. The change in errors for the COSMO-SAC 2002 and COSMO-SAC 2006 models was larger compared to the COSMO-SAC dsp, suggesting the first two models were initially poorly optimised. Finally, the prediction for infinite dilution activity coefficients of 50 solutes in water and hexane were carried out at 298.15 K using the COSMO-SAC 2002, COSMO-SAC 2006 and COSMO-SAC dsp models. The predictive performance of the unoptimised and re-optimised versions were compared against experimental values for the infinite dilution activity coefficients for the solutes in water and hexane. The AAD values for the unoptimised COSMO-SAC 2002, COSMO-SAC 2006 and COSMO-SAC dsp models were 1.45, 2.34 and 1.16 kcal mol^{-1}, respectively. In contrast, the AAD values for the re-optimised versions were 1.29, 1.28 and 1.3 kcal mol^{-1}, respectively. Therefore there was a slight improvement for the COSMO-SAC 2002 model, a significant improvement for the COSMO-SAC 2006 model and

a slight decrease in performance for the COSMO-SAC dsp model. Therefore, there was a definite improvement in predictive performance by re-optimising the universal parameters for the COSMO-SAC models.

UNIFAC-based approaches are popular tools in the prediction of infinite dilution activity coefficients (which relate to the free energy of solvation). However, since the underlying assumptions of UNIFAC models only consider the liquid phase, a user must either assume the solute in the ideal gas phase is an ideal gas (the fugacity of the pure solute is unity) or use the model in conjunction with another model that can model the solute in the gas phase at dilute conditions. In this context, equations of state (EoSs) which apply over a wide range of thermodynamic conditions and provide a consistent platform for both the liquid and vapor phases are also of interest if they can be used predictively. However, commonly in EoSs, molecules are sometimes not modelled with a explicit structure and only with a set of parameters, isomers cannot be distinguished from one another. Furthermore, in the case of UNIFAC-based approaches, which are based on the solution-of groups concept, they are unable to account for the proximity of other groups. Therefore unless second order-groups are considered, these models are also unable to distinguish isomers. Several group-contribution cubic EoSs such as the group-contribution EoS (Skjold-Jørgensen, 1984; Skjold-Jørgensen, 1988), the group-contribution associating EoS (Gros, Bottini, and Brignole, 1996; Gros, Bottini, and Brignole, 1997), the predictive Soave-Redlich-Kwong EoS (Holderbaum and Gmehling, 1991), the volume-translated Peng-Robinson EoS (Ahlers and Gmehling, 2001; Ahlers and Gmehling, 2002) and the predictive Peng-Robinson approach (Jaubert and Mutelet, 2004) have been presented over the years.

The statistical associating fluid theory (SAFT) (Chapman et al., 1989; Chapman et al., 1990) represents a different class of EoSs that explicitly account for hydrogen bonding and stem from the first-order thermodynamic perturbation theory (TPT1) for associating fluids developed by Wertheim (1984; 1984; 1986; 1986). In the original SAFT approach (Chapman et al., 1989; Chapman et al., 1990), a homonuclear molecular approach was proposed where molecules are represented as associating chains of bonded identical spherical segments. This original approach has been recast to incorporate a heteronuclear model resulting in a group contribution EoS. In the SAFT-VR-GC (Peng et al., 2009) and SAFT-γ (Lymperiadis et al., 2007; Lymperiadis et al., 2008; Papaioannou et al., 2014) approaches, different types of monomeric segments are used to describe the different chemical functional groups comprised

in a specific molecule. This allows for the prediction of thermodynamic mixture properties from experimental pure component data alone (Lymperiadis et al., 2007; Peng et al., 2009; Papaioannou et al., 2011; Ramos et al., 2011).

In a recent work by Hutacharoen et al. (2017), the SAFT-γ Mie EoS was used to predict solvation free energies of *n*-alkanes and alcohols in water. In the case of *n*-alkanes, there was excellent agreement with the experimental solvation energies for carbon numbers under 11. They note that the uncertainty in the experimental data is large for carbon numbers 11 and above due to uncertainties carried forward from solubility measurements. In the case of alcohols, the agreement between the predicted and experimental solvation free energies improved as the number of hydrocarbons in the solute molecules increased. This is attributed to the assumption that the transferability of functional groups is less applicable for the smaller polar molecules due to proximity effects (Hutacharoen, 2017).

As mentioned previously, Borhani et al. (2019) carried out a review of predictive tools for the free energy of solvation. Of note, Borhani et al. highlighted critical QSPR studies for the free energy of solvation; however, the models developed have limited applicability for the choice of solute and solvent as these models were developed for a single solute in a range of solvents or a range of solutes in a single solvent. Therefore, Borhani et al. proposed a new methodology to construct QSPR models that can be used for any combination of solute and solvent. This was done by incorporating 12 bulk solvent descriptors, and nine quantum mechanical solute descriptors such that the model had a QM description of the solute molecule and the bulk description of the solvent. These descriptions are not explicitly molecular but are properties such as dipole moments, van der Waal's volumes, or dielectric constants. Thus, since these properties are continous variables, a resulting model based on these descriptors is able to predict any combination of solute and solvent given the right solute and solvent input data. The partial least squares (PLS) and multivariate linear regression (MLR) methods were used in tandem with a database of 1800 experimental free energy of solvation data to develop the model, which included 21 solute and solvent descriptors. These models achieved a root mean square error (RMSE) of 0.52 and 0.58 kcal mol^{-1} and an MUE of 0.43 and 0.44 kcal mol^{-1} respectively when tested against a data set of 1777 experimental free energy of solvation data points.

The systematic studies highlighted in the current section have shown that there are few systematic assessments that encompasses the whole range of predictive tools. Further, in view

of recent advances in group-contribution EoSs, this is a perfect opportunity to test benchmark newer models and popular models. As such, this thesis focuses on describing some chosen predictive tools which are later compared in a systematic assessment in chapter 5. These models include: the data-driven model of Borhani et al. (2019); new QSPR models developed with an extension of the experimental database from Borhani et al. (2019), which are developed in Chapter 3; the non-random two liquid model (NRTL), UNIFAC, and modUNIFAC (Do) activity coefficient models; the SAFT-γ Mie equation of state (Papaioannou et al., 2014; Dufal et al., 2014) and the hybrid COSMO-SAC (Wang, Sandler, and Chen, 2007).

2.2 Experimental database of free energy of solvation

This section focuses on introducing the experimental database of the free energy of solvation. In Chapter 1, we provided a definition for the free energy of solvation according to Cramer (2004) using the notation $\Delta G^{solv}_{i,j}$. The definitions have changed in nuance throughout the years as the industry standards have evolved over time. The conventional definitions for the free energy of solvation are categorised as the standard free energies of solution. It is important to discuss these definitions with any relevant equations, discussing the origins of the database, and the classification of solute and solvent molecules.

2.2.1 Standard states and scales

Conventional definitions for the standard free energies of solution

A commonly used definition for the standard free energy of solution, dubbed the standard transfer free energy, $\Delta G^{\circ,\infty,m}_{t,i}$, is the transfer of a solute i in the ideal gas standard state (hypothetical 1 atm (101325 Pa), P°) at a given temperature T (in K) to the hypothetical 1 mol/kg standard state at the solution pressure P and the same temperature (Ben-Naim, 2006). The standard free energy of transfer is obtained at the infinite dilution limit where the liquid phase is essentially pure solvent j. The thermodynamic quantity is expressed as:

$$\Delta G^{\circ,m,\infty}_{t,i} = RT\left[\ln\frac{P}{P^\circ} + \ln\hat{\varphi}^{\infty}_{i,j} + \ln M^\circ_j m^\circ\right] \tag{2.1}$$

where the superscripts $\circ$, ∞, and m denote the use of a standard state, the infinite dilution limit and the molality scale, respectively. The terms $\hat{\varphi}^{\infty}_{i,j}$, M°_j, R and m° are the infinite

dilution mixture fugacity coefficient of a solute i in solvent j (with a 1 mol/kg solution reference) which is dimensionless, the molar mass of a solvent j in kg/mol, the universal gas constant with units of kcal mol^{-1} K^{-1} and a standard molality of 1 mol/kg, respectively. The fugacity coefficient can be evaluated as $\gamma_{i,j}^{\infty} P^{\circ}/P$ where $\gamma_{i,j}^{\infty}$ is the infinite dilution activity coefficient of a solute i in a solvent j assuming the vapour phase is ideal.

There exists another definition for the standard free energy of solution, dubbed the standard free energy of solvation (the preferred definition in this study), which involves the transfer of a solute i in the ideal gas standard state (hypothetical 1 mol/L) at a given temperature T to the hypothetical 1 mol/kg standard state at the solution pressure P and the same temperature. The standard free energy of solvation is expressed as:

$$\Delta G_{s,i}^{\circ,m,\infty} = RT \left[\ln \frac{P}{P^{\circ}} + \ln \hat{\varphi}_{i,j}^{\infty} + \ln M_j^{\circ} m^{\circ} - \ln \frac{\overline{R} T c^{\circ}}{P^{\circ}} \right] \tag{2.2}$$

where $\overline{R}$ is the universal gas constant is the universal gas constant in J mol^{-1} K^{-1}, c° is 1000 mol/m^{-3}, P° is 1 atm (101325 Pa) and the dimensionless term $\ln \frac{RTc^{\circ}}{P^{\circ}}$ is the conversion of a hypothetical 1 atm standard state to a hypothetical 1 mol/L standard state (Guthrie and Povar, 2009; Ho, Klamt, and Coote, 2010; Struebing, 2011).

All of the free energy of solvation data reported in this work uses the $\Delta G_{s,i}^{o,m,\infty}$ definition seen in equation (2.2). However, some experimental sources of data or other predictive tools use the definition seen in equation (2.1). Thus, this section has outlined a path to convert between these two definitions. This is important as the difference in values can result in significant offsets in predicted values.

2.2.2 Database of experimental data

Following from the previous section, the experimental data considered in this study is binary standard state Gibbs free energies of solvation, at 298 K and 1 atm. The corresponding standard states and scales are an ideal gas at 1 mol/L for the gaseous phase and an ideal solution at 1 mol/L with an infinitely dilute reference state for the solution phase. Therefore, all forms of the free energy of solvation is converted into the same scale as the $\Delta G_{s,i}^{o,m,\infty}$ found in equation (2.2). In terms of experimental values of $\Delta G_{s,i}^{o,m,\infty}$, the Minnesota Solvation database Marenich et al., 2012 is used. It provides 2353 solute/solvent pairs. In addition, the Compsol database Moine et al., 2017 provides an extra 11 solute/solvent pairs at the same

temperature and pressure of 298 K and 1 bar. From this point forward, the free energy of solvation will be referred to using $\Delta G^{o,m}_{s,i,j}$. The m superscript refers to the molality scale, whereas the s subscript refers to the ideal solution state of 1 mol/L in the gas phase. Further, the i and j subscripts are used to indicate a solute i and a solvent j. Since all of the free energy of solvation data considered in this thesis is at the infinite dilution, it is omitted from notation.

In order to develop any form of data-driven model, the experimental free energy of solvation values, $\Delta G^{o,m,exp}_{s,i,j}$ values must be used in tandem with a set of descriptor variables. Borhani et al. (2019) compiled 12 bulk descriptors related to solvent and nine quantum mechanical descriptors related to the solute for the 1777 experimental data points in their work. These descriptors were then matched to the newly compiled database of 2364 experimental data points. The 12 bulk solvent descriptors include the boiling temperature, molecular weight, relative permittivity, surface tension, refractive index, enthalpy of vaporization, liquid molar volume, octanol/water partition coefficient, critical temperature, critical pressure, critical volume and dipole moment of the solvent. The nine quantum mechanical solute descriptors include the dipole moment, electronic basicity, electronic acidity, molecular van der Waals volume, HOMO energy, LUMO energy, electronic energy, isotropic polarisability, ideal gas entropy and the dipole moment of the solute. The corresponding symbols and units for these solute and solvent descriptors can be found in table 2.1. The full database can be found in tables B.1 to B.79 in appendix B. Any experimental $\Delta G^{o,m,exp}_{s,i,j}$ data found in the aforementioned tables has been derived from the Minnesota Solvation database and the Compsol database.

As mentioned in section 1.1, the free energy of solvation cannot be determined directly, and thus has to be derived from other thermodynamic quantities. Experimental partition coefficients, Henry's constants or vapour pressures have been compiled over a long period of time and were then converted into free energies of solvation in the MNSol database. The MNSol database documentation (Marenich et al., 2012) details the various thermodynamic paths for converting these quantities into solvation free energies.

The $\Delta G^{o,m}_{s,i,j}$ part of the experimental database is used for the comparison and development of predictive tools, whereas the descriptor part of the database is used specifically for the development of the data-driven models. These are included in this section to allow any other potential users to use the same data set to develop models.

TABLE 2.1: Solute and solvent properties used in Borhani's Partial Least Squares model (Borhani et al., 2019). "Type" indicates whether the descriptor belongs to the solute, solvent or the overall system.

Description	**Type**	**Property**	**Units**
Boiling Point at 1 atm	Solvent	T_b	K
Molecular Weight	Solvent	M_w	g mol^{-1}
Relative Permittvity at 298 K and 1 atm	Solvent	ϵ	-
Surface Tension at 298 K	Solvent	σ	mN m^{-1}
Refractive Index at 298 K and 1 atm	Solvent	n_D	-
Enthalpy of Vaporization at normal boiling point	Solvent	ΔH_v	kJ mol^{-1}
Liquid Molar Volume at 298 K and 1 atm	Solvent	V_m	m^3 kmol^{-1}
Partition Coefficient in Octanol/Water at 298 K, 1 atm and infinite dilution	Solvent	logK_{ow}	-
Critical Temperature	Solvent	T_c	K
Critical Pressure	Solvent	P_c	MPa
Critical Volume	Solvent	V_c	m^3 kmol^{-1}
Dipole Moment at 298 K and 1 atm	Solvent	μ^j	Debye
Electronic basicity[a]	Solute	q^-	-
Electronic acidity[a]	Solute	q^+	-
Molecular van der Waals volume	Solute	V_{mc}	Å^3
Energy of the HOMO	Solute	ε_H	a.u.
Energy of the LUMO	Solute	ε_L	a.u.
Electronic energy	Solute	E	a.u.
Isotropic Polarizability	Solute	π	Bohr^3
Ideal gas entropy at 298 K	Solute	S	cal mol^{-1} K^{-1}
Dipole moment	Solute	μ^i	Debye
Free Energy of Solvation at 298 K, 1 atm pressure with a reference states of 1 mol/L in both the gas and solution phase and the molality concentration scale	System	$\Delta G^{o,m}_{s,i,j}$	kcal mol^{-1}

[a] As calculated using the approach of Famini and Wilson (1993)

2.2.3 Experimental uncertainty of $\Delta G^{o,m}_{s,i,j}$ database

The validity of this comparative study relies on the experimental data that will be compared against the considered predictive tools. If the experimental data has a large error, there is a broader scope for error in the predicted values from the models. Thus, it is vital to assess the experimental uncertainty of the database. The MNSol database which makes up the majority of the database used in this work has documentation (Marenich et al., 2012) that reports an experimental uncertainty of 0.2 kcal mol^{-1}. It is important to note that the MNSol database has been through revisions and upgrades since 2003. In one of the earlier works from the Minnesota group, Thompson et al. (2004) estimate the typical uncertainty for the

free energy of solvation of neutral solutes to be 0.2 kcal mol^{-1}. This shows that on average, the experimental uncertainty has not changed over time according to the Minnesota group. However, it must be noted that the collection of data used in this study and the MNSol database is inherited through numerous works over the course of 90 years.

Extensive compilation works such as the work of Abraham et al. (1990) contain hundreds of solvation free energies for water and hexadecane which account for about 30% of the database. In their work, a range of partition coefficients for water and hexadecane are collected from other sources or measured experimentally. The air/solvent partition coefficients can be converted into free energies of solvation by using the following expression:

$$\Delta G^{\circ,m,\infty}_{B/air} = -2.303RT \log_{10}(P_{B/air}), \tag{2.3}$$

where the subscripts B/air refers to the air/water partition and $P_{B/air}$ is the air/solvent partition coefficient. Free energies of transfer between water and hexadecane can be obtained by subtracting the free energies of solvation of water and hexadecane solvent using hexadecane/water partition coeffecients. This is written as:

$$\Delta G^{\circ,m,\infty}_{water/hexadecane} = \Delta G^{\circ,m,\infty}_{water/air} - \Delta G^{\circ,m,\infty}_{hexadecane/air} \tag{2.4}$$

Abraham et al. noted that the expected experimental errors for the hexadecane partition coefficients are about 0.03 log units whereas the water partition coefficients have more substantial errors. Abraham et al. cites two sources (Hine and Mookerjee, 1975; Mackay and Shiu, 1981) that for the halogenated alkanes, the $\log P_{aq/air}$ (the logarithm of the water partition coefficients) values differ by about 0.1 log unit, and for hexachloroethane, the recorded values differ by 1 log unit. They further note that the error in their transfer energies is estimated to be 0.2 kcal mol^{-1}. The error for the hexadecane solvation free energies according to their estimated error can be obtained by substituting the 0.03 log units into equation (2.3) which yields 0.04 kcal mol^{-1}. The same process applies when considering the error for water solvation free energies. The range of errors for the halogenated alkanes is 0.136 to 1.36 kcal mol^{-1} for hexachloroethane. Given the sum of the transfer energy errors is 0.2 kcal mol^{-1} where the error for the hexadecane solvation free energies is very small, the average estimated error for the water solvation free energies is roughly 0.2 kcal mol^{-1}. While this can be interpreted as a form of confirmation for the current database, it is an uncertainty that is derived from a

lack of information.

Another example where an estimated 0.2 kcal mol^{-1} is carried forward is in the work of Nicholls et al. (2008), where they note that a "reasonable" estimate for the experimental uncertainty is roughly 0.2 kcal mol^{-1}. They also note that several studies have compared measurements for free energies of hydration and found variations for same compounds in the range of 0.01-0.1 kcal mol^{-1}. Further, they note that estimates for the uncertainty have been suggested to be around 0.2 kcal/mol but sometimes larger. This criticism is not directed at any specific author but is meant to elucidate how difficult to estimate experimental uncertainties as they are regularly small or very difficult to extract.

In the work of Buttery et al. (1969), the air water partition coefficients of a selection of ketones, aldehydes, ketones and esters were measured and accompanied with a series of experimental errors. For example, the air/water partition coefficent of acetone is 1.6 $\times$ 10^{-3} with an error of 0.2 $\times$ 10^{-3}. The free energy of solvation can be calculated using equation (2.3) using the partition coefficient, where as the error was estimated using the following expression:

$$\sigma_{\Delta G^{\circ,m,\infty}_{aq/air}}(kcalmol^{-1}) = RT\frac{\sigma_{P_{aq/air}}}{P_{aq/air}} \tag{2.5}$$

After substituting the values into the equations, the free energy of solvation for acetone in water is -3.81 $\pm$ 0.07 kcal mol^{-1}. The listed value in the database is -3.85 kcal mol^{-1} and is within experimental error. The same process was repeated for 23 more partition coefficients and an average experimental error of 0.08 kcal mol^{-1} was found with a range of 0.007 to 0.507 kcal mol^{-1}. Therefore, the same observations from Abraham et al. (1990) and Nicholls et al. (2008) can be seen here with very small errors with some larger errors.

In some works, there are cases where uncertainties or errors in measurement were not reported or were given as average estimated uncertainties. Unfortunately, due to time constraints, an exhaustive estimation of the experimental uncertainty is not possible. It must be noted that even if an estimation was carried out, there would be gaps in the study, and the estimation would have to be treated as a blanket value. Therefore, until a more accurate estimation of the experimental error is carried out, a value of 0.2 kcal mol^{-1} will be accepted as the experimental error.

2.2.4 Classification of solute and solvent molecules

The classification of molecules is a key endeavour as one needs to select a scheme that will allow enough generalisation of the molecules without muddying the physics of the problem. For example, there are molecules which belong to individual families such as alkanes, esters, or alcohols. In this context, multifunctional species can be especially difficult to allocate to a single class. Reichardt and Welton (2014) proposed five schemes to classify molecules, but because of broad definitions, there is some overlap between them. These schemes include i) classification according to the chemical constitution in which molecules are classified according to their chemical bonds such as covalent bonds, ionic bonds or metallic bonds; ii) classification using physical constants, which characterises the properties of a solvent, such as refractive index, dielectric constant, or surface tension; iii) classification based on acid-base behaviour following the Brønstead-Lowry, or Lewis theories that categorise on how protic a molecule is on a relative acidity/basicity scale iv) classification in terms of specific solute/solvent interactions which divides molecules according to their specific interactions with cations and anions often characterised by relative permittivities and dipole moments; and v) classification using multivariate statistical methods which classify molecules according to their molecular properties. The final scheme specifically uses multivariate statistical methods to classify molecules rather than using physical knowledge to attribute them beforehand.

The preferred scheme in this thesis is one that is intuitive or familiar to a reader. While the classification of molecules according to their chemical bonds is useful, it does not offer much when discerning the effects between a solute and a solvent as the bonds are internal to the molecule. The classification scheme according to physical constants, acid-base behaviour, and specific interactions with cations and anions is difficult to use as it is less intuitive. The classification scheme using multivariate statistical methods also requires a large set of data to ensure a fair comparison. Further, the classification is subject to change depending on the molecules used. Therefore, this thesis follows an approach similar to the classification of molecules in terms of specific solute/solvent interactions from Moine et al. (2017). The molecules are sorted by the type of interactions they may form. An example of the assignation of the types considered in the classification for these molecules is given in table 2.2. Self-associating (SA) molecules such as water contain labile hydrogen atoms and a lone pair of electrons that allow for self-association. Non-associating (NA) molecules lack either labile

hydrogen atoms or lone pairs of electrons and the lone pair of electron (E) class applies to molecules with lone pairs of electrons only. This classification scheme enables a user to assess how a model is performing concerning a particular type of molecule and also when considering solute-solvent pairs; one can qualitatively and quantitatively assess the type of interaction formed between these types of molecules (e.g. acetone in water results in hydrogen-bonding, or n-hexane in water results in only dispersive interactions). This classifcation scheme is both intuitive and is also particularly well-suited to SAFT-γ Mie as the molecules have explicit descriptions for being able to interact. The molecules are also classified into individual chemical families such as alkanes, esters or alcohols. A list of the molecules used in this study with their corresponding interaction types and classes can be found in tables A.1 and A.2.

TABLE 2.2: Classification scheme based on the presence of labile hydrogen atoms or lone pairs of electrons for molecules.

Type of molecule	**Label**	**Example**
Self-associating	SA	Water
Non-associating	NA	n-Hexane
Has at least one lone pair of electrons	E	Acetone

2.3 Selected predictive tools for the prediction of $\Delta G^{o,m}_{s,i,j}$

In this section, a brief description of the models used in the systematic assessment in Chapter 5 is presented. First, selected data-driven empirical models (PLS, QPLS, ALAMO) will be introduced, followed by semi-empirical models (NRTL, UNIFAC, modUNIFAC (Do), SAFT-γ Mie) and an ab initio model, COSMO-SAC. Data-driven models are linear or non-linear methods that regress an experimental response variable (in this case, $\Delta G^{o,m,exp}_{s,i,j}$) to an experimental set of descriptor variables (here, the 12 bulk solvent and nine QM solute descriptors). Semi-empirical models have model parameters that are either molecule or functional group-specific which are parameterised from experimental data, and ab initio models are models that are based on quantum-mechanical theory.

2.3.1 Data-driven empirical models

Current state of empirical data-driven models for the prediction of solvation free energies

There have been several attempts at developing data-driven solvation models for the Gibbs free energy of solvation. Data-driven models are preferred as they benefit from their ease of use and with the correct set of descriptors, any solute/solvent system can be modelled. Any solute/solvent system may include unstable compounds such as transition states and mixed solvents; however, these examples require QM descriptors or experimental descriptors that pertain to these compounds. In the case of data-driven solvation models, the performance of such models depends on the quantity, quality and variety of data used to train and validate the models. Further, the robustness of the methodology to fit the descriptors to the desired variable also needs to be considered. Generally, studies aiming to obtain such models focused on a single solute in a range of solvents or a range of solutes in a single solvent. Further, these models can be classified into two main types: experiment-based and theory-based. Experiment-based models utilise experimentally derived chemical descriptors such as solvatochromic and Hildebrand parameters (Abraham et al., 1987a; Abraham et al., 1987b). Conversely, theory-based models utilise molecular descriptors such as topological indices, quantum-mechanical (QM) or thermodynamic descriptors (Borhani, Bagheri, and Manan, 2013; Borhani, Afzali, and Bagheri, 2016). These QM-derived descriptors have been highlighted in several chemometric studies (Cramer, Famini, and Lowrey, 1993; Lowrey et al., 1995; Karelson, Lobanov, and Katritzky, 1996; Katritzky et al., 2010).

Data-driven models utilise statistical and regression methods to relate the Gibbs free energy of solvation to a choice of experimental or theory-based descriptors. For example, Michielan et al. (2008) carried out predictions of the aqueous free energy of solvation of 271 organic molecules using a model that combined 12 autocorrelation molecular electrostatic potential (auto MEP) descriptors with response surface analysis (RSA). They divided their data set into a training set of 248 training points and a testing set of 23 data points. Their model had a coefficient of determination (R^2) of 0.990 and a root mean square error (RMSE) of 0.069 kcal mol^{-1} for the training set and an R^2 of 0.92 and an RMSE of 0.084 kcal mol^{-1} for the testing set. In another example, Delgado and Jaña (2009) showed predictions for the free energy of solvation of 147 components in 1-octanol using multiple linear regression (MLR)

model based on three descriptors. They achieved an R^2 of 0.930 and a standard deviation of 0.570 kcal mol^{-1}.

Katritzky et al. (2003; 2003) proposed an MLR approach in a two-part study which focused on developing two sets of models; one for a series of solutes in a single solvent and the other, a single solute in a series of solvents. The first part of the study was to model the Ostwald solubility coefficient for a series of solutes in a single solvent. This approach was applied to 69 solvents which resulted in 69 solvent-specific MLR models where the number of solutes included in the regression for any given solvent varied from 14 to 226. The 69 models only used solute descriptors as the solvent was constant in each model. The resulting R^2 of the models ranged from 0.837 to 0.998, with a standard deviation of 0.060 to 0.8 kcal mol^{-1}. The second study also focused on building models for predicting solubilities but with a single solute in a series of solvents. In this case, 80 MLR models were derived from solubility data of 80 solutes for which data were available across a range of 15 or more solvents. Conversely to the first study, the 80 models were regressed to solvent descriptors only as the solute was constant in each model. These models achieved excellent performance with R^2 values of 0.604 to 0.996 and a standard deviation of 0.020 to 0.610 kcal mol^{-1}. From these examples, it has been shown that data-driven approaches perform well over a broad range of solutes in solvents or a single solute in a series of solvents.

In a recent study by Borhani et al. (2019) two data-driven models were proposed, an MLR and a partial least squares (PLS) model, which uses both experimentally derived and QM descriptors to model the Gibbs free energy of solvation for a variety of solutes in a range of solvents. A comprehensive data set of 1777 Gibbs free energies of solvation with a corresponding set of 12 bulk solvent descriptors and 9 QM solute descriptors, as shown in table 2.1, were compiled for the training and testing of the MLR and PLS models. The data set included various molecular classes and the resulting models excluded any solute/solvent systems with water as a solvent, except for the self-solvation of water. This exclusion is because water has a vastly different behaviour than other molecules, and thus, the model is designed to handle nonaqueous solvents only. The MLR model included three QM solute descriptors, the polarizability, the energy of the lowest unoccupied molecular orbital (LUMO), and the electrostatic acidity, and two bulk solvent descriptors, namely the heat of vaporisation and the octanol-water partition coefficient. In contrast, the PLS model utilised all 21 descriptor variables. Comparison against experimental Gibbs free energies of solvation yielded an R^2 of

0.88 and an RMSE of 0.59 kcal mol^{-1} for the MLR model. The PLS model that included six latent variables yielded an R^2 of 0.91 and an RMSE of 0.52 kcal mol^{-1}.

Empirical data-driven methodologies

The data-driven empirical methodologies presented in this subsection all serve as a basis for the development of new solvation models found in chapter 3. While these models are discussed briefly in this section, there will be a further elaboration on the details of the methodologies and the new solvation models in chapter 3. The data-driven models found in current work are developed using the experimental $\Delta G_{s,i,j}^{o,m,exp}$ data from the database found in tables B.1 to B.79. Previously in section 2.2.1, it was shown that the solvation free energy can be reported in different standard states and scales. However, since the data-driven will be developed using a database of experimental data in the desired form of $\Delta G_{s,i,j}^{o,m}$, the predicted values from the data-driven models do not need to be converted into another standard state or scale.

Partial Least Squares

In partial least squares or projection to latent structures, a set of descriptors, $\mathbf{X}$, is related to a response variable, $\mathbf{Y}$, through their latent spaces. The latent space is a set of orthogonal projected axes that are a function of the solute and solvent descriptors. The axes are projected such that the covariance between the $\mathbf{X}$ and $\mathbf{Y}$ variables are maximised. In this methodology, a linear expression is extracted between the $\mathbf{X}$ and $\mathbf{Y}$ projected subspaces such that the covariance is maximised (Wold, 1973). In chapter 3, we further elaborate on the specifics of the PLS methodology as it will be used to develop a new set of predictive solvation models. The new PLS models are regressed to the database of experimental free energies of solvation and descriptor variables mentioned earlier. It follows the methodology of Borhani et al. (2019) by using the bulk solvent and QM solute descriptors as variables. Ultimately, the Gibbs free energy of solvation is expressed as a function of the solute and solvent descriptors, $\mathbf{X}$.

Quadratic Partial Least Squares

The quadratic PLS (QPLS) model is a modification to the linear PLS model made by Wold et al. (1989) to account for any nonlinearities between the response variable $\mathbf{Y}$ and descriptor variables $\mathbf{X}$. Similarly to the linear PLS model, the response and descriptor variables are projected on to new axes. However, the main difference lies in the model used to maximise

the covariance between the projected axes. According to Wold et al. (1989), potentially any function is viable; however, the underlying assumption for using a nonlinear model was the idea that the response and descriptor variables are nonlinearly related. In this work, we use an algorithm developed by Garcia-Mũnoz et al. which is presented in chapter 3.

Automatic Learning of Algebraic Models for Optimisation

The automatic learning of algebraic models for optimisation (ALAMO) is a computational methodology developed by the Sahinidis group (Cozad, Sahinidis, and Miller, 2014; Cozad, Sahinidis, and Miller, 2015; Wilson and Sahinidis, 2017) meant for the modelling of complex black-box simulations by fitting low complexity surrogate models and sampling data points at which the surrogate models break down. It is an iterative process which attempts to find a surrogate model that fits over an unknown functional over its applicable domain. The low complexity surrogate models are built from a set of potential basis functions such as bilinear, logarithmic, linear, or polynomials by adding them linearly together. The surrogate model generation is supported by optimisation and statistical methods to find the best subset of basis functions that accurately model the black box without overfitting. However, when applied to the development of a new solvation model, there is no need for the sampling of data points at which the model fails because all of the experimental data have already been specified. Therefore, in chapter 3, the development of new ALAMO solvation models, will only use the surrogate model generation aspect of ALAMO.

2.3.2 Activity coefficient models

The activity coefficient models found in this subsection all calculate the activity coefficient in the mole fraction scale, γ. Further, the activity coefficient models employ the 1 atm reference state for the gas phase. In this current work, the desired form of the solvation free energy utilises the 1 mol/L reference state for the gas phase and the molality scale. Thus, the activity coefficients outputted by the models are in the right reference to be used in equation (2.2). The activity coefficient is defined as a ratio of the mixture fugacity coefficient and the pure solute fugacity coffecient, expressed as such:

$$\gamma_{i,j}^{\infty} = \frac{\hat{\varphi}_{i,j}^{\infty}}{\varphi_i^{\circ}} \tag{2.6}$$

where $\gamma_{i,j}^{\infty}$ is the infinite dilution activity coefficient of a solute i in a solvent j and φ_i° is the fugacity coefficient of pure solute i. However, a limitation of activity coefficient models is that they are restricted to mixtures in the liquid phase and cannot be used to calculate fugacity coefficients of pure compounds. Thus, with regards to activity coefficient models, the current study uses the same strategy as the one seen in Fingerhut et al. (2017) by assuming the vapour phase is an ideal gas. Thus, φ_i° is set to 1. Therefore, the desired form of the free energy of solvation (equation (2.2)) can be calculated.

Non-random two liquid model

Local composition models are based on the concept of local composition which assumes that the composition on a molecular level differs from the bulk composition. Other molecules preferentially surround molecules depending on their size, shape or interaction energies such that the energy level of the system is minimised. The non-random two liquid (NRTL) model is part of the family of local composition models and developed for describing liquid-liquid equilibria (Renon and Prausnitz, 1968). In this thesis, only binary systems are considered; therefore, the activity coefficient of a given species i in a binary system is defined by equation 2.7.

$$\ln\gamma_i = x_j^2 \left[\tau_{ji} \left(\frac{G_{ji}}{x_i + x_j G_{ji}} \right)^2 + \frac{\tau_{ij} G_{ij}}{(x_j + x_i G_{ij})^2} \right] \tag{2.7}$$

where i is the solute, j is the solvent and τ_{ij} and G_{ij} are defined below:

$$G_{ji} = \exp\left(-\alpha_{ji}\tau_{ji}\right) \tag{2.8}$$

$$\tau_{ii} = \tau_{jj} = 0 \quad \text{and} \quad G_{ii} = G_{jj} = 1 \tag{2.9}$$

$$\tau_{ji} = \frac{g_{ji} - g_{ij}}{RT} \tag{2.10}$$

The binary system NRTL model utilises three parameters. The first two are g_{ji} and g_{ij}, which are energy interaction parameters between the solute i and solvent j. These parameters are obtained through the parameterisation of the model by using experimental binary system data. The other parameter is α_{ij}, which is defined as a non-randomness parameter that

describes the order of the molecules surrounding a given molecule on a scale of 0 to 1. An α_{ij} value of 0 describes a completely random phase and a value of 1 is considered an ordered phase. As an example, in a liquid that contains two species that do not interact with one another, the molecules surrounding a given molecule will be essentially random, therefore having no interaction energy and an α_{ij} value of 0. In contrast, if the molecules do form bonds, the non-randomness parameter has a value above zero and the interaction parameters have a non. Given the formulation of semi-empirical models, these parameters are regressed against experimental data which can be found in databases such as the Dortmund Data Bank (DDBST GmbH, 2018). The parameters are summarised in table 2.3.

TABLE 2.3: Definitions of parameters used in the NRTL model.

Parameter	**Units**	**Description**
g_{ij}	-	Energy parameter characteristic of the $i-j$ interaction
g_{ji}	-	Energy parameter characteristic of the $j-i$ interaction
α_{ij}	-	Measure of the nonrandomness of the mixture (where a value of 0 indicates a completely random mixture)

Universal quasichemical functional activity coefficient model

Another member of the family of local composition models is the UNIFAC subfamily of activity coefficient models. The original universal quasichemical functional-group activity coefficient model (UNIFAC) was developed by Fredenslund et al. (1977) as a combination of the Wilson solution-of-groups concepts and the universal quasichemical (UNIQUAC) model (Muzenda, 2013). In comparison to the earlier entries in the local composition series of models, UNIFAC is less reliant on experimental data, and it can be applied to a broader range of molecules due to the use of chemical functional groups. In the UNIFAC mode, the excess Gibbs energy is partitioned into a combinatorial term which describes the effects of molecular size and shape, $\ln \gamma^C$, and a residual term which describes the effects due to group-group interactions, $\ln \gamma^R$ (Fredenslund et al., 1977). The combinatorial term is the Staverman-Guggenheim term (Staverman, 1950; Guggenheim, 1952). Therefore, the activity coefficient for a solute i in a solution of solvent j is expressed as:

$$\ln \gamma_{i,j} = \ln \gamma^R_{i,j} + \ln \gamma^C_{i,j} \tag{2.11}$$

Fredenslund et al. (1977) noted some limitations to the application of the UNIFAC family of models: (i) they are restricted to mixtures in the liquid phase; (ii) they can only model low-pressure systems unless it is paired with an equation of state (e.g. Predictive Soave-Redlich Kwong/Predictive Peng-Robinson). Furthermore, since it is based on the solution-of-groups concept, it does not take into account the proximity of other groups and is therefore unable to distinguish isomers.

Modified UNIFAC Dortmund version (modUNIFAC (Do)

The success of the original UNIFAC model spawned newer versions of the UNIFAC model that modify the parameters, or the formulation has been proposed. One such modification was the modified UNIFAC (Do) model (Weidlich and Gmehling, 1987; Gmehling, Li, and Schiller, 1993; Gmehling et al., 1998; Gmehling et al., 2002; Jakob et al., 2006; Wittig, Lohmann, and Gmehling, 2003) which had a number of changes to the original model such as: (i) the introduction of temperature-dependent group interaction parameters; (ii) expanding the set of groups to cyclic alkanes and reclassifying alcohols into primary, secondary, and tertiary groups with their own set of parameters; and (iii) the fitting of group interaction parameters to infinite dilution activity coefficient data, vapour-liquid equilibria and excess enthalpies to allow for better handling of the infinite dilution limit. Furthermore, the combinatorial term was modified to better model asymmetric mixtures (Muzenda, 2013). As seen in section 2.1, the modUNIFAC (Do) model has markedly excellent performance for the featured comparative studies and model is highly applicable to a broad range of systems due to the large set of functional groups. The corresponding parameters and can be found in the set of modUNIFAC (Do) papers (Weidlich and Gmehling, 1987; Gmehling, Li, and Schiller, 1993; Gmehling et al., 1998; Gmehling et al., 2002; Jakob et al., 2006; Wittig, Lohmann, and Gmehling, 2003).

2.3.3 SAFT-γ Mie

In the SAFT-γ Mie approach, (Papaioannou et al., 2014; Dufal et al., 2014) molecules are represented as heteronuclear chains formed from fused spherical segments which correspond to different chemical functional groups. The model uses the Mie intermolecular potential (Mie, 1903) and a high-temperature perturbation expansion to third order (Lafitte et al., 2013) is implemented to provide a high level accuracy in the Helmholtz free energy and its derivatives.

In the SAFT-γ Mie approach, molecules are modelled as heteronuclear, comprising spherical groups, each contributing S_k to the free energy, and are characterised by the diameter σ_k and dispersion energy ε_k. The SAFT-γ Mie model treats hydrogen bonding and strong polar interactions explicitly via association sites that exist as "H" or "e" sites depending on the functional group but does not directly account for long-range electrostatics. A more in-depth explanation and the relevant parameters can be found in the this set of papers. (Papaioannou et al., 2014; Dufal et al., 2014; Burger et al., 2015; Papaioannou et al., 2016; Sadeqzadeh et al., 2016; Hutacharoen, 2017; Haslam et al., 2020)

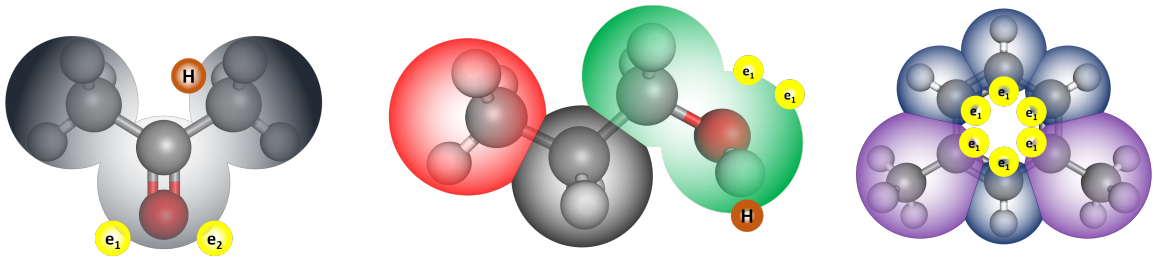

FIGURE 2.1: Examples of the SAFT-γ decomposition of molecules into functional groups (from left to right): acetone is a single group molecule comprising three fused spherical segments (grey) with three association sites; (brown and yellow) 1-propanol is made up of three functional groups, one CH_3 (red), one CH_2 (black) and one CH_2OH group comprising two fused spherical segments (green) with three association sites (brown and yellow); 1,3-dimethylbenzene is made up of six functional groups, two $acCH_3$ (purple) groups and four acCH groups (blue).

An advantage of SAFT-γ Mie is its foundation in statistical mechanics and the group contribution aspect that allows for the modelling for a wide range of molecules. However, molecules are represented as a chain of Mie segments rather than having a defined branched or ring structure. This means that isomers cannot be distinguished from one another unless specific functional groups are created.

The SAFT-γ Mie equation of state is written in terms of the Helmholtz free energy as a function of the temperature T, the volume V and the composition vector $\mathbf{N}$ $(N_1, N_2, ...)$. The SAFT-γ Mie equation of state uses the mole fraction concentration scale with a reference state of 1 atm pressure in the gas phase and 1 mol/L in the solvent phase. Therefore, the Gibbs free energy of solvation in the context of SAFT-γ Mie is the Gibbs free energy of transfer, $\Delta G_{t,i}^{o,x}$, in the mole fraction scale. In order to calculate the Gibbs free energy of transfer $\Delta G_{t,i}^{o,x}$, the residual chemical potential μ_i^{res} is given by:

$$\mu_i^{res}(T,P,\mathbf{x}) = \left.\frac{\partial A^{res}(T,V,\mathbf{N})}{\partial N_i}\right|_{T,V,N_{j\neq i}} - RT\ln Z(T,P,\mathbf{x}) \tag{2.12}$$

where R is the universal gas constant, the compressibility factor $Z = p\nu_p/RT$, ν_p is the molar volume for a given pressure, the mole fraction vector $\mathbf{x} = \mathbf{N}/N$ and N is the total number of moles. The infinite dilution fugacity coefficient $\hat{\varphi}_{i,j}^{\infty}$ of a component i in solvent j can be obtained via the infinite dilution residual chemical potential:

$$\ln \hat{\varphi}_{i,j}^{\infty}(T,P) = \frac{\mu_{i,j}^{res,\infty}(T,P)}{RT} \tag{2.13}$$

so that the Gibbs free energy of solvation can be obtained via the residual chemical potential of a solute i in a solvent j at infinite dilution (McCabe, Galindo, and Cummings, 2003). Therefore, the free energy of solvation, $\Delta G_{s,i,j}^{o,m}$, can be obtained.

2.3.4 Conductor-like screening model segmented activity coefficients

In the conductor-like screening model segmented activity coefficients (COSMO-SAC) (Lin and Sandler, 2002; Wang, Sandler, and Chen, 2007; Wang, Song, and Chen, 2011; Hsieh, Lin, and Vrabec, 2014) model, molecules to be a collection of equally sized surface segments, with their interactions determined from the screening changes they acquire on ideal solvation. The model is based on the assumptions that these segments are paired in the solution, and that segment interactions are confined within each pair such that there is no interaction among segments of different pairs (Lin and Sandler, 2002). The general equation for the COSMO-SAC model is shown below:

$$\ln \gamma_{i,j}(T,P,\mathbf{x}) = \ln \gamma_{i,j}^{R}(T,P,\mathbf{x}) + \ln \gamma_{i,j}^{C}(T,P,\mathbf{x}) \tag{2.14}$$

Although the partitioning of the activity coefficient contributions is the same here and in the modUNIFAC (Do) model, the residual contribution of the COSMO-SAC approach considers the permanent electrostatic interactions between molecules in the mixture. These interactions are based on quantum chemistry and COSMO solvation calculations (Klamt and Schüürmann, 1993). These calculations are only carried out once per molecular species and have been stored in databases like VT-database (Mullins et al., 2006; Mullins et al., 2008) to be used later for thermophysical property and phase behaviour calculations (Lin et al., 2004; Wang, Hsieh, and Lin, 2015; Hsieh and Lin, 2012; Lin et al., 2011; Hsieh et al., 2011; Hsieh, Sandler, and Lin, 2010; Hsieh and Lin, 2009; Hsieh and Lin, 2008). A user can also perform

their own calculations to be used in the model; however, a difference in the quality of the calculation or the program used can affect the predicted value of the activity coefficients.

The surface charge distribution of a molecule i obtained from the QC/COSMO calculation is averaged using a semi-theoretical equation (Lin and Sandler, 2002) and then used to generate σ-profiles, $p_i(\sigma_m)$. This can be thought of the probability of finding a surface segment with charge density σ_m on molecule i. The σ-profile is also known as the molecular surface shielding charge density distribution and is unique for every molecule. The combinatorial contribution considers the molecular size and shape effects between molecules in the mixture via the Staverman-Guggenheim (SG) combinatorial term (Staverman, 1950; Guggenheim, 1952).

Several modifications have also been made to the original 2002 COSMO-SAC: (i) Wang et al. (2007) updates their definition of hydrogen bonding; (ii) a more substantive update, COSMO-SAC10 (Wang, Song, and Chen, 2011) sought to treat electrostatics as a temperature-dependent parameter and differentiated hydrogen bonding with respect to hydroxyl groups; (iii) COSMO-SAC-dsp (Hsieh, Lin, and Vrabec, 2014) is the latest version of the COSMO-SAC family where it includes a dispersion term derived from molecular dynamics simulations to account for dispersive forces. The version of COSMO-SAC employed in this study is the 2002 version of COSMO-SAC (Lin and Sandler, 2002). The relevant parameters for each of these properties can be found in their respective references.

The Gibbs free energy of solvation is calculated using the same set of equations as the activity coefficient models because the COSMO-SAC model shares the same reference standard states and concentration scale. Thus, the activity coefficient obtained from COSMO-SAC can be converted into $\Delta G_{s,i}^{o,m,\infty}$ using the same methodology found in section 2.3.2.

2.3.5 Conclusion

In this section, a review of the current state of systematic studies was presented. In the selection of critical comparative studies, the studies of interest only tested models from the same class of tools.. As such, several models have been selected for a systematic assessment that spans the range of predictive tools. We have also discussed the details of the experimental database of free energy of solvation data and the relevant standard states and scales. We have also presented how to convert the selected models into the chosen form of the free energy of solvation, $\Delta G_{s,i}^{o,m,\infty}$. As a reminder, the free energy of solvation will only be referred to with

$\Delta G^{o,m}_{s,i,j}$. In chapter 3, we cover the development of new PLS, QPLS, and ALAMO models that will feature in the systematic assessment in chapter 5.

Chapter 3

The development and testing of data-driven solvation models

3.1 Objective

In section 2.3.1, the state of data-driven models was introduced. There were methods with limited scope that focused on modelling a single solute in many solvents or many solutes in a single solvent. These approaches would result in numerous models for each solute and solvent. Borhani et al. (2019) developed a generalised form for the solvation model that can accept any solute in any solvent. Despite some benefits to this approach, it could potentially be improved by adopting a bigger data set for training and validation and by utilising different mathematical frameworks for the model building. The expanded database has already been introduced in section 2.2.2 and the frameworks are chosen to be the PLS, QPLS and ALAMO models. Therefore, this chapter will focus on developing new data-driven models with the expanded database and new frameworks.

3.2 Data-driven methodologies

Data-driven empirical models relate a set of descriptor variables, $\mathbf{X}$, and a response variable, $\mathbf{Y}$. The mathematical relationship used to relate these variables can be linear or nonlinear where any resulting models are defined as linear or nonlinear. Further, the complexity of the mathematical relationships can extend past a linear or quadratic equation. For example, in regression-focused methodologies such as projection to latent spaces (or partial least squares), the model is inherently linear, making it easy to use; however, the descriptor variables have been projected onto new axes to capture the direction of maximum variance. Thus, these

new axes (or projected variables) can be used more efficiently at describing the response variable, with some loss in precision. In contrast, for complex methodologies such as neural networks, nonlinear mathematical functions are mixed together to create a convoluted model that can describe the intricacies of the response variable. However, this convolution comes at the expense of model complexity. There also exist methodologies such as the automatic learning of algebraic models for optimisation (ALAMO) which employs optimisation to regress parameters of a sum of a best of linear and nonlinear mathematical functions determined by the model. Therefore, in data-driven models, any mathematical framework is a means to describe the response variable, $\mathbf{Y}$ through a set of descriptor variables, $\mathbf{X}$; however, the trade-off between model complexity and precision is a crucial aspect that needs to be considered.

In section 2.2.2, the experimental database of solvation free energies, $\Delta G^{o,m}_{s,i,j}$, was presented. Given a number N of data points, a number M of descriptor variables, $\boldsymbol{x}_m$, and K response variables, $\boldsymbol{y}_k$, this results in two data blocks of sizes $(N \times M)$ for $\mathbf{X}$ and an $(N \times K)$ for $\mathbf{Y}$. The $\mathbf{X}$ acts as the input data set while $\mathbf{Y}$ acts as the output data set. In this section, we elaborate on the partial least squares (PLS), quadratic partial least squares (QPLS) and automatic learning of algebraic models for optimisation (ALAMO) methodologies as mentioned in section 2.3.1.

3.2.1 Linear and Quadratic Partial Least Squares

In the PLS methodology, the $\mathbf{X}$ and $\mathbf{Y}$ data blocks are projected onto new axes (latent subspaces), $\boldsymbol{t}$ and $\boldsymbol{u}$, which are the input and output scores. The scores are then regressed to each other using a linear or nonlinear function, resulting in a model that relates the $\mathbf{X}$ and $\mathbf{Y}$ subspaces. The benefit of following this approach is to minimise any potential noise and maximise the colinearity between the $\mathbf{X}$ and $\mathbf{Y}$ data blocks.

The $\mathbf{X}$ and $\mathbf{Y}$ matrices can be decomposed into several rank-one matrices in the same way as principal component analysis (PCA). The decomposition is defined as the product between the input and output scores vectors $\boldsymbol{t}_a$ and $\boldsymbol{u}_a$ and corresponding input and output loadings vectors, $\boldsymbol{p}_a$ and $\boldsymbol{q}_a$. These scores can be interpreted as the orthogonal distance of the projected data points from their respective projected axes where the loadings vectors describe the projected axes. Each axes is a function of the variables in their respective data blocks and is orthogonal to all other axes. Thus, this allows for the representation of the $\mathbf{X}$ and $\mathbf{Y}$ matrices through the sum of several smaller independent parts:

$$\mathbf{X} = \sum_{a=1}^{L} \boldsymbol{t}_a \boldsymbol{p}_a{}^{\mathbf{T}} + \mathbf{E} \tag{3.1}$$

$$\mathbf{Y} = \sum_{a=1}^{L} \hat{\boldsymbol{u}}_a \boldsymbol{q}_a{}^{\mathbf{T}} + \mathbf{F} \tag{3.2}$$

where L is the chosen number of latent variables, E and F are residual matrices, and $\hat{\boldsymbol{u}}$ is the prediction of the scores $\boldsymbol{u}_a$ through a linear or nonlinear function given by,

$$\hat{\boldsymbol{u}}_a = f(\boldsymbol{t}_a) + \boldsymbol{h}_a, \qquad a = 1, 2, \ldots, L. \tag{3.3}$$

where $f(\boldsymbol{t}_a)$ describes any continuous mathematical function and $\boldsymbol{h}_a$ is a residual vector. In this work, the PLS and QPLS frameworks are considered and therefore, linear and quadratic function are used with the resulting model. These functions are shown below:

$$\hat{\boldsymbol{u}}_a = \boldsymbol{t}_a b_a + \boldsymbol{h}_a, \qquad a = 1, 2, \ldots, L. \tag{3.4}$$

$$\boldsymbol{u}_a = c_{0_a} + c_{1_a} \boldsymbol{t}_a + c_{2_a} \boldsymbol{t}_a \circ \boldsymbol{t}_a + \boldsymbol{h}_a, \qquad a = 1, 2, \ldots, L. \tag{3.5}$$

where $\circ$ is the Hadamard product for element-wise multiplication of two matrices. As seen in equations (3.4) and (3.5), $\boldsymbol{t}$ is a $N \times 1$ vector of the scores, therefore, $\boldsymbol{t} \circ \boldsymbol{t}$ is a $N \times 1$ vector of the squared scores. An example of the Hadamard product is if $\boldsymbol{t}_a = (1, 2, 3)$, then $\boldsymbol{t}_a \circ \boldsymbol{t}_a = (1, 4, 9)$. A user can adapt the methodology and use any function desired. Further, the linear or nonlinear functions that relate the $\mathbf{X}$ and $\mathbf{Y}$ scores are referred to as inner functions and are fitted using an ordinary least-squares approach. Thus, considering the form of the $\mathbf{X}$ and $\mathbf{Y}$ matrices, and their corresponding $\boldsymbol{t_a}$, $\hat{\boldsymbol{u}}_{\boldsymbol{a}}$, $\boldsymbol{p_a}$ and $\boldsymbol{q_a}$ arrays, the resulting PLS and QPLS models can be thought of a sum of independent linear or nonlinear functions which sum together to form $\mathbf{X}$ and $\mathbf{Y}$

The variance in the $\mathbf{X}$ and $\mathbf{Y}$ input and output blocks is usually captured well by the first L latent variables where $L < (M, K)$ while the resulting residual matrices $\mathbf{E}$ and $\mathbf{F}$ typically represent the random noise in the data. The number of latent variables L can be chosen according to different heuristic or statistical tools; here, with the aim to develop robust models for prediction (Baffi, Martin, and Morris, 1999). Therefore, cross-validation will be

used to determine the number of latent variables, L, as it minimises the risk of overfitting the models (Wold, 1973; Geladi and Kowalski, 1986; Wold, Kettaneh-Wold, and Skagerberg, 1989; Baffi, Martin, and Morris, 1999). Cross-validation involves splitting $\mathbf{X}$ and $\mathbf{Y}$ into S equal sized, non-overlapping several subsets and omitting one of the sets such that the PLS models will be trained on $S-1$ subsets of data and tested on the omitted subset. This process is repeated until each subset in S has been omitted once. The predicted residual sum of squares (PRESS) statistic is typically used to quantify the performance of a given model and is given by:

$$\text{PRESS} = \sum_{g=1}^{N} (\boldsymbol{y}_g - \hat{\boldsymbol{y}}_g)^2 \tag{3.6}$$

where $\boldsymbol{y}_g$ and $\hat{\boldsymbol{y}}_g$ are the experimental and predicted data entries in the $\mathbf{Y}$ matrix. However, in this work, the metric used to quantify the performance of the models is the mean squared error (MSE) given by:

$$\text{MSE} = \frac{1}{N} \sum_{g=1}^{N} (y_g - \hat{y}_g)^2 \tag{3.7}$$

The cross-validation MSE (MSE-CV) is the average of the MSEs over each combination of the r training and testing data subsets and is defined below:

$$\text{MSE-CV} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{N_s} \sum_{g=1}^{N_s} (y_g - \hat{y}_g)^2 \tag{3.8}$$

where s is a subset in S, and N_s is the data points belonging to subset s. For the PLS and QPLS models, S is chosen to be 20, which excludes 5% of the data for testing. A PLS model is regressed for each $a = 1, 2 \ldots, L$ and the MSE-CV is calculated at each step. The optimal number of latent variables is then chosen as that giving the set of PLS models with the lowest MSE-CV. This process is applied to both the linear and quadratic version of the PLS methodology.

The general idea behind the PLS and QPLS methodologies has been introduced. It is possible for the reader to refer to the original papers for the algorithms which derive the score and loading vectors; however, the QPLS algorithm used in this work is a modification of the original QPLS algorithm (García-Muñoz, 2020). Therefore, to understand the reason why the modification is used, it is necessary to understand the original PLS and QPLS

methodologies. The PLS methodology utilises the nonlinear iterative partial least squares (NIPALS) algorithm (Wold, 1973) which extracts sequentially the latent variables of the $\mathbf{X}$ and $\mathbf{Y}$ blocks from $a = 1 \rightarrow L$ as a combination of the input and output variables. This process is illustrated in Table 3.1. The algorithm is split into two loops, the outer loop for each pair of latent variables and an inner loop for the calculation of the scores, loadings and regression of the relation between $\boldsymbol{u}$ and $\boldsymbol{t}$. In step 0, the calculation is initialised and a is set as the first latent variable. In steps 1.1 to 1.3, the $\mathbf{X}$ matrix is projected into the scores through the use of the weight vector $\boldsymbol{w}$ which finds the direction of maximum variance in the $\mathbf{X}$ space. For steps 2.1 to 2.3, the direction of maximum variance is found in the $\mathbf{Y}$ space while accounting for the maximum variance in $\mathbf{X}$, resulting in the maximisation of covariance in both spaces. In equation 3.4, it is shown that vector $\hat{\boldsymbol{u}}$ is linearly dependent on $\boldsymbol{t}$ but since the values of constant $\boldsymbol{b}$ are immaterial because of later normalizations, $\boldsymbol{t}$ is regressed directly against $\mathbf{Y}$ (Wold, Kettaneh-Wold, and Skagerberg, 1989). Since the $\mathbf{Y}$ block only contains one response variable, the weight vector $\boldsymbol{c}$ is set to the loading vector $\boldsymbol{q}$ as there is only one direction of maximum variance in $\mathbf{Y}$. In step 3, the convergence is checked by using either the $\boldsymbol{t}$ or $\boldsymbol{u}$ scores, unless the maximum number of iterations is exceeded or the tolerance criterion is met. If convergence is not met, the algorithm returns to step 2.1. When convergence is achieved, a new $\boldsymbol{u}$ vector is calculated. The loading vector $\boldsymbol{p}$ is then calculated in step 5, and in step 6, the $\mathbf{X}$ and $\mathbf{Y}$ are deflated. The deflation process is where the predictions obtained by the product the scores and loadings, $\boldsymbol{t}\boldsymbol{p}^{\mathbf{T}}$ and $\boldsymbol{u}\boldsymbol{q}^{\mathbf{T}}$, are subtracted from the $\mathbf{X}$ and $\mathbf{Y}$ blocks, respectively. The reasoning behind this is to create a set of orthogonal a latent variables that account for a certain amount of variance in the original $\mathbf{X}$ and $\mathbf{Y}$ data blocks. Each of these latent variables is able to predict a response variable to a certain degree, and the sum of these variables represents the overall predictive capability of the PLS model. The $\mathbf{X}$ and $\mathbf{Y}$ are then set to the resulting residual matrices $\mathbf{E}$ and $\mathbf{F}$, respectively, to evaluate the next pair of latent variables. In step 7 of the algorithm, the termination occurs when $a = k$, where k is the original number of descriptors (also the maximum number of latent variables), or $a = L$ if L is determined beforehand. The optimal number of latent variables are determined by using cross-validation and the MSE-CV metric. The MSE-CV is calculated by using the predicted values of the PLS model at each a until $a = k$. The PLS model at each a is the sum of its current a and previous a variables. For example, if $a = 5$, the PLS model would be composed of the $a = 1, 2, 3, 4, 5$ variables. The process of calculating the MSE-CV values for each a

until $a = k$ result in k MSE-CV values. In doing this, the optimal number of latent variables can be chosen as the number with the lowest MSE-CV. Otherwise, if the algorithm does not terminate, the deflated **X** and **Y** data blocks, **E** and **F** are used to calculate the next set of latent variables.

TABLE 3.1: Nonlinear Iterative Partial Least Squares (NIPALS) algorithm

0: *Initialisation* - Use one column in the **X** matrix as a starting vector for $\boldsymbol{t}$
Set $a = 1$; Set $\boldsymbol{t}_{a,old}$=$\mathbf{X_{k=1}}$; Set iter = 0 (iteration counter); Set $\text{iter}^{\max} = 500$
1.1: Calculate the **X** block weight vector $\boldsymbol{w}$
$\boldsymbol{w}_a = \mathbf{X^T}\boldsymbol{t}_a / \boldsymbol{t}_a^{\mathbf{T}}\boldsymbol{t}_a$
1.2: Normalise the **X** block weight vector $\boldsymbol{w}_a$
$norm \quad \boldsymbol{w}_a : ||\boldsymbol{w}_a|| = 1$
$\boldsymbol{w}_a = \boldsymbol{w}_a / (\boldsymbol{w}_a^{\mathbf{T}}\boldsymbol{w}_a)^{0.5}$
1.3: Update the **X** block score vector $\boldsymbol{t}_a$
$\boldsymbol{t}_a = \mathbf{X}\boldsymbol{w}_a$
2.1: Update the **Y** block loading vector $\boldsymbol{q}_a$
$\boldsymbol{q}_a = \mathbf{Y^T}\boldsymbol{t}_a / \boldsymbol{t}_a^{\mathbf{T}}\boldsymbol{t}_a$
2.2: Update the **Y** block score vector $\boldsymbol{u}$
$\boldsymbol{t}_a = \mathbf{Y^T}\boldsymbol{q}_a / \boldsymbol{q}_a^{\mathbf{T}}\boldsymbol{q}_a$
3: Check convergence of the **X** block score vector $\boldsymbol{t}_a$
$\boldsymbol{t}_{a,diff} = \boldsymbol{t}_a - \boldsymbol{t}_{a,old}$
if $\boldsymbol{t}_{a,diff}^{\mathbf{T}} \cdot \boldsymbol{t}_{a,diff} < 10^{-}6$
→ go to step 4
else if iter $<$ $\text{iter}^{\max}$
→ $\boldsymbol{t}_{a,old} = \boldsymbol{t}_a$ and return to step 2.1
4: Recalculate $\boldsymbol{t}_a$ (step 2.3), the slope, b_a, and error, $\boldsymbol{h}_a$, of the inner relation ($\boldsymbol{u}_a = b_a\boldsymbol{t}_a + \boldsymbol{h}_a$)
5: Calculate **X** block loadings vector $\boldsymbol{p}$
$\boldsymbol{p}_a = \mathbf{X^T}\boldsymbol{t}_a / \boldsymbol{t}_a^{\mathbf{T}}\boldsymbol{t}_a$
6: Calculate residual matrices **E** and **F**
$\mathbf{E} = \mathbf{X} - \boldsymbol{t}_a\boldsymbol{p}_a^{\mathbf{T}}$
$\mathbf{F} = \mathbf{Y} - \boldsymbol{u}_a\boldsymbol{q}_a^{\mathbf{T}}$
7: Repeat calculation until $a = k$ where k is the original number of descriptors or
$a = L$ if L is predetermined
if $a = L$
→ Stop
else
$\mathbf{X} = \mathbf{E}$; $\mathbf{Y} = \mathbf{F}$; $a = a + 1$; iter=0
return to step 1.1

The QPLS algorithm (Wold, Kettaneh-Wold, and Skagerberg, 1989; García-Muñoz, 2020) used in this work is detailed in Table 3.2. The initialisation step uses the NIPALS algorithm to calculate a set of starting scores and loadings vectors to be used in the QPLS algorithm. In step 1.1, ordinary least squares regression is used to fit the $\boldsymbol{u}$ and $\boldsymbol{t}$ scores to a quadratic model. In comparison to the NIPALS algorithm, the $\boldsymbol{t}$ scores cannot be used directly and

therefore $\boldsymbol{q}$ is calculated using the predicted scores $\boldsymbol{r}$. The loading vector does not need to be normalised once again as $\mathbf{Y}$ is a $(N \times 1)$ matrix. In step 1.4, the $\mathbf{Y}$ is projected onto the $\boldsymbol{u}$ scores using the loading vector $\boldsymbol{q}$. For step 2, the corrections to the weight vector $\boldsymbol{w}$ are calculated by utilising a Newton-Raphson linearisation as the effect of the nonlinear inner relation needs to be accounted for in the projection of $\mathbf{X}$ onto $\boldsymbol{t}$ (Wold, Kettaneh-Wold, and Skagerberg, 1989; Baffi, Martin, and Morris, 1999). In step 2.1, a matrix $\mathbf{Z}$ is constructed which contains the Newton-Raphson linearisation of the quadratic inner relation $(c_1 + 2c_2\boldsymbol{t})$, an $(N \times 1)$ vector multiplied by each dimension of the $\mathbf{X}$ block. In steps 2.2, 2.3, and 2.4, the process developed by Wold et al. (Wold, Kettaneh-Wold, and Skagerberg, 1989) is followed. In step 2.5, only the first K elements are chosen as the corrections to the weights $\boldsymbol{w}$ where the weights are subsequently updated and normalised in steps 2.5 and 2.6, respectively. The scores $\boldsymbol{t}$ are then calculated with the newly updated weights $\boldsymbol{w}$ and the $\mathbf{X}$ matrix resulting in a projection that accounts for the quadratic inner relation. Salvador García-Muñoz introduced step 3.2, which is not a part of the original QPLS algorithm, to account for the effect of the nonlinear inner relation in the $\mathbf{Y}$ block. Therefore, the predicted scores $\boldsymbol{r}$, loadings $\boldsymbol{q}$, and scores $\boldsymbol{u}$ are recalculated to propagate this effect (García-Muñoz, 2020). The convergence is then evaluated on $\mathbf{X}$ scores vector $\boldsymbol{t}$ with a tighter convergence criterion of 10^{-7} and a larger maximum number iterations, due to the nonlinear nature of the QPLS model. The loadings $\boldsymbol{p}$ are calculated in step 5, and the $\mathbf{X}$ and $\mathbf{Y}$ data blocks are deflated in the same way as the PLS algorithm in step 6. The termination condition of the QPLS algorithm is the same as the PLS algorithm and the optimal number of latent variables are calculated in the same way. and resulting residual matrices are then calculated in steps 5,. The procedure is then repeated if another pair of latent variables is desired.

Therefore, the general idea behind the PLS methodology is made up of two parts, the outer mapping and inner mappings of the model. Table 3.3 illustrates this idea and is adapted from the work of Baffi et al. (1999). The linear inner and outer mapping describe the translation of the $\mathbf{X}$ and $\mathbf{Y}$ data blocks into their respective latent variables and the inner mapping describes the linear or nonlinear regression of the $\mathbf{u}$ and $\mathbf{t}$ scores. The matrix $\mathbf{R}$ is defined as:

$$\mathbf{R} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \tag{3.9}$$

where $\mathbf{W}$ and $\mathbf{P}$ are matrices that contain the weights $\boldsymbol{w}_a$ and $\boldsymbol{p}_a$ for the latent variables

TABLE 3.2: Quadratic PLS algorithm(García-Muñoz, 2020)

0: *Initialisation* - Use one column in the $\mathbf{Y}$ block as the starting vector for $\boldsymbol{u}$ and calculate starting vectors for $\boldsymbol{w}$ and $\boldsymbol{t}$ by a linear PLS calculation

Set $a = 1$; $\boldsymbol{t}_{old} = \boldsymbol{t}$; $\boldsymbol{w}_{old} = \boldsymbol{w}$; $\boldsymbol{c} = 0$; $iter = 0$; $iter^{\max} = 1000$

1.1: Estimation of coefficients $\boldsymbol{c}$ for the $\boldsymbol{u}$-$\boldsymbol{t}$ inner relation by least squares

$\boldsymbol{u} = c_0 + c_1\boldsymbol{t}_{old} + c_2\boldsymbol{t}_{old} \circ \boldsymbol{t}_{old} + \boldsymbol{h}$

1.2: Calculate $\boldsymbol{r}$ predictions of $\boldsymbol{u}$-$\boldsymbol{t}$ inner relation

$\boldsymbol{r} = \hat{\boldsymbol{u}} = c_0 + c_1\boldsymbol{t}_{old} + c_2\boldsymbol{t}_{old} \circ \boldsymbol{t}_{old}$

1.3: Calculate and normalise $\mathbf{Y}$ block weight vector $\boldsymbol{q}$

$\boldsymbol{q} = \mathbf{Y}^{\mathbf{T}}\boldsymbol{r}/\boldsymbol{r}^{\mathbf{T}}\boldsymbol{r}$

$\boldsymbol{q} = \boldsymbol{q}/(\boldsymbol{q}^{\mathbf{T}}\boldsymbol{q})^{0.5}$

1.4: Update the $\mathbf{Y}$ block score vector $\boldsymbol{u}$

$\boldsymbol{u} = \mathbf{Y}^{\mathbf{T}}\boldsymbol{q}/\boldsymbol{q}^{\mathbf{T}}\boldsymbol{q}$

2: Calculate corrections to $\mathbf{X}$ block weight vector $\boldsymbol{w}$

2.1: Construct $\mathbf{Z}$ matrix where Z has a shape of $(N \times K + 3)$

$\mathbf{Z} = [(c_1 + 2c_2\boldsymbol{t}) \circ \mathbf{X}, 1, \boldsymbol{t}_{old}, \boldsymbol{t}_{old} \circ \boldsymbol{t}_{old}]$

2.2: Calculate and normalise column vector $\boldsymbol{v}$

$\boldsymbol{v} = \mathbf{Z}^{\mathbf{T}}\boldsymbol{u}/\boldsymbol{u}^{\mathbf{T}}\boldsymbol{u}$

$\boldsymbol{v} = \boldsymbol{v}/(\boldsymbol{v}^{\mathbf{T}}\boldsymbol{v})^{0.5}$

2.3: Calculate column vector $\boldsymbol{s}$

$\boldsymbol{s} = \mathbf{Z}\boldsymbol{v}/\boldsymbol{v}^{\mathbf{T}}\boldsymbol{v}$

2.4: Calculate column vector $\boldsymbol{b}$

$\boldsymbol{b} = \boldsymbol{s}^{\mathbf{T}}\boldsymbol{u}/\boldsymbol{s}^{\mathbf{T}}\boldsymbol{s}$

2.5: Calculate correction to $\mathbf{X}$ block weight vector $\boldsymbol{w}$, $\mathbf{d}\boldsymbol{w}$

$\mathbf{d}\boldsymbol{w} = \boldsymbol{b}\boldsymbol{v}$ (for the first K elements)

2.6: Update and normalise $\mathbf{X}$ block weight vector $\boldsymbol{w}$

$\boldsymbol{w} = \boldsymbol{w}_{old} + \mathbf{d}\boldsymbol{w}$

$\boldsymbol{w} = \boldsymbol{w}/(\boldsymbol{w}^{\mathbf{T}}\boldsymbol{w})^{0.5}$

3.1: Calculate $\mathbf{X}$ block score matrix $\boldsymbol{t}$

$\boldsymbol{t} = \mathbf{X}^{\mathbf{T}}\boldsymbol{w}/\boldsymbol{w}^{\mathbf{T}}\boldsymbol{w}$

3.2: Recalculate $\mathbf{Y}$ block vectors $\boldsymbol{q}$ and $\boldsymbol{u}$

$\boldsymbol{r} = \hat{\boldsymbol{u}} = c_0 + c_1\boldsymbol{t} + c_2\boldsymbol{t} \circ \boldsymbol{t}$

$\boldsymbol{q} = \mathbf{Y}^{\mathbf{T}}\boldsymbol{r}/\boldsymbol{r}^{\mathbf{T}}\boldsymbol{r}$

$\boldsymbol{u} = \mathbf{Y}^{\mathbf{T}}\boldsymbol{q}/\boldsymbol{q}^{\mathbf{T}}\boldsymbol{q}$

4: Check convergence on $\mathbf{X}$ block scores $\boldsymbol{t}$

if $|(\boldsymbol{t}^{\mathbf{T}}\boldsymbol{t})^{0.5} - (\boldsymbol{t}_{old}{}^{\mathbf{T}}\boldsymbol{t}_{old})^{0.5}| \;/\; (\boldsymbol{t}_{old}{}^{\mathbf{T}}\boldsymbol{t}_{old})^{0.5} < 10^{-7}$

$\rightarrow$ go to step 5

else iter $<$ iter$^{\max}$

$\rightarrow \boldsymbol{t}_{old} = \boldsymbol{t}, \boldsymbol{w}_{old} = \boldsymbol{w}$ and return to step 1.1

5: Calculate $\mathbf{X}$ block loadings vector $\boldsymbol{p}$

$\boldsymbol{p} = \mathbf{X}^{\mathbf{T}}\boldsymbol{t}/\boldsymbol{t}^{\mathbf{T}}\boldsymbol{t}$

6: Calculate residual matrices $\mathbf{E}$ and $\mathbf{F}$

$\mathbf{E} = \mathbf{X} - \boldsymbol{t}\boldsymbol{p}^{\mathbf{T}}; \mathbf{F} = \mathbf{Y} - \boldsymbol{u}\boldsymbol{q}^{\mathbf{T}}$

7: Repeat calculation until $a = k$ where k is the original number of descriptors or $a = L$ if L is predetermined

if $a = L \rightarrow$ Stop

else $\mathbf{X} = \mathbf{E}$; $\mathbf{Y} = \mathbf{F}$; $a = a + 1$; iter=0; return to step 1.1

$a = 1 \rightarrow L$. The matrix $\mathbf{R}$ can be thought of the direct projection of the $\mathbf{X}$ onto $\mathbf{T}$ (the matrix that contains the scores $\boldsymbol{t}_a$ for $a = 1 \rightarrow L$) since the latent variables are calculated sequentially using deflated versions of the $\mathbf{X}$ matrix. Matrices $\hat{\mathbf{U}}$ and $\mathbf{Q}$ are the counterparts to the $\mathbf{T}$ and $\mathbf{P}$ matrices for the $\mathbf{X}$ block whereas the $\mathbf{B}$ matrix is a diagonal matrix that contains the inner regression coefficients b_a, where the off-diagonal elements are equal to zero.

TABLE 3.3: Summary of the general PLS approach

Linear input outer mapping	$\mathbf{X} \rightarrow \mathbf{T}$	$\mathbf{T} = \mathbf{XR}$
	$\mathbf{T} \rightarrow \hat{\mathbf{X}}$	$\hat{\mathbf{X}} = \mathbf{TP^T}$
Inner mapping	$\mathbf{T} \rightarrow \mathbf{U}$	$\hat{\boldsymbol{u}}_k = f_k(\boldsymbol{t}_k)$
Linear output outer mapping	$\hat{\mathbf{U}} \rightarrow \hat{\mathbf{Y}}$	$\hat{\mathbf{Y}} = \hat{\mathbf{U}}\mathbf{Q^T} = \mathbf{TBQ^T}$
	$\hat{\mathbf{Y}} \rightarrow \hat{\mathbf{U}}$	$\hat{\mathbf{U}} = \hat{\mathbf{Y}}\mathbf{Q}(\mathbf{Q^TQ})^{-1}$

3.2.2 Automatic learning of algebraic models for optimisation

The automatic learning of algebraic models for optimisation (ALAMO) (Cozad, Sahinidis, and Miller, 2014; Cozad, Sahinidis, and Miller, 2015; Wilson and Sahinidis, 2017) is a regression and classification model learning methodology developed to generate algebraic models for experiments, simulations or any black box systems. It consists of a surrogate model builder that uses a set of data points and an adaptive sampler that updates the data set where the surrogate model fails. The combination of the builder and the sampler ensures that the surrogate model can give the best or a high quality estimate the sample set of data. In this work, the input data set is determined beforehand and no new experiments are conducted; therefore, only the surrogate model builder of ALAMO is required to build predictive models of interest. This surrogate model builder utilises an integer-programming-based "best subset" technique with the aim to identify relevant functional forms from a broad set of linear or nonlinear mathematical transformations of the original input descriptor variables in order to fit the response variable.

ALAMO contains a set of default linear and nonlinear transformations, or basis functions, such as monomial, bilinear, ratio, exponential, logarithmic and trigonometric functions. Examples of these transformations can be seen in Table 3.4. The term $X_j(\boldsymbol{x})$ is defined as the j^{th} basis function of the set of input variables $\boldsymbol{x}$. In the context of this work, the input variables are the 21 solute and solvent descriptors introduced in section 2.2.2. The exponent α and the scalar γ are further used to modify the basis functions with values that usually represent

TABLE 3.4: List of potential basis function forms

Category	$X_j(\boldsymbol{x})$
Polynomial	$(\boldsymbol{x})^{\alpha}$
Multinomial	$\Pi_{d \in D' \subseteq D}(x_d)^{\alpha_d}$
Exponential	$\exp(\frac{x_d}{\gamma})^{\alpha}$
Trigonometric	$\sin(\frac{x_d}{\gamma})^{\alpha}$
Expected bases	Known mathematical transformations

physically reasonable or common statistical fitting functions (e.g. $\alpha = \{\pm 0.5, \pm 1, \pm 2, \pm 3, \pm 4\}$ and $\gamma = \{0.1, 1, 10\}$) (Cozad, Sahinidis, and Miller, 2014). A user can also designate more complicated custom transformations such as sigmoid, Arrhenius, and Gaussian relationships if the transformations are known to describe the response variable effectively (Wilson and Sahinidis, 2017). The surrogate model builder selects the best set of basis functions by minimising a model fitness metric in a mixed-integer programming formulation, shown in equation (3.10) below.

$$
\begin{aligned}
&\min_{r=1,\ldots,k} FM(r) \\
&\text{where} \\
&\qquad FM(r) = \quad \min \Sigma_{i=1}^{N} (z_i - \mathbf{X}_i \beta)^2 + C(r) \\
&\qquad s.t. \quad \Sigma_{j=1}^{k} y_j \leq r \\
&\qquad\qquad -M y_j \leq \beta_j \leq M y_j, j = 1, \ldots, k \\
&\qquad\qquad y_j \in \{0, 1\}, j = 1, \ldots, k
\end{aligned}
\tag{3.10}
$$

where $FM(r)$ is a choice of possible fitness metrics, z_i is the response variable at data point i, $\mathbf{X}_i$ is a vector of mathematical transformations of the original input descriptor variables (basis functions) at data points $i = 1, \ldots, N$, and β is a vector of corresponding coefficients for the mathematical transformations. The term $C(r)$ is a metric-dependent complexity penalty. A binary variable y_j is used in conjunction with big-M constraints to include or exclude certain basis functions. The β_j terms are the individual coefficients for each basis function $j = 1, \ldots, k$ in β. Further, the cardinality constraint $\sum_{j=1}^{k} y_j \leq r$ ensures that the model contains no more than r basis functions.

The model fitness metrics found in Wilson and Sahinidis (2017) are a set of model performance indicators developed from information and statistical theory to balance the bias-variance trade-off. Examples of these metrics include the corrected Akaike information criterion (AIC^c) (Akaike, 1974), the mean squared error (MSE) (Park and Klabjan, 2013), and the Bayesian information criterion (BIC) (Neath and Cavanaugh, 2012). Each of these examples may have different formulations for both the first term of $FM(r)$ and the second term $C(r)$; however, every fitness metric utilises the sum of squared errors (SSR), seen in equation (3.11). The SSR is equivalent to the PRESS metric (equation (3.6)). It helps quantify the quality of fit to the training set and penalise the corresponding fit directly based on the number of nonzero regression coefficients (Wilson and Sahinidis, 2017).

$$SSR = \sum_{i=1}^{N} \left(z_i - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \tag{3.11}$$

A potential issue with using the surrogate model builder is determining the initial selection of basis functions. Increasing the size of the initial set increases the number of possible models combinatorially. In this work, several tests are carried out to select a fitness metric and a set of input basis functions.

Selecting a fitness metric using an independent data set

In this work, an independent data set was used to determine a suitable fitness metric. The data set is adapted from the work of Cherkassky, Gehring and Mulier (1996), who compared statistical and neural network methodologies for function estimation. The function of interest (equation (3.12) is formed of four uncorrelated and randomly distributed inputs, x_i (for $i = 1, 2, 3, 4$) ranging from -0.25 to 0.25, for 500 data points. The minimum value of the function y is 0.71, with a maximum of 1.42 and an average of 1.01. Therefore, the behaviour of the function is only dependent on its expression and has no correlation between its four inputs. Thus, the function can be used as an example to showcase the flexibility and precision of the ALAMO framework.

$$y = \exp(2x_1 \sin(\pi x_4)) + \sin(x_2 x_3) \quad \text{where} \quad x \in [-0.25, 0.25] \tag{3.12}$$

The fitness metrics considered are the metrics found in the work of Wilson and Sahinidis

(2017). These include the BIC, Mallow's Cp (Mallows, 1973) (Cp), AICc, Hannan-Quinn information criterion (HQIC) (Hannan and Quinn, 1979), the mean squared error (MSE) (Park and Klabjan, 2013), and the risk inflation criterion (RIC) (Foster and George, 1994). The fitness metric will be chosen based on the RMSE, R^2 and the model size as the goal is to have an accurate model while maintaining low model complexity. An initial set of polynomial and two term multinomial (binomial) basis functions is chosen with different values of α. The list of basis functions is found in Table 3.5. The terms x_d and x_e represent the input parameters where $d = 1, 2, 3, 4, e = 1, 2, 3, 4$ and $d \neq e$. Thus the total number of basis functions is 32 as there are 20 functions from the polynomials (four functions per power) and 12 from the binomials (C_2^4 per power).

TABLE 3.5: Initial basis functions for the selection of a fitness metric

Category	Symbol	Power (α)
Polynomial	$(x_d)^\alpha$	-1, 1, 2, 3, 4
Binomial	$(x_d x_e)^\alpha$	-1, 1

TABLE 3.6: Comparison of fitness metrics using the basis set found in Table 3.5

Metric	RMSE	R^2	Model size
BIC	0.00698	0.997	6
Cp	0.02490	0.962	2
AICc	0.01230	0.991	4
HQIC	0.01230	0.991	4
MSE	0.01230	0.991	4
RIC	0.02490	0.962	2

In Table 3.6, it can be seen that the model obtained with the BIC metric has an RMSE value of 0.00698 in comparison to an average value of y of 1.01 over the data set. These results suggest a near-perfect fit of the input data with some residual error due to the random noise. The BIC model has a model size of 6, and is given by:

$$\hat{y} = 0.38x_1^2 + 0.34x_4^2 + 0.22 \times 10^{-6}(x_2x_3)^{-1} + 6x_1x_4 + 0.99x_2x_3 + 0.99 \tag{3.13}$$

As seen in equation (3.13), it can be seen that only the quadratic terms x_1^2, x_4^2 and the bilinear terms x_1x_4, x_2x_3, and $(x_2x_3)^{-1}$ are used with corresponding weights in the model. For the AICc, HQIC and MSE metrics, the RMSE values are roughly double the BIC, and the Cp and RIC metrics are nearly quadruple. The R^2 values for the AICc, HQIC and MSE

metrics are slightly lower but there is a larger difference when comparing the Cp and RIC metrics. A further test is carried out to assess the effect of removing a basis function. The basis function $(x_d x_e)^{-1}$ is removed as $(x_2 x_3)^{-1}$ is used in equation (3.13). The modified basis set can be seen in Table 3.7 and the results with the reduced basis set can be seen in Table 3.8. The following model is obtained:

TABLE 3.7: Modified initial basis functions for the selection of a fitness metric

Category	Symbol	Power (α)
Polynomial	$(x_d)^\alpha$	-1, 1, 2, 3, 4
Binomial	$(x_d x_e)^\alpha$	1

TABLE 3.8: Comparison of fitness metrics using the reduced basis set in Table 3.7

Metric	RMSE	R^2	Model size
BIC	0.00709	0.997	5
Cp	0.02490	0.962	2
AICc	0.00703	0.997	8
HQIC	0.00709	0.997	5
MSE	0.00707	0.997	6
RIC	0.02490	0.962	2

$$\hat{y}_{reduced} = 0.38x_1^2 + 0.35x_4^2 + 5.9x_1x_4 + 0.99x_2x_3 + 0.99 \tag{3.14}$$

The BIC is chosen as the fitness metric due to the performance in the first test as well as the second test. The resulting BIC model found in equation (3.14) has does not contain the $(x_2 x_3)^{-1}$ term found in equation (3.13) but is otherwise very similar. There was only a negligible change in performance. The results found in Table 3.8 show improvements for the AICc, HQIC, and MSE metrics as their RMSE values have decreased to about 0.007 at the cost of an increase in model size. The Cp and RIC metrics do not change in performance as both the RMSE and R^2 values remain the same from equation (3.13) to equation (3.14). These test results suggest that not every basis function captures the same amount of variance in the experimental data set. In this case, the removal of the $(x_2 x_3)^{-1}$ term only resulted in a negligible change in performance. This observation is reinforced by the fact that the coefficients of the other terms only changed slightly. Another observation is that the choice of fitness metric is sensitive to changes in the basis sets.

3.3 Development of data-driven models using the experimental $\Delta G^{o,m}_{s,i,j}$ database

3.3.1 Data set used for the development of the data-driven models

In the development of predictive models, aside from the mathematical methodology that supports the model, experimental or observed data are required train the model to the system of interest. In this work, all models are nonaqueous, meaning that water is excluded as a solvent, with the exception of the self-solvation of water. Thus the database of 2364 data points found in Section 2.2.2 is reduced to 2167 data points. Usually, a database of experimental points is divided into a training data set for model development and a test data set for model validation. In developing robust, predictive models, the input data set of 2167 points is divided into a training set (for calibrating the model) and a testing set (for validating the model). When considering a specific data set, a decision needs to be made as to what proportion of the data set must be used in the training and testing sets, as some splits of the data set may lead to overfitting. Overfitting refers to when the model predicts the training set of data with high accuracy but does not predict the testing set or independent data with low accuracy. The optimal proportion of training to testing data is investigated in this work. Another aspect of dividing the data set is the distribution of the solute/solvent systems equitably across both groups of data in terms of molecular representation. If some types of solute/solvent systems exist solely in the testing set, the developed data-driven model will have no information on how to represent those kinds of systems. An example of this is if all of the alcohol/alkane data points existed in the testing set, a user would not be able to train the developed model to handle those systems and thus, it may have poor predictive ability for those kinds of systems. Therefore, it is crucial to have a good spread of the types of data points in both the training and testing sets. This problem is addressed in Section 3.3.2.

3.3.2 Cross-validation

In this work, cross-validation is used as a model validation technique to address the overfitting problem. A simple example of the technique is leave-one-out cross-validation (LOO CV) in which a single data point from the original data set of N data points and evaluates the regression model N times, omitting a new data point is omitted each time. Then, the

model performances (usually using PRESS) of the N regressions are aggregated to provide an estimate of the overall model performance using the data set. This technique allows a user to intuitively understand how well a model would perform using a certain data set. However, as N becomes sufficiently large, a user would have to evaluate a larger number of models, and this becomes computationally intensive. The generalised form of LOO CV is leave-p-out cross-validation (LpO CV) in which p data points are omitted from the training set. In LpO CV, the model is evaluated C_p^N times, where C_p^N is the binomial coefficient. Similarly to LOO CV, the issue with LpO CV is that with larger values for N and $p > 1$, C_p^N becomes very large, making this approach computationally infeasible. For example, with $N = 2167$ and $p = 10$, $C_{10}^{2167} \approx 6 \times 10^{26}$. LOO and LpO CV are exhaustive forms of cross-validation. They provide a fully quantified view of the effect of the training and testing sets on the models; however, they are computationally intensive.

In contrast, in non-exhaustive cross-validation, not all possible combinations of C_p^N are computed, which reduces the computational load while deriving an approximation of LpO CV. Two examples of non-exhaustive cross-validation are Monte Carlo and k-fold validation. In the Monte Carlo technique, the input data set is split randomly into k equal parts k times, meaning there is a chance some data points may never be in either the training set or in the testing set. There may also be repeating points across the number of k splits, which can be detrimental as it may cause bias in the model. For the k-fold technique, the input data set is also split randomly into k equal parts k times. However, the data are split such that each data point will appear in the testing set just once, with no repetition meaning each data point is used both in training and validating the model during the process. Examples of the Monte Carlo and k-fold techniques can be seen in Table 3.9. The example shows the partitioning of $N = a$ integer data points (1 to a) with $k = 3$, resulting in three training/testing sets for the Monte Carlo and k-fold techniques. It can be seen that there are no repetitions in the testing sets for the k-fold techniques, but data points '5' and '8' repeat in the Monte Carlo testing sets. In both these techniques, k models are regressed using $N = k$ data points and validated using k data points, and the performance metrics of all models are then averaged.

The decision to use k-fold or Monte Carlo validation for $\Delta G_{s,i,j}^{\circ,m}$ modelling is made by carrying out a test using a series of splits from 2 to 20 and by using the PLS methodology. The test only compares model performance using the testing data set as this represents the predictive capability of the models. In this test, several performance metrics are used in the

TABLE 3.9: Examples of the training and testing data splits in the Monte Carlo and k-fold cross-validation techniques on a range of integers from 1 to 9, where $k = 3$

Monte Carlo			k-fold		
Model Number	Training set	Testing set	Model Number	Training set	Testing set
1	[1, 2, 3, 4, 6, 7]	[5, 8, 9]	1	[2, 4, 5, 6, 8, 9]	[1, 3, 7]
2	[1, 3, 5, 6, 7, 9]	[2, 4, 8]	2	[1, 3, 4, 5, 6, 7]	[2, 8, 9]
3	[2, 3, 4, 7, 8, 9]	[1, 5, 6]	3	[1, 2, 3, 7, 8, 9]	[4, 5, 6]

cross-validation process. These metrics are calculated for each model L in the k validation. These include the coefficient of determination (R^2_L) for model L, the root mean squared error (RMSE_L), and the model, $Bias_L$. Two other metrics, the optimal number of PLS components and the model size, are also considered. Variable R^2_L is given by:

$$R^2_L = \left(\frac{\sum_{n \in N_{L,test}} \left(\Delta \underline{G}^{\circ,m,exp}_{s,n} - \Delta \underline{G}^{\circ,m,exp}_{s,mean}\right)\left(\Delta \underline{G}^{\circ,m,pred}_{s,n} - \Delta \underline{G}^{\circ,m,pred}_{s,mean}\right)}{\sqrt{\sum_{n \in N_{L,test}} \left(\Delta \underline{G}^{\circ,m,exp}_{s,n} - \Delta \underline{G}^{\circ,m,exp}_{s,mean}\right)^2 \sum_{n \in N_{L,test}} \left(\Delta \underline{G}^{\circ,m,pred}_{s,n} - \Delta \underline{G}^{\circ,m,pred}_{s,mean}\right)^2}} \right)^2 \tag{3.15}$$

where $N_{L,test}$ is the subset of test data used in the L validation, whereas m is the individual data point in $N_{L,test}$. Variables RMSE_L and $Bias_L$ are defined below:

$$\text{RMSE}_L = \frac{1}{N_{L,test}} \sum_{n=1}^{N_{L,test}} (\Delta \underline{G}^{\circ,exp}_{s,n} - \Delta \underline{G}^{\circ,pred}_{s,n})^2 \tag{3.16}$$

$$Bias_L = \frac{1}{N_{L,test}} \sum_{n}^{N_{L,test}} \Delta \underline{G}^{\circ,m,exp}_{s,n} - \Delta \underline{G}^{\circ,m,pred}_{s,n} \tag{3.17}$$

In the PLS models, the optimal number of principal components or latent variables is given by the number of components with the lowest MSE-CV, whereas in the ALAMO models, the model size is the number of terms in the model with the lowest objective value from the optimisation.

Figures 3.1 and 3.2 contain the results of the cross-validation study. It can be seen that the R^2_L values in both figures plateau after 4 splits; however, the R^2CV values have small ranges of 0.775 to 0.780. Similarly, the RMSE CV values have a small range of 0.860 to roughly 0.866 kcal mol^{-1}, which is negligible compared to most values of the free energy of solvation, which are three orders of magnitudes larger. The bias values are minimal and are a direct result of the linear regression in the PLS methodology. The optimal number of latent

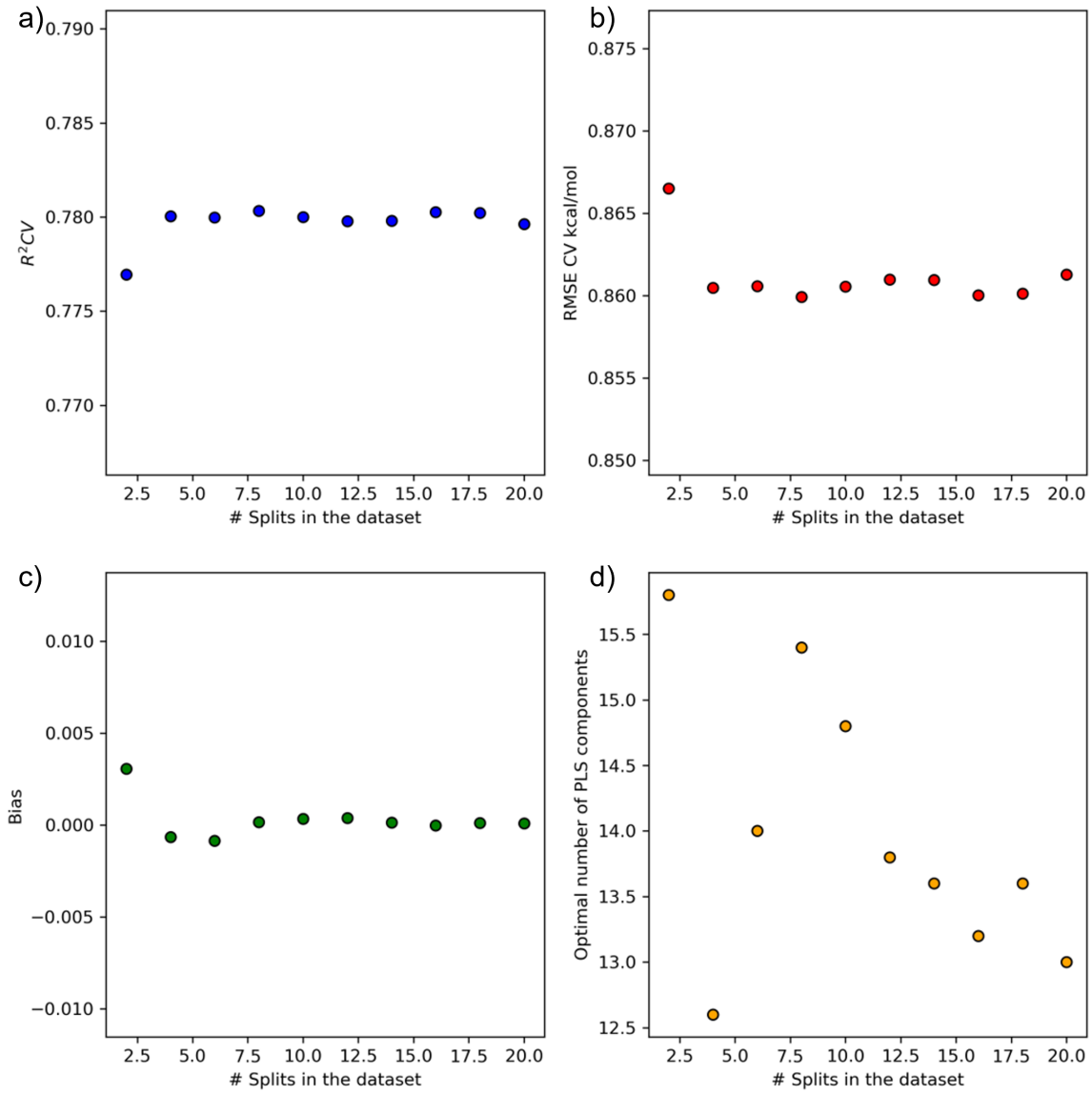

FIGURE 3.1: k-fold cross-validation results for the PLS methodology with the performance metrics R^2CV (a), RMSE (b), Bias (c) and optimal number of components (d).

variables fluctuates between 12 to 16 in both cases. It can be concluded, that for the system of interest, there is effectively no difference in using the k-fold validation or the Monte Carlo methodologies, at least in the case of PLS models. The k-fold cross-validation methodology was chosen to develop subsequent models because each data point is tested at least once.

3.3.3 Correlation between target variable $\Delta G^{o,m}_{s,i,j}$ and solute/solvent descriptors in the nonaqeuous experimental database

Several steps have been taken to curate the experimental data that will be used in the development of data-driven models. In section 3.3.1, the experimental database was reduced

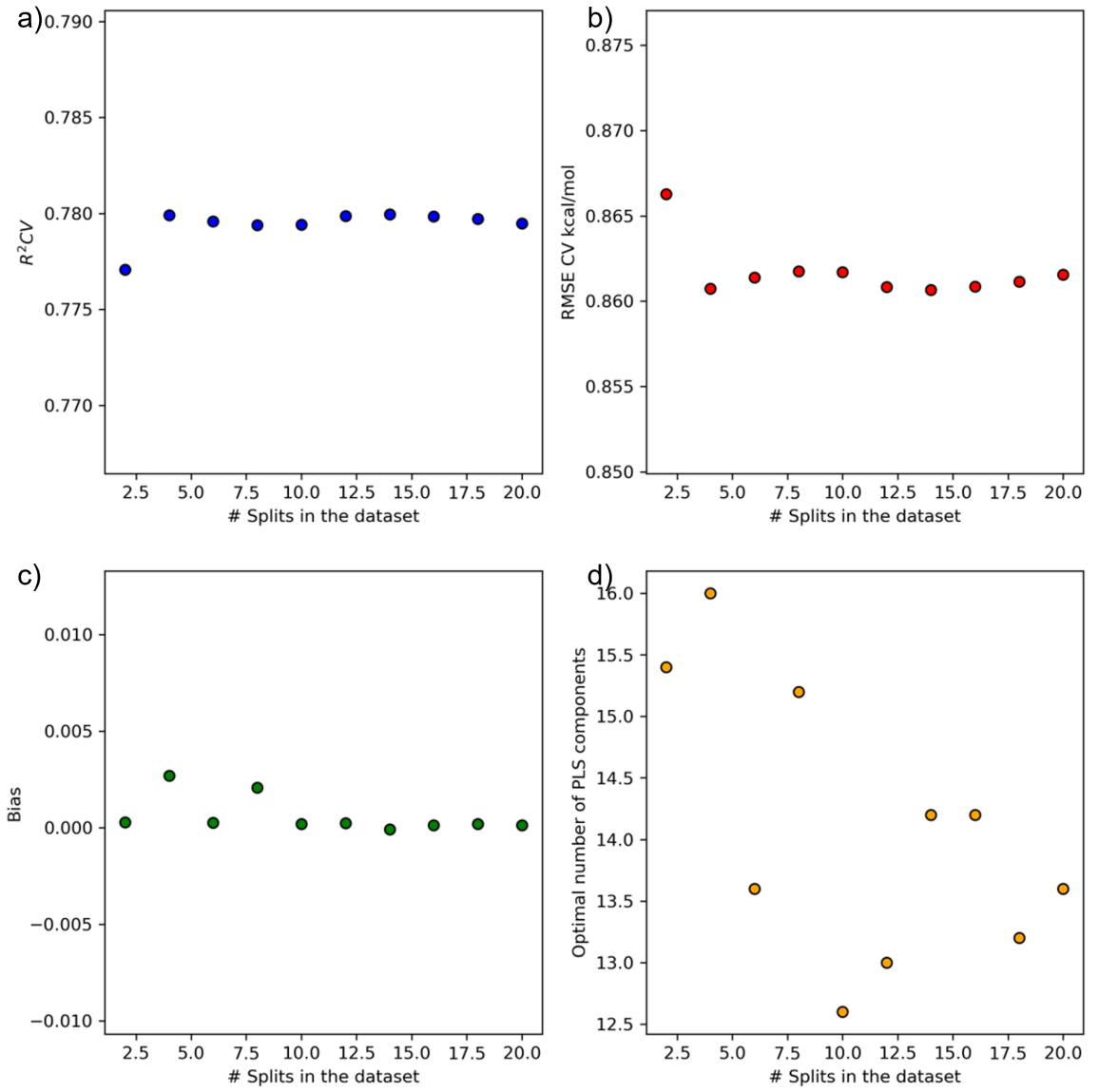

FIGURE 3.2: Monte Carlo cross-validation results for the PLS methodology with the performance metrics R^2CV (a), RMSE (b), Bias (c) and optimal number of components (d).

from 2364 to 2167 data points to reduce the impact of water on model performance whereas, in section 3.3.2, k-fold and Monte Carlo cross-validation techniques were used to determine if the size of training and testing sets affected the performance of PLS models. It was found that over the the range of splits tested, there was only a negligible change in performance. A further examination of the data is carried out in this section to study the relationships between the target variable $\Delta G^{o,m}_{s,i,j}$ and the solute/solvent descriptor variables.

In the nonaqueous database of 2167 data points, there are 294 solute molecules and 210 solvent molecules. The first point of the study is to see if the number of solute or solvent molecules affects the correlations between the target variable $\Delta G^{o,m}_{s,i,j}$ and the descriptor

variables. The k-fold cross-validation technique is employed to randomly shuffle each solute/solvent data point in the nonaqueous database into splits of 2, 3, 5, 10, 15, and 20. The correlation coefficient is calculated between the target variable and each descriptor variable for each split of training and testing data. This analysis highlights the relationship between the number of solute and solvent molecules and the correlation coefficients between the target and descriptor variables. The Pearson correlation coefficient is used to measure the linear relationship between two data sets (in this case the target variable and each descriptor variable, resulting in 21 correlation coefficients) and is expressed in equation (3.18).

$$r = \frac{\sum_n^N (x_n - m_{n,x})(y_n - m_{n,y})}{\sqrt{(\sum_n^N (x_n - m_{n,x})^2 \sum (y_n - m_{n,y})^2}} \tag{3.18}$$

where m_x is the mean of the vector x and m_y is the mean of the vector y. The target variable is used in the x position whereas the 21 descriptor variables are used interchangeably in the y position. The range of values for the Pearson correlation coefficients is -1 to +1, where higher absolute values indicate stronger positive or negative relationships between variables and 0 indicates no correlation. Figure 3.3 contains four plots that show the average number of solutes and solvents, and the averaged correlation coefficients of the 21 descriptor variables against the target variable $\Delta G_{s,i,j}^{o,m}$ for both training and testing sets against a number of splits. The average number of solutes and solvents diverges as the number of splits increases for both the training and testing sets. This divergence shows that the number of unique solutes and solvents varies proportionally to the size of the data set. As the size of the training set increases, the difference between the number of solutes and solvents widens. In contrast, the size of the testing set decreases with the number of splits, the average number of solutes and solvents become closer to one another. However, it must be noted that this trend in diversity is not guaranteed and is dependent on how the data is shuffled.

Several interesting points can be drawn when observing the bottom plots that show the averaged correlation coefficients for both the training and testing sets. First, there is almost no difference in the correlation coefficient values for the training and testing sets. It can be seen that only the correlation coefficient of the molecular volume, V_m, increases in value as the size of the testing set decreases. The correlation coefficients of the other descriptors remain roughly constant for all splits. A point to note is that the correlation coefficients can vary when comparing individual cases; however, the average values confirm that the

distribution of solutes to solvents has little to no effect on the correlations between the target and descriptor variables. The solute descriptors (which are described in section 2.1) have the most influence on the free energy of solvation as S, V_{mc}, and π have coefficients whose absolute values exceed 0.5, which is considerably larger than the solvent descriptor with the largest coefficient, $\log K_{ow}$, at 0.25. The second-largest solvent descriptor, μ^j, has a similar value and is comparable to the other solute descriptors whose absolute values are near or larger than 0.2. The rest of the solvent descriptors are evenly distributed over a range of -0.2 to 0.2. This shows that while the numbers of solutes and solvents do not affect the correlation coefficients, the solute descriptors are more correlated with the free energy of solvation.

Therefore, these observations show that as the average number of solutes converge towards the number of solvents, model performance does not improve. However, this analysis does not consider the distribution of solute and solvent molecules. While the number of solutes outnumbers the solvents, some solvents may have 20 solutes whereas some may only have one. The issue of an uneven distribution of solutes to solvents can only be addressed by adding more experimental data points to represent each solute and solvent.

3.3.4 Determining basis sets for the ALAMO methodology

It was demonstrated in Section 3.2.2 that the ALAMO framework is robust in selecting the best subset of basis functions while being able to capture the variance of the data set with high precision. As such, for any set of data with a given number of descriptors, the ALAMO framework can be used with a wide variety of initial basis functions to always obtain the best subset. However, by supplying a large number of initial basis functions, the number of possible combinations of basis functions is combinatorial and becomes computationally intensive. Therefore, reducing the initial set of basis functions, reduces the required amount of computational power, while providing insight into the linear or nonlinear relationships in the data set. Therefore, in this section, the most appropriate set of basis functions for ALAMO is presented using the data set of 2167 points, and cross-validation is used to evaluate the best set of basis functions. The free energy of solvation data contains 21 solute and solvent descriptors for 2167 data points. A preliminary assumption about the basis functions is that some categories of functions can better capture the experimental data set variance than others. This assumption was shown to be correct in Section 3.2.2 as removing the inverse binomial term only decreased the RMSE value by a small amount. However, while this is the case, it

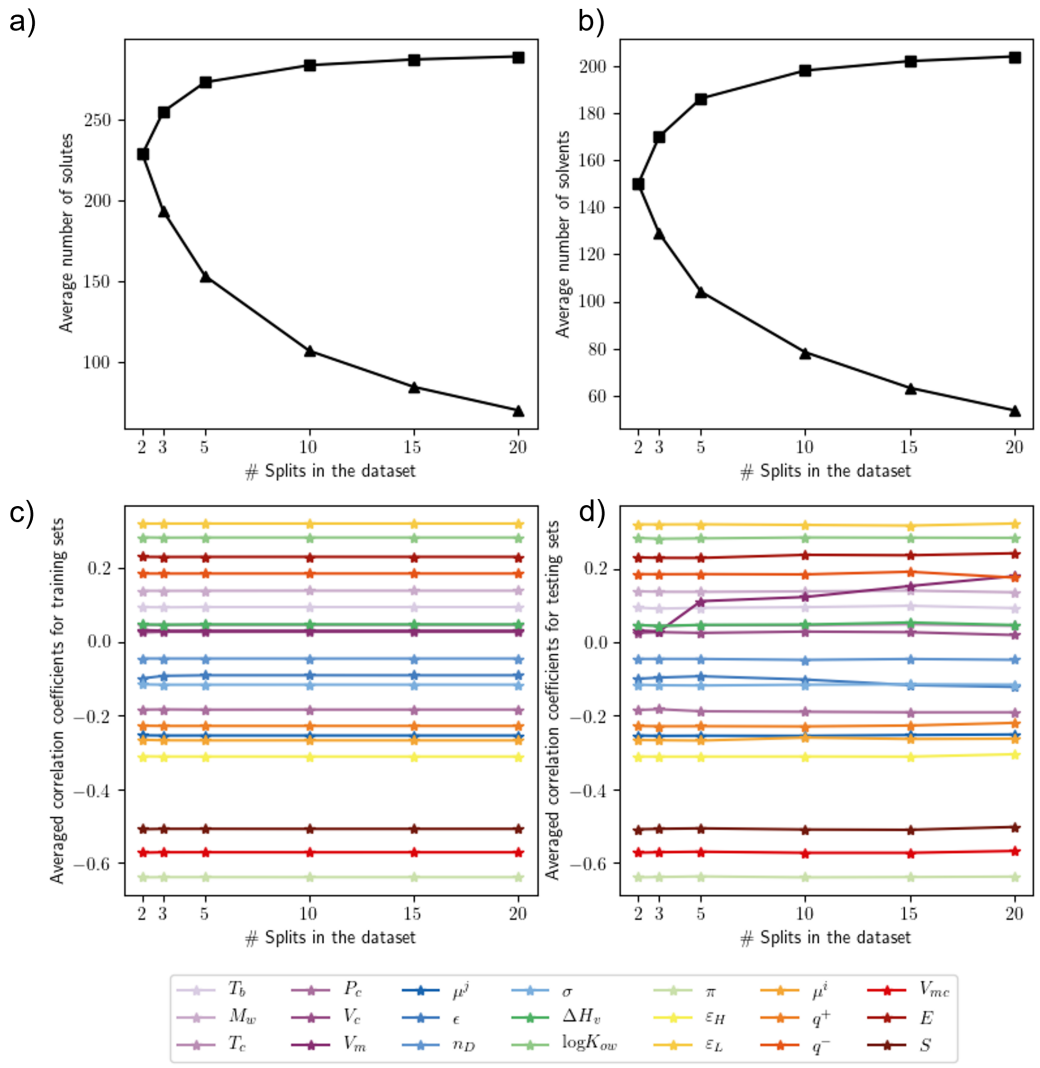

FIGURE 3.3: Plots of the average number of solutes (a) and solvents (b), and the averaged correlation coefficients of the 21 descriptor variables with the target variable $\Delta G^{o,m}_{s,i,j}$ for both training (c) and testing (d) sets across the number of splits. The square and triangle markers in the upper plots are the training and testing data, respectively.

is difficult to ascertain if combinations of categories of basis functions will have a synergistic or detrimental effect on the prediction of the experimental free energy of solvation data. Therefore, in this study, I will test categories of basis functions individually and combine the best performing categories in a second test. The tests will also use 10-fold cross-validation to average out any bias in the models. The outline of this study can be found in Table 3.10.

After the data set is split into ten 9:1 training/testing sets, 10 ALAMO models are regressed for every power of every basis set. This means both the individual and combined basis sets utilise the same data set to ensure a fair comparison between the sets of models.

Table 3.10: Outline for determining an appropriate set of basis functions for the ALAMO model.

Step	Action
1	Split the data into 10 equal parts for 10-fold cross-validation
2	Regress model using 10-fold CV for the individual functions found in Table 3.11
3	Combine the best performing sets and regress the ALAMO model using 10-fold CV
4	Compare the results and determine the most appropriate set of basis functions

Table 3.11: Individual functions considered for an appropriate set of basis functions

Basis set	Power (α)
x_d^α	-1, 1, 2, 3
$(x_d x_e)^\alpha$	-2, -1, 1, 2 ,3
$\sin(x_d)$	-
$\cos(x_d)$	-
$\exp(x_d)$	-
$\log(x_d)$	-

Therefore, if the polynomial term x_d^α seen in Table 3.11 has four powers, -1, 1, 2, and 3, ten models are regressed for each power, resulting in four sets of ten ALAMO models whose performance metrics are then averaged to give four averaged models of -1, 1, 2, and 3 powers. This process is repeated for each basis set. The results for the individual basis functions are shown in table 3.12. The performance metrics are the RMSE, R^2 and the model size.

Table 3.12: Averaged metrics for individual basis functions considered for an appropriate set of basis functions

Model name	Basis set	RMSE (kcal mol^{-1})	R^2	Model size
L	x_d	0.8686	0.7692	10.1
L-1	x_d^{-1}	1.223	0.5422	8.8
L2	x_d^2	0.9343	0.7329	14.3
L3	x_d^3	1.064	0.6525	13.3
S	$\sin(x_d)$	1.404	0.3977	9.8
CS	$\cos(x_d)$	1.530	0.2841	9.5
EXP	$\exp(x_d)$	1.489	0.3222	7.1
LOG	$\log(x_d)$	1.313	0.4724	9.4
BN	$(x_d x_e)$	0.511	0.9200	125.3
BN-2	$(x_d x_e)^{-2}$	1.133	0.6064	56.9
BN-1	$(x_d x_e)^{-1}$	0.8885	0.7586	75.7
BN2	$(x_d x_e)^2$	0.5382	0.9114	114.7
BN3	$(x_d x_e)^3$	0.7441	0.8307	90.1

In Table 3.12, the S, EXP and LOG models do not perform well and are no longer considered a part of the input basis functions. It is clear that the binomial terms, $(x_d x_e)^\alpha$ can

accurately describe the free energy of solvation of data for $\alpha = 1, 2$ with slightly worse performance from $\alpha = -1$ and 3. However, according to the model size, models that utilise binomial terms effectively contain from 75 to 125 bilinear terms. Conceptually, the binomial models have less variance captured per term. In contrast, the linear models L, in Table 3.12, capture 76.92% of the variance of the experimental data with just 10 terms on average. These results suggest the experimental free energies of solvation are related to the 21 solute and solvent descriptors mostly linearly and bilinearly.

To investigate any further synergistic or detrimental combinations, some combined basis sets that include quadratic, logarithmic, and inverse terms are considered. A list of combined basis sets can be found in Table 3.13. A potential user may also input all of the basis sets found in Table 3.11; however, since the number of possible biliear terms per power is 210, including all the terms above combinatorially increases the required computation time. Therefore, in the interest of time, the combined basis sets seen in Table 3.13 are considered.

Table 3.13: List of combined basis sets for the ALAMO model

Model type	Basis functions
A	x_d, x_d^2
B	x_d, x_d^2, $\log(x_d)$
C	x_d, x_d^{-1}
D	x_d, x_d^{-1}, $(x_d x_e)$
E	x_d, x_d^{-1}, $(x_d x_e)$, $(x_d x_e)^{-1}$
F	x_d, $(x_d x_e)$
G	x_d, $(x_d x_e)$, $(x_d x_e)^2$
H	x_d, $(x_d x_e)$, $(x_d x_e)^2$, $(x_d x_e)^3$

Table 3.14: Combined basis functions considered for the final set of basis functions

Basis set	RMSE (kcal mol^{-1})	R^2	Model size
A	0.8093	0.7998	28.2
B	0.8094	0.7994	28.4
C	0.8250	0.7918	24.8
D	0.4997	0.9236	121.1
E	0.4298	0.9434	154.9
F	0.5100	0.9205	114.5
G	0.4011	0.9508	175.7
H	0.3603	0.9603	211.7

With the combined basis sets available, step 3 in Table 3.10 can now be carried out. A similar analysis to Table 3.12 is found in table 3.14. The linear models L have an R^2 value

of 0.7692, compared to the quadratic models L2 which have an R^2 value of 7329. Using both these basis functions should result in models which have a significantly larger R^2 value; however, the model A only has a value of 0.7998. This small increase in R^2 value shows that while there is a synergistic effect, there is overlap in the variance captured by the linear and quadratic effects. The same effect is seen when comparing model C, which has an R^2 value of 0.7918. This suggests that the linear effects outweigh the quadratic and inverse effects. There is essentially no change in performance from models A to B, suggesting the logarithmic terms in this combination of functions do not contribute anything towards the performance of the regressed model. Among the linear, quadratic, inverse and logarithmic terms, it seems that the linear terms represent the experimental data effectively, with support from the quadratic and inverse terms. The dynamic changes when binomial terms are added into the initial set of basis functions. The models BN achieve an RMSE value of 0.511 kcal mol^{-1} and a R^2 value of 0.92 whereas models F achieve an RMSE value of 0.51 kcal mol^{-1} and a R^2 value of 0.9205, indicating that the linear terms only offer a small improvement in the performance compared to the binomial terms. However, an interesting observation is made when comparing model sizes. Models L and BN have model sizes of 10.1 and 125.3, respectively, whereas model F has a model size of 114.5. This shows that while the performance of the model did not increase significantly, there is a notable synergistic effect between the linear and binomial terms as the model size decreased. Model D achieves a lower RMSE value of 0.4997 and an R^2 value of 0.9236; however, this shows that the inverse terms do not contribute significantly to the model performance, as the average model size from model F to model D increases by 6.1 which is close in value to the average model size of model L-1 (8.8). Regarding the G and H sets, there is a constant drop in RMSE value as the binomial squared and cubed terms are added to the basis set; however, the model complexity also increases. for basis sets B and C, the inclusion of the binomial and inverse binomial terms significantly improves the performance relative to model A.

The analysis of the combined data sets suggest that the experimental solvation free energy is well represented by the binomial terms, and a synergistic effect can be achieved by using incorporating linear terms. These models have been optimised such that the BIC fitness metric is always minimised, preventing any overfitting in the models. Therefore, while averaged, it is important to note that the initial number of binomial terms is 210 compared to the 21 linear terms, indicating that the variance captured by each term is on average larger for each linear

term, but this fact is overshadowed by the large number of binomial terms. Therefore, it would be interesting to study how the ALAMO model performs with respect to the training and testing sets instead of an overall look.

3.3.5 Determining the number of k-fold cross validation splits for the development of PLS, QPLS and ALAMO models

This section focuses on using k-fold cross validation to determine the optimal number of splits and the spread of data in the training/testing sets for the PLS, QPLS, and ALAMO frameworks. The approach detailed in Section 3.3.3 is applied to the three modelling approaches. The experimental data sets used for this cross-validation study are the same data sets from the correlation study in section 3.3.3. This results in the experimental database being divided into training and testing sets for splits of 2, 3, 5, 10, 15 and 20. The same performance metrics are used, namely the R^2, RMSE, and Bias with the optimal number of components for the PLS and QPLS models and the ALAMO model size. For the ALAMO methodology, several models are regressed using the basis sets found in section 3.3.2. The performance of the PLS, QPLS and ALAMO models against their respective testing data is found in figures 3.4, 3.5, and 3.6 respectively.

In figure 3.4, the RMSE values of the PLS models decrease from 0.860 to 0.850 kcal mol^{-1} from 2 to 20 splits. This trend may suggest that 20 splits are optimal; however, the range of RMSE values span only 0.001 kcal mol^{-1}. This range shows the models essentially have no difference in performance, and the number of splits in the training and testing sets is irrelevant for the PLS methodology. In contrast, the QPLS models in figure 3.5 has the worst performance for 2 splits, an optimum of 5 and worsening performance for 10 to 20 splits. The range of errors is slightly larger with RMSE values of 0.866 kcal mol^{-1} to 0.837 kcal mol^{-1}, but the difference is also negligible. For the PLS and QPLS methodologies, the proportion of data in the training and testing sets does not affect model development significantly.

For the ALAMO methodology, models were cross-validated using the basis sets found in Table 3.13 using the same performance metrics seen earlier. Figure 3.6 shows the trends for basis sets except for basis set H as it was challenging to converge the models with this basis set for all of the 20-fold splits. Some data points were cropped out from the plots as the performance was too poor. The tabulated version of the values can be found in Appendix C.1 in Tables C.1, C.2, C.3, and C.4. An example of this are the R^2 and RMSE values for basis

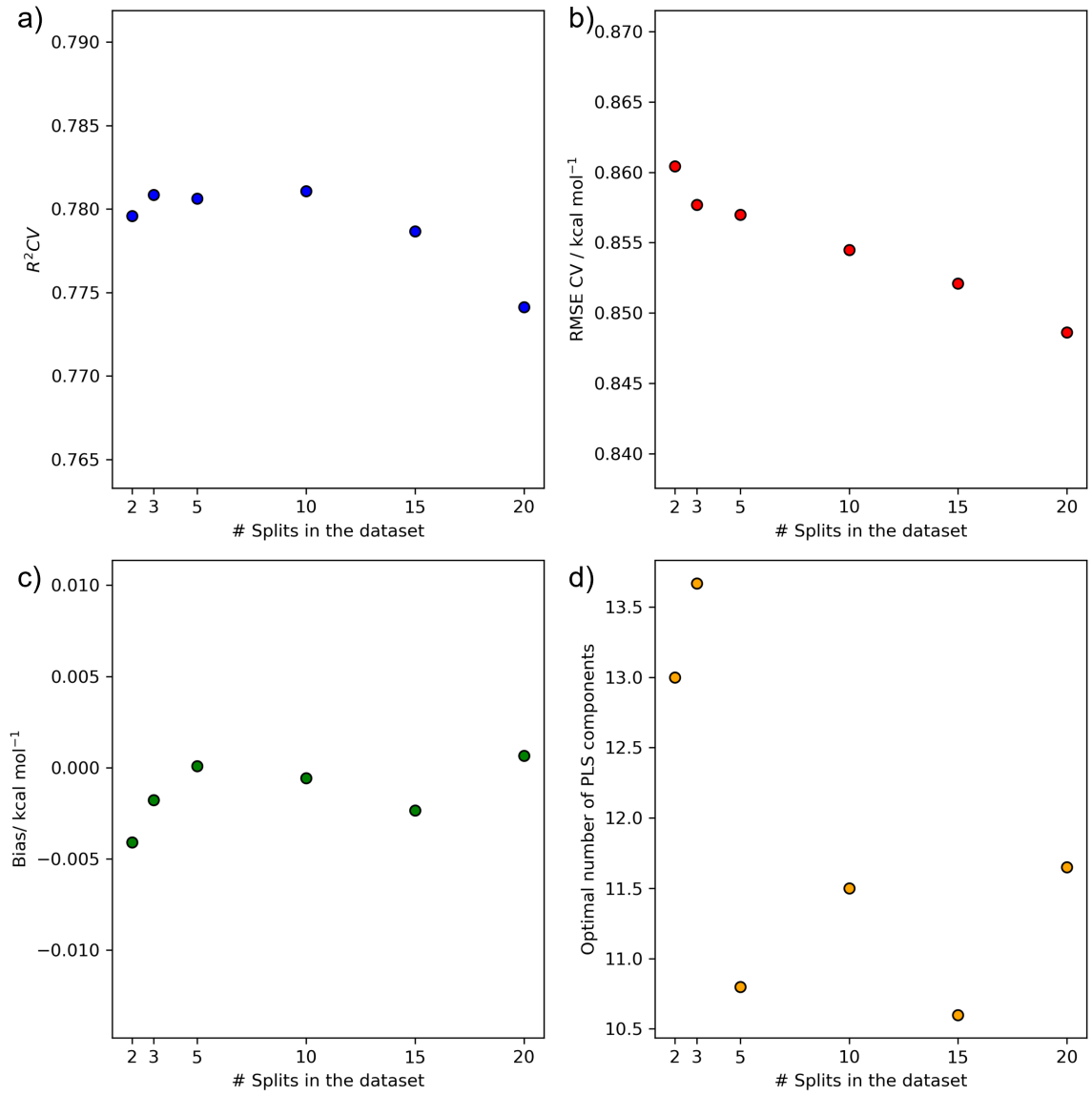

FIGURE 3.4: k-fold cross-validation results for the PLS methodology with with the performance metrics R^2CV (a), RMSE (b), Bias (c) and optimal number of components (d)..

sets D, F, and G for 2 splits in the two top panels in Figure 3.6. This poor performance is attributed to the lack of some solutes or solvents in the training set. A split of 2 means the database is in two parts, where all data points of a solute or solvent may be shuffled into only the testing set. This shuffling would mean that the ALAMO model is not trained on some molecules. As a result, the ALAMO models have no information about how to represent said molecules. This finding shows that the ALAMO framework has poor extrapolation capacity, which is a common flaw among data-driven models. However, this poor performance is not observed in the PLS and QPLS models of Figures 3.4 and 3.5. A possible reason for this is that the PLS and QPLS framework projects data points onto new axes that capture the data trend.

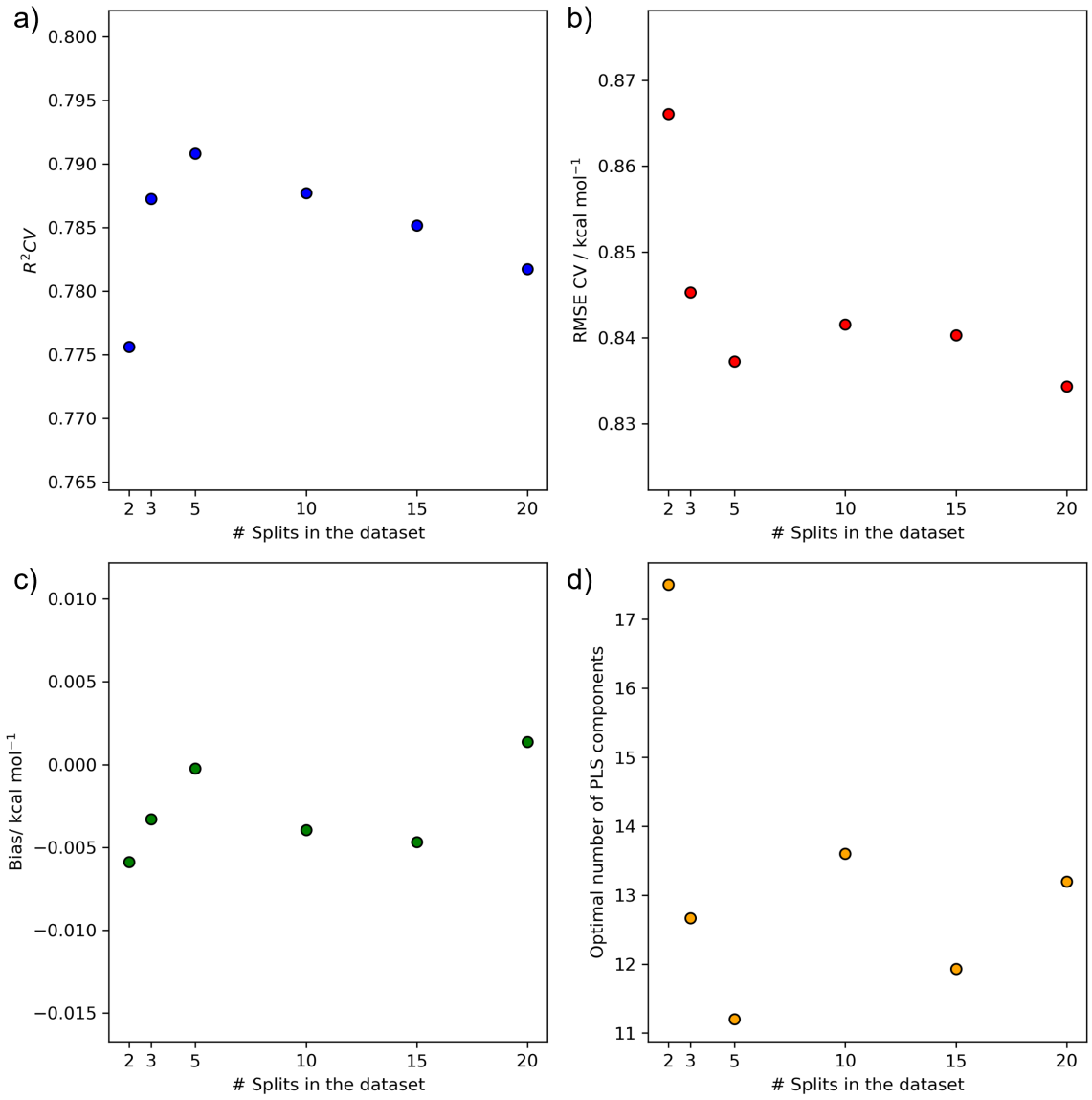

FIGURE 3.5: k-fold cross-validation results for the QPLS methodology with the performance metrics R^2CV (a), RMSE (b), Bias (c) and optimal number of components (d).

These new axes follow the direction of maximum variance in the data, which may account for the data points the ALAMO model failed to capture. Conversely, the ALAMO framework allows for a closer approximation of the data points as seen with the more complex basis sets. However, this may cause the framework to be more sensitive to outliers or molecules not found in the training set. This poor performance could be addressed is by adding a second layer of validation by repeating the test several times and averaging those results. This process is not covered in this work. Otherwise, the R^2 performance of the basis sets is relatively consistent across the number of splits with a range of 0.80 to 0.87. For the RMSE values, basis sets A, B, and C have similar performance with no difference across the splits and the RMSE

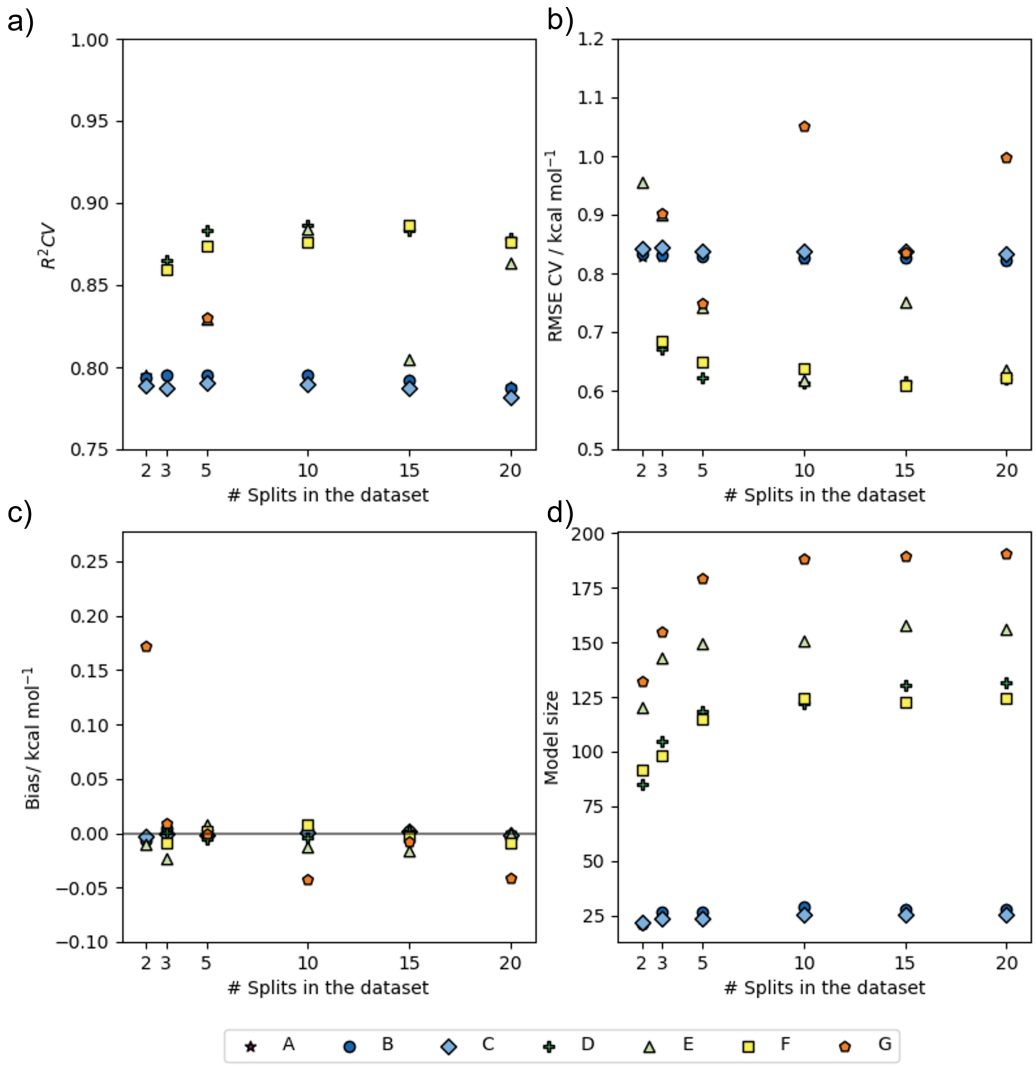

FIGURE 3.6: k-fold cross-validation results for the ALAMO methodology with the performance metrics R^2CV (a), RMSE (b), Bias (c) and optimal number of components (d).

remains around 0.84 kcal mol^{-1}. This result is similar to the RMSE values found in the PLS and QPLS models from figures 3.4 and 3.5. The D and F basis sets have the best performance of approximately 0.65 kcal mol^{-1} for splits 3, 5, 10, 15 and 20. Basis set E also has the best performance for splits 10 and 20; however, it has a higher error for splits 2, 5 and 15. The same case is seen for basis set G as the RMSE ranges from 0.75 to 1.05 kcal mol^{-1}; however, it has the largest model size.

An important observation is that the average number of terms in the model increases with the number of splits, but this trend does not guarantee an increase in model performance. In Table 3.11, it has been shown that the inverse and squared bilinear terms are known to

have good performance in isolation. The basis sets D, E, F, and G basis sets in Table 3.13 can be used to show show how these terms can have positive or negative effects on model performance when combined together and with other terms. For basis sets D and E, the E set has an extra inverse bilinear term compared to basis set D but has worse performance overall compared to basis set D. The D basis set also has a lower number of terms compared to basis set E. This shows that the addition of the inverse bilinear term does not improve performance. The same case in observed in the F and G basis sets, where the G basis set has a squared bilinear term but has worse performance and a significantly larger number of terms. Therefore, as the number splits is inconsequential when using the PLS and QPLS models, the number of splits is chosen based on the model performance of the ALAMO model with basis set D. A comparison between the PLS, QLPS and ALAMO D models with 10 splits can now be carried out.

3.3.6 Comparison of PLS, QPLS and ALAMO models

For the comparison of the PLS, QPLS and ALAMO methodologies, the 10th split is chosen for this comparison of models. PLS and QPLS models are marked as P-10 and Q-10 whereas ALAMO models using basis set D in this split are marked with A-D10 from now on. Tables 3.15, 3.16, and 3.17 contain the testing, training and overall performance metrics for the P-10, Q-10, and A-D10 models. The "set" column in the tables represents the set of data in each split in the 10 data splits. For the training set metrics found in table 3.16, the same trend is observed for the relative performance of the models as as for the testing set but with more minor errors and ranges of errors. For example, the model bias values are negligible with magnitudes of 10^{-14} to 10^{-5}. Interestingly, the P-10 and Q-10 models achieve lower overall biases in the training set, contrary to the testing set trend. The R^2 values for the A-D10 models are consistently higher with a range of 0.916-0.923 compared to the P-10 (0.780-0.791) and the Q-10 (0.786-0.799) models. The RMSE values mirror the R^2 trend as the A-D10 models have a range of 0.508 to 0.531 kcal mol^{-1}, the P-10 models have a range of 0.707-0.748 kcal mol^{-1}, and the Q-10 models have a range of 0.680-0.730 kcal mol^{-1}. It is now clear from the superior performance of the A-D10 model shows that the ALAMO methodology produces better models. The overall performance for each model for the 2167 experimental data points is shown in table 3.17. It can be seen that the A-D10 models have the best performance for the R^2 and RMSE metrics for each set, followed by the Q-10 and

the P-10 models. However, there is negligible difference (0.0139 to 0.0353 kcal mol^{-1}) between the Q-10 and P-10 models. From the selection of overall models, the A-D10-10 has the lowest RMSE value of 0.516 kcal mol^{-1} and an R^2 value of 0.921. In the testing set, the ALAMO RMSE values are lowest, followed by the Q-10 and P-10 values. The A-D10 model bias ranges from -0.082 to 0.092 kcal mol^{-1}, which is negligible. In contrast, the P-10 and Q-10 models have bias values that reach 0.178 and 0.159 kcal mol^{-1}. The A-D10 R^2 values range from 0.857-0.914 and are higher than the P-10 and Q-10 models, ranging 0.733 to 0.809. These values show that the ALAMO framework provides excellent performance that outclasses the PLS and QPLS frameworks. The small A-D10 RMSE range of 0.529-0.745 kcal mol^{-1} indicates superior performance relative to the Q-10 and P-10 models (0.593-0.862 and 0.600-0.864 kcal mol^{-1} respectively) for the testing set. Overall, the A-D10 models are the best choice, followed by the Q-10 and then the P-10 models. Set 2 in the A-D10 models has the lowest RMSE value of 0.529 kcal mol^{-1}. This model is marked as A-D10-2.

TABLE 3.15: Table of performance metrics for the 10 $\Delta G^{o,m}_{s,i,j}$ testing sets in 10-fold cross-validation splits for the PLS, QPLS and ALAMO methodologies.

Set	R^2			RMSE / kcal mol^{-1}			Bias / kcal mol^{-1}		
	P-10	Q-10	A-D10	P-10	Q-10	A-D10	P-10	Q-10	A-D10
1	0.775	0.778	0.857	0.695	0.688	0.672	-9.5E-02	-1.2E-01	-5.5E-02
2	0.798	0.805	0.914	0.640	0.619	0.529	1.8E-01	1.6E-01	8.1E-02
3	0.784	0.800	0.891	0.680	0.629	0.593	1.5E-01	1.1E-01	9.2E-02
4	0.797	0.809	0.884	0.680	0.642	0.627	2.0E-02	1.0E-02	5.8E-02
5	0.774	0.775	0.900	0.741	0.738	0.572	-5.6E-02	-4.7E-02	-1.8E-02
6	0.783	0.784	0.866	0.864	0.862	0.745	-1.6E-01	-1.3E-01	-8.2E-02
7	0.784	0.798	0.910	0.833	0.777	0.591	-9.2E-02	-7.8E-02	1.4E-02
8	0.733	0.753	0.864	0.847	0.782	0.666	-2.7E-02	-3.0E-02	-6.8E-02
9	0.787	0.779	0.900	0.747	0.777	0.603	7.3E-02	8.3E-02	-2.6E-03
10	0.796	0.798	0.900	0.600	0.593	0.544	3.5E-03	4.8E-03	-6.1E-02

The PLS model of Borhani et al. (2019) attains an RMSE value of 0.55 kcal mol^{-1} for their testing set of 356 data point, which is significantly lower than the RMSE values seen in the P-10 and Q-10 models found in this study. The difference in RMSE values can be attributed to the use of different training sets. However, the testing set differs in the number of data points (356 in Borhani et al. (2019) to approximately 217). Further, the $\Delta G^{o,m}_{s,i,j}$ database has not only increased in size (1777 in Borhani et al. (2019) to 2167 in this work), but also in the types of molecules and their proportions in the data set. Therefore, the Borhani et al. (2019) model was used to predict the 2167 data points to compare to the newly developed PLS

TABLE 3.16: Table of performance metrics for the 10 $\Delta G^{o,m}_{s,i,j}$ training sets in 10-fold cross-validation splits for the PLS, QPLS and ALAMO methodologies.

Set	R^2			RMSE / kcal mol^{-1}			Bias / kcal mol^{-1}		
	P-10	Q-10	A-D10	P-10	Q-10	A-D10	P-10	Q-10	A-D10
1	0.786	0.795	0.922	0.726	0.696	0.516	2.2E-14	-3.7E-09	-5.7E-07
2	0.780	0.786	0.920	0.744	0.723	0.520	1.1E-14	-4.6E-09	-3.2E-06
3	0.786	0.796	0.922	0.727	0.693	0.515	1.9E-14	-1.2E-09	-8.0E-07
4	0.778	0.783	0.923	0.746	0.730	0.508	1.5E-14	4.4E-12	3.6E-07
5	0.787	0.792	0.917	0.719	0.704	0.531	1.3E-14	6.4E-10	5.3E-06
6	0.786	0.794	0.916	0.707	0.678	0.527	1.9E-14	1.3E-16	-2.6E-06
7	0.785	0.791	0.918	0.709	0.689	0.521	1.5E-14	3.6E-09	-1.4E-05
8	0.791	0.799	0.922	0.707	0.680	0.514	2.1E-14	-2.0E-10	1.0E-07
9	0.786	0.794	0.922	0.718	0.691	0.512	1.7E-14	9.6E-10	3.4E-07
10	0.781	0.788	0.923	0.748	0.724	0.512	1.3E-14	-3.4E-09	-3.1E-06

TABLE 3.17: Table of performance metrics for the 2167 $\Delta G^{o,m}_{s,i,j}$ data points in 10-fold cross-validation splits for the PLS, QPLS and ALAMO methodologies.

Set	R^2			RMSE / kcal mol^{-1}			Bias / kcal mol^{-1}		
	P-10	Q-10	A-D10	P-10	Q-10	A-D10	P-10	Q-10	A-D10
1	0.785	0.793	0.915	0.723	0.695	0.533	-9.5E-03	-1.2E-02	-5.5E-03
2	0.782	0.788	0.919	0.733	0.713	0.521	1.8E-02	1.6E-02	8.1E-03
3	0.785	0.796	0.919	0.722	0.687	0.524	1.5E-02	1.1E-02	9.2E-03
4	0.780	0.786	0.919	0.739	0.721	0.521	2.0E-03	1.0E-03	5.8E-03
5	0.786	0.790	0.915	0.721	0.707	0.535	-5.6E-03	-4.7E-03	-1.8E-03
6	0.785	0.793	0.909	0.723	0.697	0.552	-1.6E-02	-1.3E-02	-8.2E-03
7	0.786	0.793	0.917	0.722	0.698	0.528	-9.2E-03	-7.8E-03	1.4E-03
8	0.786	0.795	0.916	0.721	0.690	0.531	-2.6E-03	-2.9E-03	-6.8E-03
9	0.786	0.792	0.919	0.721	0.700	0.522	7.3E-03	8.3E-03	-2.6E-04
10	0.782	0.789	0.921	0.733	0.711	0.516	3.4E-04	4.8E-04	-6.1E-03

model. A comparison of Borhani et al.'s model and model 10 of the P-10 models (P-10-10) can be found in Table 3.18. The two models were trained and validated on different data sets, and therefore a fair comparison can only be drawn from the overall performance. However, it is useful to see the performance of the models with their original training and testing sets. In the table, the training set RMSE value is 0.52 kcal mol^{-1} whereas the testing set RMSE value is 1.349 kcal mol^{-1} for the Borhani et al. model. The reason for the large RMSE value of Borhani's PLS model is due to new points being added to the training set that the original model was not validated on. The P-10-10 model has a lower RMSE value of 0.600 kcal mol^{-1} for the testing set and an RMSE value of 0.748 kcal mol^{-1} for the training set. The spread of the errors is more even in the P-10-10 model. Further, the overall RMSE value for P-10-10 model is lower than the overall RMSE value from the Borhani PLS model. Therefore, this

section has improved on the PLS methodology of Borhani et al. (2019), solely by using a new training and testing data set.

Table 3.18: Comparison of the Borhani et al. (2019) PLS model and the PLS model developed in this section

	Borhani et al. (2019)		P-10-10	
	Data points	RMSE / kcal mol^{-1}	Data points	RMSE / kcal mol^{-1}
Overall	2167	0.889	2167	0.733
Testing	746	1.349	205	0.600
Training	1421	0.520	1950	0.748

Nait Saidi et al. (2020) also compared their re-optimised Mullins et al. COSMO-SAC model with the PLS model of Borhani et al. (2019) using the data set of 1777 data points found in the latter work. Borhani et al. reported a mean unsigned error (MUE) or absolute average deviation (AAD) of 0.44 kcal mol^{-1} whereas Nait Saidi et al. reported an AAD of 0.37 kcal mol^{-1}. The P-10-10 and Q-10-10 models achieved an AAD value of 0.45 and 0.45 kcal mol^{-1}, respectively. These values similar to the AAD values from Borhani et al. The best model in this section was found to be the ALAMO model with basis set D, using the 10th data split, marked as A-D10-10. This model was also used to calculate the AAD over the 1777 data points and achieved a comparatively lower value of 0.31 kcal mol^{-1}. This result highlights the excellent performance of the data-driven models developed using the ALAMO framework. However, while this result has a lower error value, the difference in performance is 0.06 kcal mol^{-1} compared to the COSMO-SAC model and is essentially negligible. Nait Saidi et al. note that the COSMO-SAC model is used in conjunction with saturated pressures and densities to obtain the free energy of solvation; however, the A-D10-10 model is standalone.

In this section, a range of PLS, QPLS and ALAMO models were developed for the purpose of creating a robust and predictive data-driven model for the prediction of free energies of solvation. A cross-validation study was carried out to assess the optimal proportion of data in the training and testing sets, where the model with the lowest testing RMSE value was chosen as the best model using this methodology. The PLS, and QPLS models are an improvement over the models found in literature; however, the ALAMO model has significantly superior performance to the PLS and QPLS models. This excellent performance is because the ALAMO framework allows for a customised fit for the shape of the data cloud between the response and descriptor variables. However, the number of terms in the PLS and QPLS

models are significantly lower compared to the ALAMO model with 7 terms in each model compared to the 132 terms in the A-D10-10 model. The stark difference in the number of terms is characteristic of the PLS, QPLS and ALAMO approaches. The PLS and QPLS frameworks reduce the number of terms through nondimensionalisation whereas the ALAMO framework has a large number of terms to closely approximate the experimental data. The best model, A-D10-10, was found to outperform the PLS model of Borhani et al. (2019) in AAD by 0.11 kcal mol^{-1} and the re-optimised Mullins et al. COSMO-SAC model of Nait Saidi et al. (2020) with a difference of 0.06 kcal mol^{-1}. The model details can be found in appendix C.2 in table C.5.

3.4 Conclusion

In this chapter, the PLS, QPLS and ALAMO frameworks were used to develop data-driven models by relating the experimental $\Delta G^{o,m}_{s,i,j}$ values to 9 quantum-mechanical solute descriptors and 12 bulk solvent descriptors. Cross-validations studies were carried out to determine the effect of the size of training and testing sets on the predictive capabilities of developed PLS, QPLS and ALAMO models. It was found that there was negligible effects for the PLS and QPLS models; however, there was a more pronounced difference between ALAMO models. The ALAMO models also depended on the choice of basis functions as some were more correlated to the experimental free energy of solvation data. In particular, linear and binomial functions had the best performance among the selection of basis functions. A further study was carried out using different combinations of basis functions to obtain the best set. It was found that a mix of linear, inverse and binomial terms with a training/testing split of 9:1 yielded the best models.

After a PLS, QPLS and a selection of ALAMO models were developed, these models were benchmarked against the experimental free energy of solvation data to derive performance metrics and determine the best data-driven model. The ALAMO framework was not only shown superior to the PLS and QPLS frameworks, but it also provided excellent predictive capability and high customisation for models. The A-D10-10 model achieved an overall RMSE value of 0.516 kcal mol^{-1}, a R^2 value of 0.921 with negligible bias. Thus, a potential user is encouraged to utilise the A-D10-10 model in the prediction of solvation free energies as it has a wide range of applicability (2167 solute/solvent pairs) and is shown to be predictive. In

conclusion, the ALAMO framework significantly improved the performance of a data-driven generalised solvation model.

Chapter 4

The development and testing of hybrid quantum mechanical/data-driven solvation models

4.1 Development of data-driven models using a combined QM and data-driven approach

In the previous chapter, data-driven models were developed using the nonaqueous free energy of solvation database of 2167 data points with the PLS, QPLS and ALAMO methodologies. In section 3.3.6, a comparative study showed the ALAMO models to have superior performance to the PLS and QPLS models, most likely due to a better framework to account for nonlinear behaviour between reponse and descriptor variables. Using the nonaqeous database also showed the best ALAMO models achieved excellent performance with a testing set RMSE value of 0.614 kcal mol^{-1} and an R^2 value of 0.886 on average. Conversely, the training set RMSE values of 0.518 kcal mol^{-1} and an R^2 value of 0.918 on average, where both sets of metrics indicate a robust and predictive model.

The models chosen in section 3.3.6 only differ from the framework of Borhani et al. (2019) in the regression methodology (QPLS and ALAMO) and the sets of data used to train the models. Despite the excellent performance found in that combination, this study aims to improve on the framework itself. Notably, Borhani et al. (2019) proposed and succeeded in developing a generalised approach that combines both the description of a solute and the solvent in the same model through quantum-mechanical (QM) and bulk descriptors, respectively. However, the framework does not consider any interaction between solute and solvent

as the QM descriptors are for solutes in a vacuum. In contrast, the solvent descriptors are for the macroscopic properties of a pure solvent.

In general, in the development of predictive tools not specific to the solvation free energy, any form of data can be used as descriptors to describe any phenomena. However, some descriptors are correlated more closely with a response variable and are therefore preferred as it reduces dependence on experimental data and offers better predictions. Mathematical or physical theory can be used to reduce further the reliance on empirical data, often at the expense of computational power. In the predictive tools discussed previously, such as activity coefficient models (ACMs), or equations of state (EoSs), the molecular interactions are described using specific interaction parameters between functional groups. These interaction parameters are obtained by regressing the underlying physical theory (local composition or intermolecular potential) against experimental data. In contrast, for data-driven methodologies such as PLS, ALAMO and even more complex examples like neural networks, there is no underlying physical theory to account to describe molecular interactions. Therefore, data-driven methodologies utilise variables that are closely or inherently related to such effects. This could be achieved by carrying out some pre-processing of the response variable or including further descriptors. While this offers an avenue for improving the data-driven framework, they come with the challenges.

A descriptor could be required for each unique solute/solvent pair but not many experimental properties of the infinite dilution state exist that are not equivalent to the solvation free energies. Predictive tools such as ACMs, EoSs or QM modelling software can instead supply several properties; however, the specific choice of property or properties is significant, and subsequently, selecting an appropriate tool is difficult. One can also pre-process the descriptor variables through mathematical transformation or by combining two or more descriptors in an algebraic manner, but that would require some supporting theory. Further, some mathematical pre-processing of the variables is already handled by the PLS, QPLS and ALAMO methodologies. In PLS and QPLS, the data are first mean-centred and scaled, then non-dimensionalised to measure the variance. In contrast, in the ALAMO methodology, the descriptor variables undergo a variety of mathematical transformations and the best subset of the transformed variables are selected. This methodology allows for a nonlinear fit of the data space. Interestingly, a possible solution to the challenge of introducing descriptors that relate solute and solvent can be found in the response variable itself, as it is a property related

to the solute and solvent. The free energy of solvation is a comparison of two thermodynamic states, the solute in vacuum and the solute in solvent. By partitioning the free energy of solvation into these effects, it is easy to see how the QM solute descriptors are related to the solutes in a vacuum, whereas the solvent in the latter can be accounted for by the bulk solvent descriptors. However, this is not the only means of partitioning the solvation free energy to account for the interaction between the solute and solvent. The proposed approach involves partitioning the solvation free energy to account for a QM description of the solute-solvent interaction.

4.1.1 Quantum-mechanical models

Following the works of Huron and Claverie (1972; 1974; 1974), the solvation process can conceptually be divided into three steps, namely: cavitation, dispersion-repulsion and electrostatics as seen in equation (4.1).

$$\Delta G_s^{\circ,m} = \Delta G^{cav} + \Delta G^{vw} + \Delta G^{ele} \tag{4.1}$$

This partitioning implies an order in the process of solvation, to describe the move of a solute in the gas phase to the solution phase. First, a cavity forms in the solvent to accommodate the solute molecule, where the energy required to do so is represented by ΔG^{cav}. Next, the dispersion-repulsion (or the van der Waals) forces act between the solute and solvent molecules, represented by ΔG^{vw}. The cavitation and van der Waals effects are typically grouped as the nonelectrostatic contribution. Finally, the charge distribution of the solute is modified in the solvent to account for electrostatic effects, ΔG^{ele}. These effects include the work required to transfer the solute charge distribution from the gas phase to the liquid phase (solvent) and the work done by the solvent to polarise the solute charge distribution. The reader is encouraged to explore this concept further in comprehensive reviews (Orozco and Luque, 2000; Tomasi, Mennucci, and Cammi, 2005b). This partitioning of the solvation free energy for an implicit inclusion of the solute-solvent interaction in our modelling efforts while maintaining the ALAMO data-driven model formulation. Before describing this approach, a brief review of QM methods is presented.

In QM methodologies, the solute molecule is almost always explicitly modelled quantum-mechanically (except for larger molecules such as polymers where only a portion of the solute

is modelled), whereas the solvent may be accounted for differently. QM methodologies can be classified into three categories: (i) explicit, (ii) implicit, and (iii) hybrid. The solvent description can range from many solvent molecules to an infinite dielectric continuum (Jalan et al., 2010). Figure 4.1 contains representations of the different approaches to modelling in solvation models. For cases (a, b, c), the solute and solvent molecules are treated explicitly, where several solvent molecules surround the solute molecule. Case (a) is a pure QM approach, where every molecule is modelled using quantum-mechanics and (b) is a purely classical approach where molecular mechanics forcefields are used to describe the solute and solvent. Cases (a) and (b) belong to the explicit category but use different methods to calculate the free energy of solvation. In case (c), it is a hybrid approach where the solute molecule is modelled using QM, and the solvent using a forcefield. Cases (d, e, f) have a continuum which represents the solvent, which means a homogeneous background described by some solvent properties, influencing any solute and solvent molecules. This representation of the solvent is called implicit solvation. Case (d) is an extension of (c), and case (e) is an extension of (a) where an extra layer of solvent is added to represent the solvent effects throughout all space. Cases (d) and (e) are both hybrid solvation methodologies. Finally, case (f) represents an implicit solvation methodology where the solvent is modelled solely as a continuum through a infinite isotropic dielectric, and sometimes through other solvent properties.

Nicholls et al. (2008) demonstrated that implicit methods are less expensive than explicit methods at the cost of slightly lower accuracy. This tradeoff is essential as it reduces the amount of time required to develop and reduces the complexity of the hybrid QM/data-driven model. The reader is encouraged to read further into the different types of QM models if interested.

To calculate the free energy of solvation as the difference in the liquid phase free energy and the gas-phase free energy of solute i in solvent j using an implicit solvation model, $\Delta G^{\circ,m}_{s,i,j}$ is defined (Ho, Klamt, and Coote, 2010) as

$$\Delta G^{\circ,m}_{s,i,j} = (E^{el,L}_{i,j} + G^{NES,L}_{i}) - E^{el,IG}_{i} \tag{4.2}$$

where $E^{el,L}_{i}$ is the liquid phase electronic energy at $T = 0$ K and $c^{\circ,L} = 1$ mol dm^{-3}, $E^{el,IG}_{i}$ is the gas phase electronic energy at $T = 0$ K and $p^{\circ,IG} = 1$ atm and $G^{NES,L}_{i}$ is the non-electrostatic term for species i at $T = 289$ K. The reference standard state for the gas phase

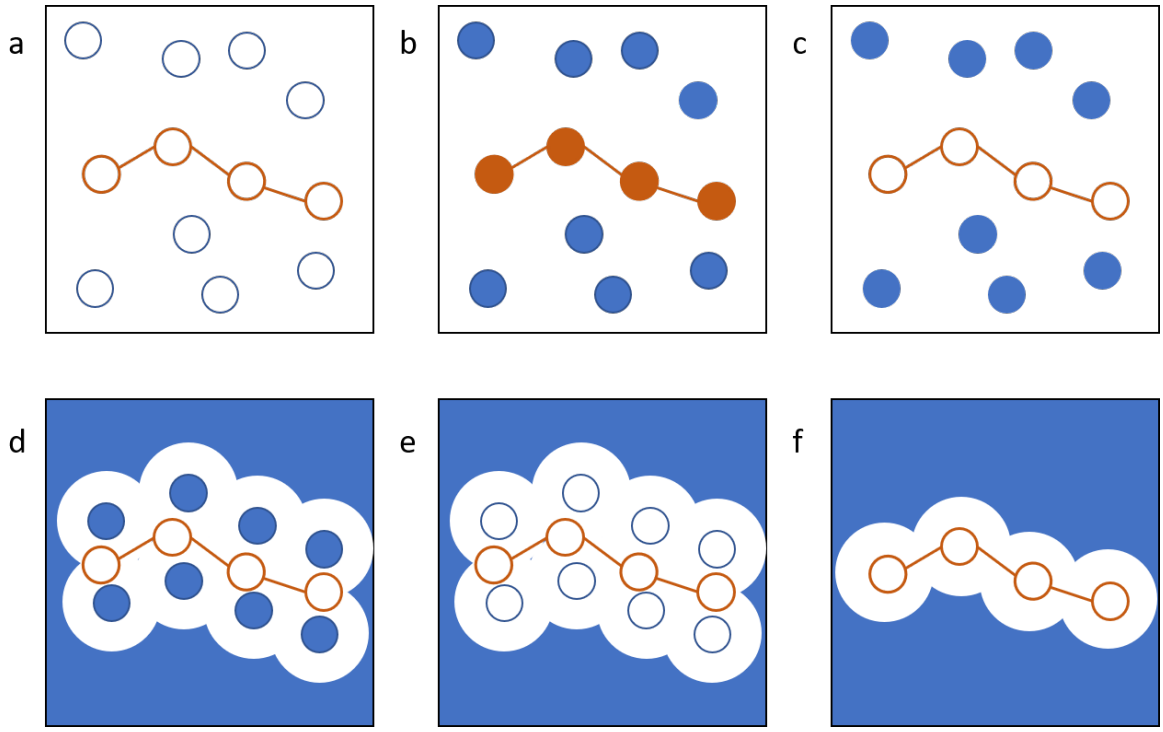

FIGURE 4.1: A representation of the different types of solvation models, from discrete to continuum models. The solute and solvent are represented using the blue colours and orange colours, respectively. A filled background represents a solvent continuum. Empty-faced dots represent molecules or atoms calculated using quantum-mechanics, and coloured dots by force field. Case (a) is the purely QM approach where both the solute and solvent are explicitly modelled using QM approaches, whereas case (b) represents the purely classical approach where the solute and solvent molecules are described using molecular mechanics forcefields. Case (c) is the hybrid case where the solute molecule is modelled through QM and the solvent is modelled using a forcefield. Case (d) is an extension of case (c) where the there is a surrounding continuum to represent the solvent effects throughout all space. Case (c) is an extension of case (a) is the same manner where the solvent effects throughout all space is modelled using a continuum. Case (f) is the implicit solvation approach where the solvent is treated as a continuum and the solute is modelled using QM.

electronic energy, $E_i^{el,IG}$, can be converted using the logarithm term found in equation (2.2). This equation is closely related to equations (2.2) and (4.1). In comparison to equation (2.2), both equations show that the free energy of solvation $\Delta G_{s,i,j}^{\circ,m}$ is the difference in chemical potentials of the solute in the liquid and gas phase. Similarly to equation (4.1), the solvation free energy is partitioned into electrostatic effects where $\Delta G^{el} = \Delta E^{el} = E^{el,L} - E^{el,IG}$, and nonelectrostatic effects wehre $\underline{G}_i^{NES,L} = \Delta G^{cav} + \Delta G^{vw}$. While QM is directly used to calculate the gas phase electronic energy, an implicit solvation model is required to calculate the liquid phase electronic energy. Thus, as long as the electronic energies for the liquid and gas phases can be calculated, the nonelectrostatic effects can be accounted for using the 21 descriptor variables outlined in this study. A brief review of implicit solvation models is provided in this section to explore this contribution further.

Implicit solvation models

Implicit solvation models or continuum solvation models consider the solute at the QM level. The solvent is a uniform polarisable medium, where the solute is placed in a suitably shaped cavity. Jensen et al. (2007) outline five aspects that differentiate continuum solvation models. These are: (i) the definition of the size and shape of the cavity, (ii) the calculation of the cavity/dispersion contribution, (iii) the description of the dielectric medium, (iv) the definition of the charge distribution, (v) the modelling of the solute either through QM or MM.

In initial solvation models, electrostatic effects were considered by characterising the solvent as a scalar, static dielectric constant and assuming a linear response from the solvent to a changing electric field (Born, 1920; Kirkwood, 1934; Born, 1936). While these models provide satisfactory descriptions of the bulk phase, they can not represent the solvents at the first-solvation-shell level. The behaviour of the solvent molecules that directly surround the solute molecule is different from those at the bulk level. This difference in behaviour further occurs at the second-solvation-shell and so on until the molecules are sufficiently far away. Therefore, Lee and Richards (1971) proposed the solvent-accessible surface (SASA) approach to tackle this difference. The SASA is defined as the area traced out by a centre of a sphere, whose radius is half the effective width of the first solvent shell, rolling over the solvent surface. Figure 4.1 contains an example of the SASA approach in case (f), where the white area surrounding the solute molecule represents the SASA. The SASA concept is widely used in solvation models to this day, where the calculated area differs from model to model (Cramer and Truhlar, 1999).

When referring to solute polarisation in solvation, a crucial concept is the reaction field, which is the electric field exerted by the solvent on the solute. Including the reaction field in the solute Hamiltonian alter the electric moments of solute, resulting in further changes in the polarisation of the solvent, which then iterates to self-consistency. This iterative process is referred to as the self-consistent reaction field (SCRF).

In the polarisable continuum model (PCM) (or DPCM for dielectric PCM), the SCRF is combined with a boundary element problem with apparent surface charges (ASC) (Miertuš, Scrocco, and Tomasi, 1981). In the PCM model, the van der Waals cavity is considered as one formed by overlapping spheres with empirically determined radii. The ASC in the integral equation formalism version of the PCM model (IEF-PCM) is calculated using the

electrostatic potential instead of the normal component of the electric field found in DPCM (Cancès, Mennucci, and Tomasi, 1997).

The COSMO model is another popular entry in QM solvation models as it describes the surrounding medium as a perfect conductor, meaning the dielectric constant is infinite. This description is used to obtain the ASCs, which are then scaled down by a function of the actual dielectric constant of the solvent (Klamt and Schüürmann, 1993). The COSMO model was then extended to COSMO-RS, where RS stands for real solvents. The post-processing step of the COSMO-RS model involves processing the COSMO calculations for the solutes and solvents with the statistical thermodynamical treatment of interacting "surface segments" (Klamt, 2011).

The final example is the series of SMx models developed by Cramer and Truhlar (Cramer and Truhlar, 1991; Cramer and Truhlar, 2006; Cramer and Truhlar, 2008; Marenich, Cramer, and Truhlar, 2009), who developed a series of generalised Born type solvation models under the name SMx, where x represents the different versions. A feature of the SMx series of models is the partitioning of the free energy of solvation into two parts, instead of the three parts found in equation (4.1). The first of the two terms account for shorter range polarisation and nonelectrostatic effects, like cavitation (C), dispersion (D) and changes to the solvent structure (S) caused by the solute, denoted as ΔG^{CDS}, and the second term accounts for changes in electronic structure of the solute and for changes in nuclear coordinates, denoted as ΔE^{ele}.

The two models most directly relevant to this work are the IEF-PCM and SMD solvation models as they follow the partitioning of the solvation free energy found in equation (4.2) more closely. These models will be reviewed and will be shown how they factor into the hybrid model framework in the following sections.

IEF-PCM

While the IEF-PCM model (Cancès, Mennucci, and Tomasi, 1997; Mennucci, Cancès, and Tomasi, 1997; Mennucci and Tomasi, 1997; Tomasi, Mennucci, and Cancès, 1999) is but one of the various PCM formulations to exist, the effects of the solvent phase on the solute molecules are calculated using these implicit solvation PCM models. The liquid phase free energy G_i^L, which is equivalent to the liquid phase electronic energy added to the nonelectrostatic free energy term, is defined in equation (4.3).

$$G_i^L = E_i^{el,L} + G_i^{NES,L} = E_i^{el,L} + G_i^{cav,L} + G_i^{dis,L} + G_i^{rep,L} \tag{4.3}$$

where $E_i^{el,L}$ is the liquid phase electronic energy and $G_i^{cav,L}$ is the free energy associated with cavity formation for the solute molecule in the solvent phase. $G_i^{dsp,L}$ is the free energy related to dispersion interactions in the liquid phase and $G_i^{rep,L}$ is the free energy associated with repulsive forces in the liquid phase. In the IEF-PCM model, the calculation of the cavitation formation free energy is based on a method developed by Pierotti (Pierotti, 1963; Pierotti, 1976). The dispersion and repulsion free energies are based on empirical dispersion and repulsion atom-atom potentials (Floris and Tomasi, 1989; Tomasi, Mennucci, and Cammi, 2005a). Further, the IEF-PCM model has been implemented in Gaussian03 (Frisch et al., 2004) and Gausian09 (Frisch et al., 2016), where Gaussian09 is the model suite of choice.

In the IEF-PCM model (Tomasi, Mennucci, and Cancès, 1999), the molar free energy of solvation is calculated via equation (4.2) where the $G^{NES,L}$ is equivalent to the cavitation, dispersion and repulsion free energies used in the IEF-PCM model.

Solvation Model based on Density (SMD)

The SMD model is introduced in this work as the framework serves as an inspiration for the hybrid QM/data-driven model. The partitioning of the electrostatic and nonelectrostatic effects seen in equation (4.2) is also used in the the SMD model; where $G^{NES,L}$ is defined as the cavity, dispersion and solvent effects term $G^{CDS,L}$. It is described as a 'universal' model due to the applicability of the model as long as several parameters are known (Marenich, Cramer, and Truhlar, 2009). This means that with the corresponding parameters for the solute and solvent, any combination of solute and solvent can be modelled. In the model, the required solvent parameters include the dielectric constant (ε), refractive index (n), bulk surface tension (γ), Abraham's hydrogen bond acidity (α) and basicity (β), aromaticity (ϕ), electronegative halogenicity (ψ) (Marenich, Cramer, and Truhlar, 2009). The acidity and basicity parameters are commonly referred to as the solvatochromic parameters (Abraham, 1993). Further, a series of intrinsic atomic Coulomb radii that are optimised for the IEF-PCM algorithm is also used (Marenich, Cramer, and Truhlar, 2009).

The framework for developing this model involves the calculation of the $\Delta\underline{E}^{ele}$ term by

using the IEF-PCM algorithm but turning off the nonelectrostatic contributions such as dispersion, repulsion and cavitation. Therefore, the structure of a solute molecule is optimised in the liquid phase solely using the dielectric. A calculation of the gas phase total energy is also performed to obtain $\Delta\underline{E}^{el}$. The resulting electrostatic term is subtracted from the experimental solvation free energies to derive the nonelectrostatic term. The nonelectrostatic term, $\underline{G}^{CDS,L}$, is defined as follows:

$$G^{CDS,L} = \sum_{j=1}^{N_A} \sigma_j A_j(\boldsymbol{r}, r_{Z_j}) + r_s + \sigma^{[M]} \sum_{j=1}^{N_A} A_j(\boldsymbol{r}, r_{Z_j} + r_s) \tag{4.4}$$

where N_A is the number of atoms, σ_j is the surface tension of atom j in cal mol^{-1} Å^{-2} and σ^M is the molecular surface tension cal mol^{-1} Å^{-2}. A_j is the SASA of atom j in Å^2, which depends on the geometry of solute $\boldsymbol{r}$, the atomic van der Waals radii r_{z_j}, and the solvent radius r_s. Through the atomic and molecular surface tensions, the nonelectrostatic term is regressed against a series of solvatochromic and solvent parameters. However, it is not a straightforward regression using the $\underline{G}^{CDS,L}$ as the response variable and the parameters as the descriptors. The atomic surface tension σ_j is calculated as follows:

$$\sigma_j = \tilde{\sigma}_{Z_j} + \sum_{j'=1}^{N_A} \tilde{\sigma}_{Z_j Z_{j'}} T_j(Z_{j'}, r_{jj'}) \tag{4.5}$$

where $\tilde{\sigma}_{Z_j}$ is a parameter that depends on the atomic number of atom j, $\tilde{\sigma}_{Z_j Z_{j'}}$ is a parameter that depends on the atomic number of both atoms j and j', and $T_j(Z_{j'}, r_{jj'})$ is a geometry-dependent function, a cutoff tanh. The atomic and molecular surface tensions depend on bulk solvent properties such as the refractive index, acidity and basicity. The parameters $\tilde{\sigma}_{Z_j}$ and $\tilde{\sigma}_{Z_j Z_{j'}}$ are defined as follows:

$$\tilde{\sigma}_\theta = \tilde{\sigma}_\theta^{[n_D]} n_D + \tilde{\sigma}_\theta^{[\alpha]} \alpha + \tilde{\sigma}_\theta^{[\beta]} \beta \tag{4.6}$$

where the subscript θ stands for either Z_j or $Z_j Z_{j'}$, and the corresponding solvent properties are as follows: n_D is the refractive index at 293 K, α is Abraham's hydrogen bond acidity, and β is Abraham's hydrogen bond basicity. The coefficients $\tilde{\sigma}^{[n_D]}$, $\tilde{\sigma}^{[\alpha]}$, $\tilde{\sigma}^{[\beta]}$ are regressed parameters that depend on θ. The molecular surface tension σ_M is expressed as follows:

$$\sigma_M = \tilde{\sigma}_\theta^{[\gamma]} \gamma + \tilde{\sigma}_\theta^{[\varphi^2]} \varphi^2 + \tilde{\sigma}_\theta^{[\psi^2]} \psi^2 + \tilde{\sigma}_\theta^{[\beta^2]} \beta^2 \tag{4.7}$$

where γ is the macroscopic surface tension at 298.15 K in units of cal mol^{-1} Å^{-2}, φ is the aromaticity and ψ is the halogenicity. The coefficients $\tilde{\sigma}^{[\gamma]}$, $\tilde{\sigma}^{[\varphi^2]}$, and $\tilde{\sigma}^{[\psi^2]}$ are regressed parameters independent of θ.

Equations (4.4) shows that the solvatochromic parameters are intrinsically tied to the structure of the solute through the atomic surface tension σ_j and the molecular surface tension $\sigma^{[M]}$. The subsequent equations 4.5), (4.6), and (4.7 further show the coefficients which are regressed alongside the solvatochromic parameters.

4.1.2 Proposed methodology for the hybrid QM/data-driven approach

In the development of the data-driven models for the free energy of solvation, a direct limitation of said models is the lack of information about how the solute interacts with the solvent. This is due to the lack of underlying physical theory and solute/solvent specific descriptors. The SMD formulation discussed in the previous section is a QM methodology that provides a detailed description of the solute structure and accounts for the solute/solvent interaction by regressing solvent descriptors to said solute structure. The proposed approach is to adopt aspects of the SMD framework to create a hybrid QM/data-driven model. However, this is not easily achieved while using the purely data-driven methodologies as these methodologies lack the underlying physical theory to relate the solute structure to the molecular descriptors in exactly the same way. The simplest way to introduce information about solute/solvent interaction is to partition the solvation free energy into an electrostatic contribution and nonelectrostatic contribution. The electrostatic contribution would be obtained from QM calculations and inherently provide information about the solute in vacuum and in solvent to the data-driven model. The resulting nonelectrostatic contribution can then be related to the solute/solvent properties using the PLS, QPLS or ALAMO methodologies. Therefore the nonelectrostatic contribution in the proposed approach is expressed below:

$$G_{i,j}^{CDS} = \Delta G_{s,i,j}^{o,m} - \Delta E_{i,j}^{el} = f(\beta, \boldsymbol{X}) \tag{4.8}$$

where $\boldsymbol{X}$ is the vector of descriptor variables related to $G_{i,j}^{CDS}$ through a mathematical function. The data-driven aspect of the model is achieved through the nonelectrostatic contribution, whereas the QM aspect of the model is achieved through the electrostatic contribution. This results in a model with a detailed description of the solute and a data-driven formulation

proven to have excellent predictive capabilities. Some drawbacks of this approach are the fact that there still no cross terms between solute and solvent descriptors, or how the relationship between solute structure and descriptor variables is not well defined. However, this provides an excellent starting point for the development of a QM/data-driven model. Further modifications to the partitioning or the inclusion of solute structure coordinates can be explored in future work.

The proposed approach will model the electrostatic contribution using the IEF-PCM model in a similar manner as the SMD model. The modelling is carried out in the Gaussian09 model suite (Frisch et al., 2016). The remaining nonelectrostatic contribution $G_{i,j}^{CDS}$ is then regressed against the set of 21 descriptor variables using the ALAMO methodology. The proposed approach is outlined in figure 4.2

In Figure 4.2, there are a series of steps that are followed to develop the hybrid QM/data-driven models. In Step 1, experimental free energy of solvation values, $\Delta G_{s,i,j}^{\circ,m,exp}$, are compiled and a range of electronic energies in the vacuum phase, $E_{i,j}^{el,IG}$, and the solvent phase, $E_{i,j}^{el,L}(\varepsilon)$, are calculated. The ε term is the dielectric constant of a solvent. The calculation of the electronic energies requires a separate algorithm which will be discussed later in this section. In Step 2, the experimental solvation free energies and the electronic energies are collated and manipulated according to equation (4.8) to obtain a database of $G_{i,j}^{CDS}$ values in Step 3. Step 3 also involves matching corresponding solute and solvent descriptors to the $G_{i,j}^{CDS}$ values. These $G_{i,j}^{CDS}$ values are the response variable values for the development of the hybrid models. In Step 4, data-driven model frameworks are selected as candidates for producing the hybrid models. In this step, any mathematical framework can theoretically be used; however, in this thesis, only the PLS, QPLS and ALAMO frameworks are considered. In Step 5, the $G_{i,j}^{CDS}$ database is used in conjunction with a model framework to produce a data-driven model with a separate array of predicted $G_{i,j}^{CDS,pred}$ values. In the case of the PLS and QPLS models, one model is produced per training/testing set; however, in the case of the ALAMO framework, several models can be produced depending on the combinations of basis functions. The process of developing these models results in a series of predictive models for the free energy of solvation. In Step 6, the predicted $G_{i,j}^{CDS,pred}$ values are calculated and converted into predicted free energies of solvation, $\Delta G_{s,i,j}^{\circ,m,pred}$, using the electronic energies of the vacuum $E_{i,j}^{el,IG}$, and solvent phases, $E_{i,j}^{el,L}(\varepsilon)$, obtained earlier. Therefore, for each hybrid model developed in step 7, a corresponding array of $\Delta G_{s,i,j}^{\circ,m,pred}$ values can now be utilised

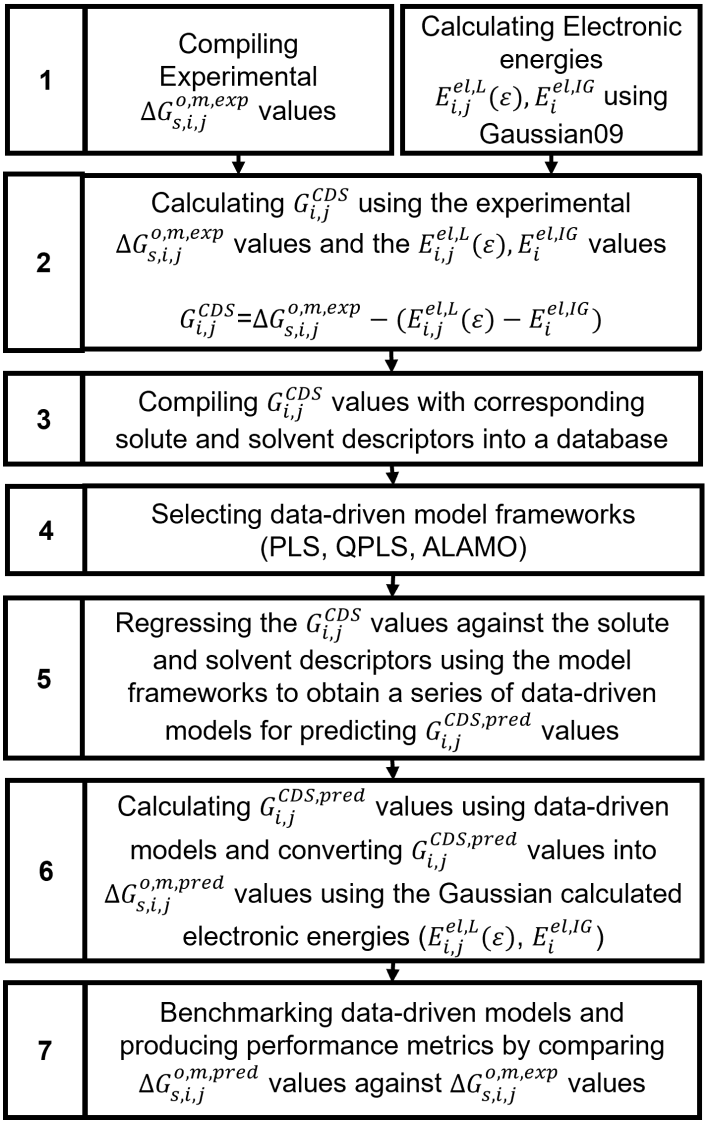

FIGURE 4.2: A description of the workflow in the hybrid quantum-mechanical/data-driven model

in the benchmarking of the models by comparing against $\Delta G_{s,i,j}^{o,m,exp}$ values. Thus, in Step 7, performance metrics can be extracted for each model and the best model can be quantitatively selected.

Therefore, it is necessary to first elaborate on how to calculate the electronic energies of the solute in the vacuum and solvent phases as this is a critical step in the development of the hybrid QM/data-driven models.

4.1.3 The calculation of electronic energies in the vacuum and solvent phases

In the previous section, the workflow for developing the hybrid QM/data-driven models was discussed. The first and crucial step of the workflow involved the calculation of electronic

energies which would provide a mathematical model with some basis in the physical world. The electronic energies would also provide a connection between the solute and solvent phase as the current set of solute and solvent descriptors used in earlier data-driven models did not yield any information about interactions between the solute and solvent. Thus, the crux of the proposed hybrid models relies on the calculation of electronic energies.

The calculation of the electrostatic energies of the solute molecule in either the vacuum or liquid phase involves a series of important choices that determines the accuracy of the calculation. A perfect example is the choice of electronic structure method as this directly influences the degree of information each atom possesses. Electronic structure methods with a higher amount of detail about the electron structure or more complex functions significantly increase the computational burden, leading to a choice in computational time versus accuracy. Another example is the choice of solute conformation in the vacuum or liquid phase when obtaining the electronic energies as the use of different conformations can lead to prominent changes in the electronic energy. Finally, there are many calculation options when using the Gaussian09 (Frisch et al., 2016) model suite which may or may not ensure true convergence in an optimisation. The following section will discuss problems and offer the most suitable choice for this work.

Calculation of electronic energies

The calculation of the electronic energies of the solute molecule in the liquid and vacuum phases are carried out using the Gaussian09 (Frisch et al., 2016) suite of models. In the SMD framework (Marenich, Cramer, and Truhlar, 2009), the electronic energies of the solute molecule in both phases were calculated using the IEF-PCM/Gaussian03 model and six electronic structure methods (ESMs), namely M05-2X/MIDI!6D, M05-2X/6-31G*, M05-2X/6-31+G**, M05-2X/cc-pVTZ, B3LYP/6-31G*, and HF/6-31G* (Hehre et al., 1986; Zhao, Schultz, and Truhlar, 2006; Easton et al., 1996; Li, Cramer, and Truhlar, 1998; Dunning, 1989; Becke, 1988; Lee, Yang, and Parr, 1988; Stephen et al., 1994). These six ESMs range from very descriptive to relatively simple methods for describing the solute molecule. In the SMD framework, the electronic energies derived from these methods are averaged to produce a model for a broader range of ESMs. This methodology can be helpful for applications such as screening possible solvents using a less descriptive ESM to save on computational resources before considering a more descriptive one. However, in this work, only one ESM is used to

save time and serve as a starting point for the development process. The middle-range ESM, X3LYP functional (Xu et al., 2005) and the 6-31G(d,p) basis set are chosen as it offers a good balance between accuracy and computational cost.

In Marenich et al. (2009), Guthrie et al. (2009) and Ribeiro et al. (2010), the electronic energy of the liquid phase is determined using a structure optimised in a vacuum. In doing this, energetic changes associated with a potential shift in the solute conformation when solvated are ignored. Struebing (2011) argues it may be a reasonable tradeoff in terms of computational resources when performing a vast number of calculations associated with a large number of solvents. In contrast, in another approach (in examples such as Stanescu & Achenie (2006) and Ashcraft et al. (2007)), the optimised vacuum conformation is re-optimised when solvated to account for the effects of the electronic field on the solute conformation through a change in its charge distribution (which may be minor or significant). Nicholls et al. (2008) demonstrated that the lowest energy in the gas phase does not necessarily correspond to the lowest-energy solute structure in the liquid phase. This results in a difference in the solvation free energies, and for example, the resulting difference may have considerable effects on the prediction of reaction rates (Struebing, 2011; Siougkrou, 2014; Grant, 2019).

Struebing (2011) presented some examples of the effect on the solvation free energy in his thesis. The first example involved predicting solvation free energies of pyridine in cyclohexane, benzene, chloroform, aniline, and water. In the example, the effect of using vacuum-optimised and "liquid-phase" optimised conformations are compared and the results are contrasted against a set of experimental data. He found average absolute errors of 0.04 kcal mol^{-1} for both approaches, which this work has shown to be negligible. However, in another example, Ribeiro et al. (2010) calculated the solvation free energies of two glycerol conformations, of which both were vacuum-optimised. The solvation free energies of conformers I and II were -11.71 kcal mol^{-1} and -12.32 kcal mol^{-1}, respectively. Struebing calculated the solvation free energies for the two glycerol conformations but in both the vacuum and liquid phases to observe if there was a substantial change. For conformer I, he found solvation free energies of -11.68 and -10.89 kcal mol^{-1} for the vacuum-optimised and "liquid-phase" optimised structures, respectively. For conformer II, he found solvation free energies of -12.30 and -11.68 kcal mol^{-1}, respectively. Therefore, the maximum possible difference in the solvation free energy can be seen by comparing the energies between Struebing's "liquid-phase" optimised and Ribeiro's conformations, which are -0.82 and -0.64 kcal mol^{-1}, respectively. Unfortunately, there was

no experimental value to confirm which approach was the most accurate. In this work, the second approach is preferred as a solvent can affect the conformations of solutes. However, it must be noted that the majority of solutes found in the database are small molecules that have limited degrees of freedom. Thus, a large difference in energy is not expected.

In Gaussian09 (Frisch et al., 2016), several user options influence the degree of convergence in the optimisation of a solute structure. These involve the calculation of force constants (Maximum Force ("Max Force"), Root Mean Square Force ("RMS Force")) throughout the structure or the density of the mesh used for the calculation of charge distribution across the molecule ("grid"). Further, a series of optimisation options such as the tightness of the convergence criteria on the forces acting on the molecule and the displacement of atoms in the molecule and whether or not a symmetry of the molecule is considered in the calculation. These options are integral to determining the quality of the electronic energy obtained at the end of the calculation. There are also options such as the maximum number of steps "maxsteps" (which determines the degree of atom displacement) and the maximum number of cycles "maxcycle" (which determines the maximum number of calculations per step). Therefore, the optimisation process works by calculating the forces and displacements of the atoms in the molecule and checking if the values are within the convergence criterion, which ranges from a forgiving "normal" to a constrictive "vtight". If not, the shape of the structure is altered until the criterion is met. This results in a molecule that has the minimum amount of forces acting upon it and therefore, is the lowest energy structure. The electronic energies are obtained from the Gaussian output file when the solute structure meets the convergence criteria, where the final structure is taken if the convergence criteria are not met.

The calculation of these electronic energies is an integral part of the hybrid models, and therefore, a careful approach is adopted to maximise the calculation quality. This approach is achieved by imposing the strictest optimisation convergence criteria on the solute structures in the vacuum phase and then the solvent phase. The first of these options includes the keyword "calcall", which calculates the force constants acting on each atom at every iteration. The "nosymm" option is included to ensure all calculations are carried out with no modes of symmetry, which may increase the number of computational steps significantly in larger molecules; however, most of the molecules considered in this study are small molecules. As such, there are no molecules that either reach polymer size or to the size of drug molecules, with some of the larger molecules only reaching carbon numbers of 16. As mentioned earlier,

the mesh density is critical as it determines the resolution used for the calculation. For the mesh density, the "integral=grid=ultrafine" option is the densest. The "vtight" convergence criteria are employed as it is the most constrictive and will ensure the lowest energy structure will be found. Further, all solute molecule optimisations do not converge, some molecules are small and the process of resolving constrictive convergence criteria with fewer degrees of freedom along the molecule may pose a challenge. The challenge comes in the form of atom displacements which oscillate close to the optimal structure each iteration. Thus, these non-convergence cases are discussed later in this section. Finally, the initial "maxstep" value is set to 20 to allow for larger atom displacements to speed up the optimisation. The larger atom displacements allow for more distinct structures to be evaluated. Thus, the process on how to obtain the lowest energy structures in Gaussian09 has been outlined by selecting the most constrictive criteria, plus the highest resolution for the calculation.

In Figure 4.2, it was shown that the calculation of electronic energies was a part of the first step. The first reason is that these electronic energies are required in the calculation of the $G_{i,j}^{CDS}$ term. In contrast, the second step is because, in the $\Delta G_{s,i,j}^{\circ,m,exp}$ database, there are several solute/solvent pairs that cannot be modelled inside the Gaussian09 IEF-PCM implementation. This deficiency is due to a lack of some dielectric constants for the solvent phase. The solvents could have been included manually; however, it was decided that they would be excluded from the database instead. Therefore, when comparing the $\Delta G_{s,i,j}^{\circ,m,exp}$ and a potential $G_{i,j}^{CDS}$ database, the latter database has a total number of 2047 data points in comparison to the 2167 data points in the $\Delta G_{s,i,j}^{\circ,m,exp}$ database.

Therefore, it is now crucial to discuss the algorithm in which the range of solute structures are submitted to Gaussian and checked for optimal structures. There are three separate algorithms; one for the calculation of electronic energies in the vacuum phase, another for the calculation of electronic energies in the solvent phase and the final is required for checking if the solute structures have been optimised. The vacuum phase algorithm is depicted in Figure 4.3.

The vacuum phase algorithm consists of several steps and a loop to calculate optimal solute structures and electronic energies. Step SV1 involves creating F input files of solute molecules in the vacuum phase. The input files follow the Gaussian09 standard with the calculation options mentioned above. "SVL" stands for "solute vacuum loop", where the F input files are iterated through the steps in the dotted box with an iteration counter, ITER,

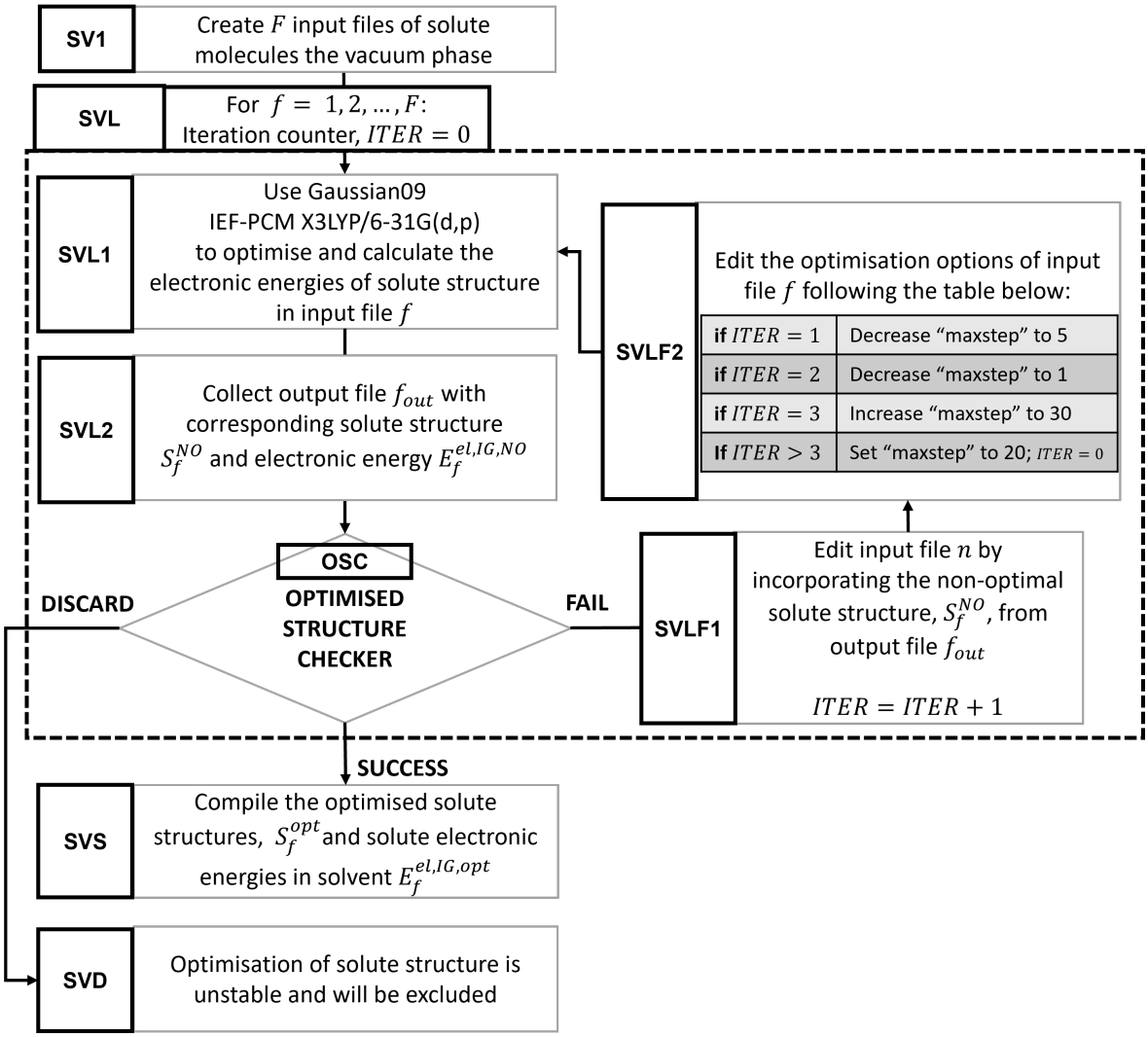

FIGURE 4.3: Algorithm for the calculation of optimised solute structures and electronic energies in the vacuum phase.

starting at 0 for each file. Therefore, each input file will be processed through an optimised structure checker, OSC, and steps SVL1, SVL2, SVLF1, and SVLF2. In Step SVL1, an input file f is submitted to Gaussian09 using the IEF-PCM model with the X3LYP/6-31G(d,p) level of theory to optimise the solute structure in the vacuum phase. The output file is then collected in Step SVL2 along with the corresponding solute structure, S_f^{NO} and electronic energy $E_f^{el,IG,NO}$, where f corresponds to the input file number and NO corresponds to a non-optimal candidate. The output file is then submitted to the OSC, which determines whether Gaussian succeeded or failed in determining the optimal solute structure. The algorithm for the OSC is detailed in Figure 4.5 and is discussed later in this section. In the case that a non-optimal structure is found, a new input file will be generated using the non-optimal structure S_f^{NO} to provide a closer starting point to the optimal. The iteration counter, ITER

will increase by one and serve as a flag to effect further changes. In Step SVLF2, the number of iterations determines what optimisation options are edited in the input file. At a counter of 1, the "maxstep" size is decreased to 5 and then 1 at a counter of 2. The decrease in step size is to initially evaluate more distinct structures when at a larger "maxstep" size and then refine any candidate structures. If no optimal structure is found within 3 iterations, the "maxstep" size is increased to 30 to search for other candidates. The process is then restarted by resetting the counter and the "maxstep" size to 20. After the input file is altered in step SVLF2, it is resubmitted to SVL1. Therefore, once an optimal structure is found, the optimised structure is compiled among other successful solute structures and the electronic energies are tabulated in step SVS. In a rare case where the optimisation never converges, the input file f is discarded in Step SVD and the solute will be excluded from the potential $G_{i,j}^{CDS}$ database. The optimised structures compiled in Step SVS are then passed on to the solvent phase algorithm to calculate the optimised solute structures and their corresponding electronic energies, $E_g^{el,L}$, as seen in Figure 4.4.

The solvent phase algorithm for calculating electronic energies of a solute molecule in solvent follows the exact same process as the vacuum phase algorithm with the addition of a few steps. The first step, SS1, involves compiling P unique solute/solvent pairs from the $\Delta G_{s,i,j}^{\circ,m,exp}$ database. In the SSP loop, the solute/solvent pairs are iterated through to check if they can be modelled by Gaussian09. To clarify, the pair cannot be modelled if the solvent is missing from Gaussian09 and will be excluded from the potential $G_{i,j}^{CDS}$ database. Otherwise, the pair p is compiled among a new list of G pairs to a new batch of G input files which use the F optimised vacuum phase solute structures as starting points in a corresponding solvent j according to the pair g. The main Gaussian loop is then exactly the same as the vacuum phase algorithm with the exception of there being a solvent dielectric field surrounding the solute molecules. If an optimal structure is found, the results are then compiled and the solute electronic energies in the solvent phase, $E_g^{el,L,opt}$, are extracted into an array. In the case an optimisation does not converge, it is excluded from the $G_{i,j}^{CDS}$ database. Therefore, by utilising the vacuum phase and solvent phase algorithms, two arrays of electronic energies corresponding to the vacuum phase and solvent phases can be obtained. Finally, the algorithm for the optimised structure checker is found in Figure 4.5.

The optimised structure checker algorithm, or OSC, is used to check if the convergence criteria of the Gaussian calculation is satisfied or if the electronic energy of the solute structure

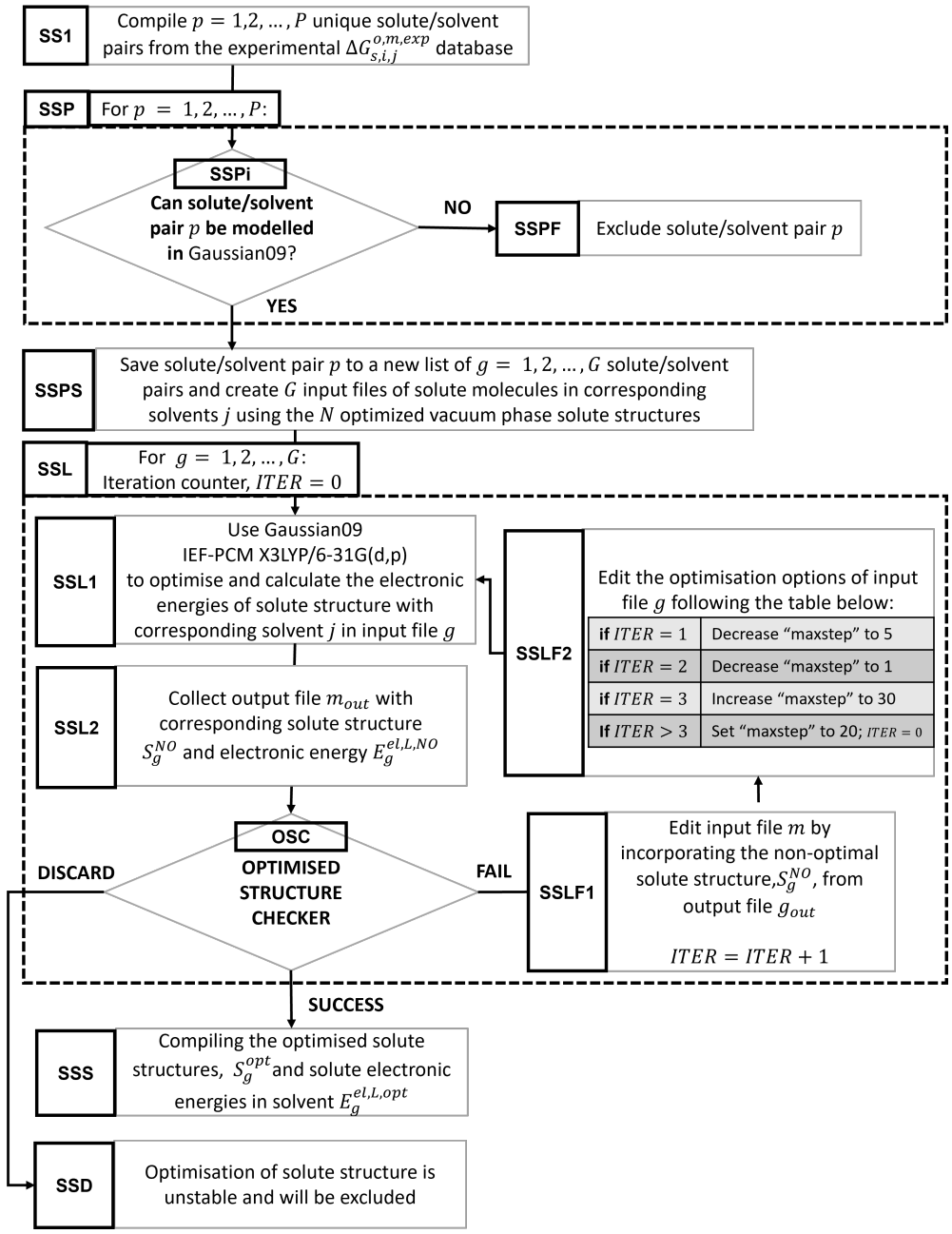

FIGURE 4.4: Algorithm for the calculation of optimised solute structures and electronic energies in the solvent phase.

has plateaued to 9 significant figures. In Step OSC1, the output files from either the vacuum or solvent phase algorithms are collected. In Step OSC2, the output files are scanned to collect convergence data from candidate structures which include the "Max Force", "RMS Force", "Max Displacement", and "RMS Displacement". The latter two data types are the maximum displacement of an atom and the root mean square displacement of an atom. The corresponding threshold values are $2 * 10^{-6}$, $1 * 10^{-6}$, $6 * 10^{-6}$ and $4 * 10^{-6}$. In most cases, the values of the convergence data for solute structures either satisfy the criteria or are very

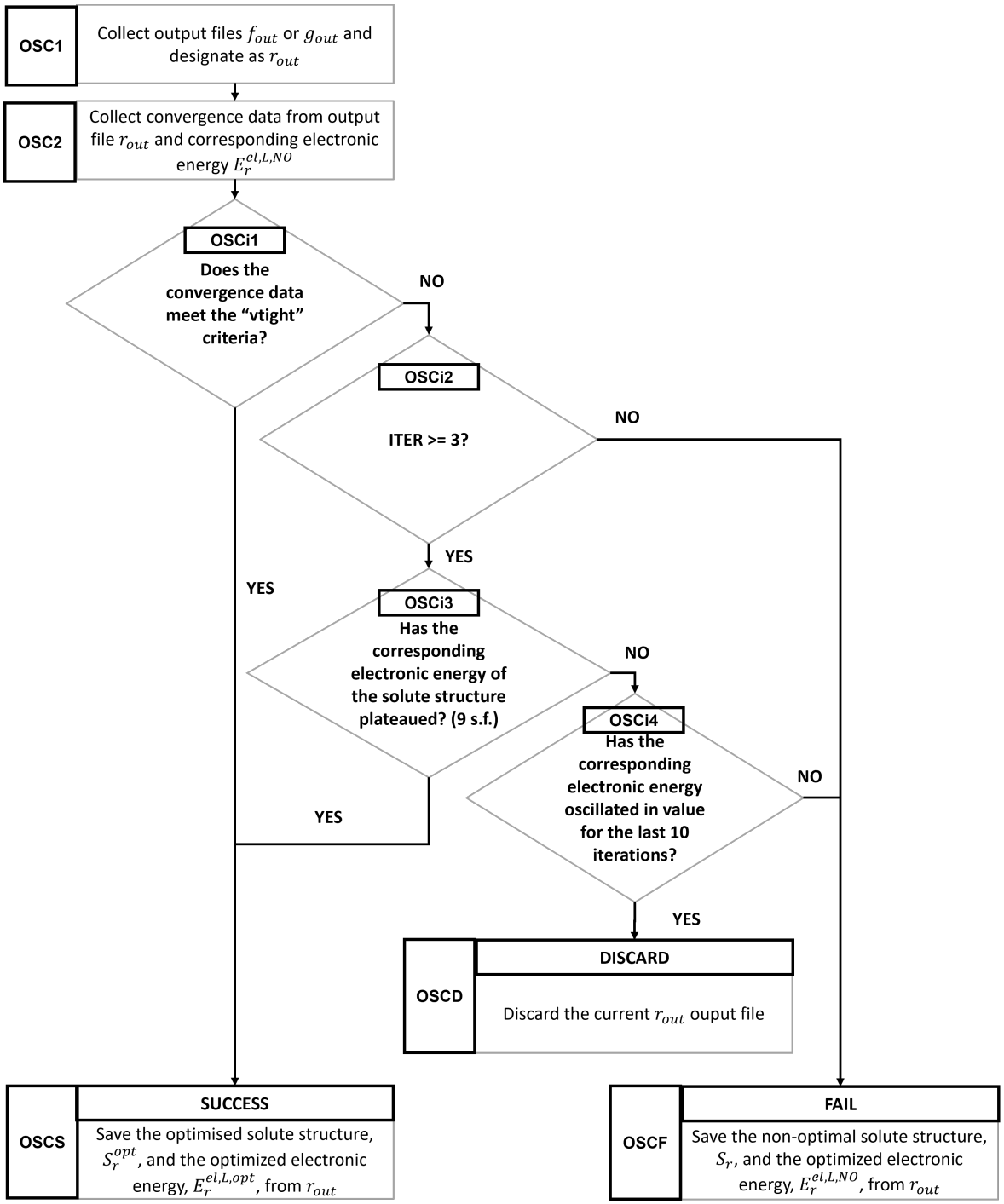

FIGURE 4.5: Algorithm for the optimal structure checker, which is used for determining optimal solute structures.

close in terms of value. The conditional OSCi1 checks for solute structures that satisfy the criteria. If the criteria is met, the solute structure is immediately designated an optimal structure in Step OSCS, where both the optimised structure S_r^{opt} and the electronic energy, $E_r^{el,L,opt}$ (where r is either f or g) is collected. Otherwise, the output file is passed to the OSCi2 conditional to check if the structure has been optimised more than three times. If

not, the output file is designated as non-optimal and will be optimised again. This is to allow enough time for Gaussian to search for an optimal structure. In the conditional OSCi3, the solute structure will have been optimised multiple times and the electronic energies of previous candidates will be compared against the current iteration. If the electronic energy has remained constant to 9 significant figures for at least 3 iterations, the solute structure is considered to be optimal regardless if the convergence criteria are met. This is because most of the molecules considered in this thesis are small molecules with little degrees of freedom. This means that these molecules have few optimal candidates. Therefore, the optimal candidate identified by Gaussian is considered to be the optimal structure. In OSCi4, the case where both the convergence criteria are not met and the electronic energy does not converge is considered. If the number of iterations is less than 10, the solute structure is resubmitted to Gaussian; otherwise, it is excluded from the new $G_{i,j}^{CDS}$ database.

These algorithms are used to calculate the electronic energies of viable solute/solvent pairs from the 2167 data points in the $\Delta G_{s,i,j}^{\circ,m,exp}$ database resulting in 2047 data points, which contains 224 solute molecules and 126 solvent molecules, in the new $G_{i,j}^{CDS}$ database. The new database is now ready to be used alongside the 21 solute and solvent descriptors in the ALAMO framework. The following section will section will focus on the development of the hybrid QM/data-driven ALAMO models.

4.1.4 Development of data-driven models using the $G_{i,j}^{CDS}$ database

In the previous sections, the workflow on how to derive a hybrid QM/data-driven model was proposed. The workflow detailed the process on how to derive a new $G_{i,j}^{CDS}$ which inherently contained a detailed QM description of the solute in the vacuum and solvent phase. In Figure 4.2, steps 1, 2 and 3 are now satisfied and a data-driven model framework can now be selected. In the previous chapter, the ALAMO models was shown to have superior performance compared to the PLS and QPLS data-driven models. In the previous chapter, several ALAMO basis sets (combinations of basis functions found in Table 3.13) were considered in the development of the ALAMO models and will be utilised again in this section. The first aim of this section is to develop a hybrid QM/data-driven model with $G_{i,j}^{CDS}$ being the target variable, the 21 solute/solvent quantities as the descriptors in the ALAMO framework. This process will encompass Steps 5, 6, and 7 or Figure 4.2. The second aim of this section is to confirm

if the new hybrid approach outperforms the models found in the previous chapter such as A-D10-10. However, this would be an unfair comparison as the models will have been developed on different training/testing sets. Therefore, the approaches can be directly compared if the same 2047 data points in the $G_{i,j}^{CDS}$ is used to develop a new $\Delta G_{s,i,j}^{\circ,m}$ data-driven model by using the corresponding $\Delta G_{s,i,j}^{\circ,m}$ data points. This approach allows for a direct comparison between the direct data-driven approach and the proposed hybrid QM/data-driven approach. Any resulting models will then be compared against once another using the same data and benchmarked using performance metrics.

The correlation study found in section 3.3.3 is repeated here as the total number of solutes and solvents decrease to 224 and 126, respectively. Further, a new target variable was introduced so it will be interesting to investigate for any new effects. Further a cross-validation study is carried out on the PLS, QPLS and ALAMO methodologies, similarly to section 3.3.2, first using the $G_{i,j}^{CDS}$ database and then the reduced $\Delta G_{s,i,j}^{\circ,m}$ database. From this point forward, new models developed using the $G_{i,j}^{CDS}$ database will be designated a "H" prefix to signify the hybrid model where as the reduced $\Delta G_{s,i,j}^{\circ,m}$ models will be designated a "R" prefix to signify the reduced models. A collection of these models will be referred to "HX" and "RX" models, respectively.

Correlation between the target variable $G_{i,j}^{CDS}$ and the solute/solvent descriptors using the $G_{i,j}^{CDS}$ database

In section 3.3.3, k-fold cross-validation was used to divide the experimental database into a range of splits to investigate the relationship between the distribution of solute and solvent molecules and the correlation of solute and solvent descriptors with the free energy of solvation. It was found that the distribution of molecules had a negligible effect on the correlations and the target variable $G_{i,j}^{CDS}$ showed greater correlation to the solute properties than to the solvent properties. The new $G_{i,j}^{CDS}$ is a modified subset of the database used in section 3.3.3, with 224 solute molecules and 126 solvent molecules. This represents a stark decrease from the original nonaqueous database. Further, the modified subset of the database also has an alternative reduced $\Delta G_{s,i,j}^{\circ,m}$ data set. The correlation study is repeated using the $G_{i,j}^{CDS}$ data set to observe if any changes when a new target variable is introduced. The data are divided accordingly into training and testing sets following splits of 2, 3, 5, 10, 15, and 20. The average

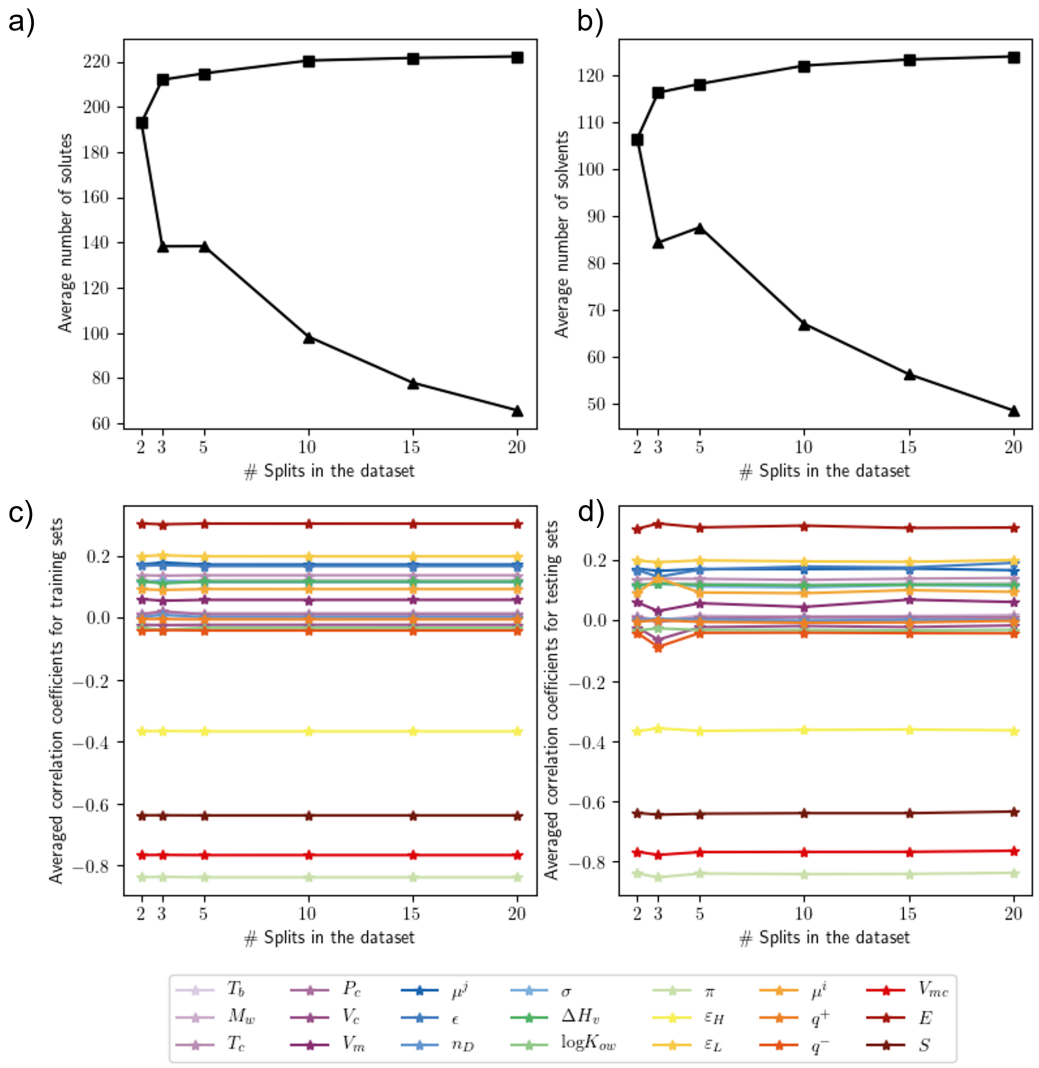

FIGURE 4.6: Plots of the average number of solutes (a) and solvents (b), and the averaged correlation coefficients of the 21 descriptor variables with the target variable $G_{i,j}^{CDS}$ for both training (c) and testing (d) sets across the number of splits. The square and triangle markers in the upper plots are the training and testing data, respectively.

numbers of solutes and solvents are shown, along with the averaged correlation coefficients across the splits in Figure 4.6.

In figure 4.6, the average number of solutes and solvents in the training sets diverge, with the solutes and solvents in the testing set converging to 60 and 50 at the 20th split. Two key observations can be extracted from the average correlation coefficient plots. First, the number of solutes and solvents do not affect the correlation coefficients, confirming that the experimental data or target variable data is well distributed (as the data can be sampled in any way and not affect the relationship between the target and descriptor variables). Next, in

comparison to the correlation coefficients found in figures 3.3 and 4.10, the difference in solute and solvent descriptors is more apparent. The polarizability (π), van der Waal's volume (V_{mc}), total entropy (S), and energy of the highest occupied molecular orbital (ε_H), and the total energy (E) have significantly larger correlation coefficient values of -0.84, -0.76, -0.63, -0.36, and 0.31, respectively; in comparison to the other descriptor variables which now range from -0.05 to 0.2. These correlation coefficients are also notably larger than the ones found in figure 3.3. These results suggest that the partitioning of the electrostatic and nonelectrostatic terms has further shifted the correlation to the target variable in favour of the solute descriptors.

Cross-validation using the $G_{i,j}^{CDS}$ database

Figures 4.7, 4.8 and 4.9 contain the cross-validation results for the PLS, QPLS and ALAMO models using the testing data sets for the hybrid models. The cross-validation process utilises the same training/testing data sets found in the correlation study in section 4.1.4. For the PLS models in figure 4.7, split 3 has the best RMSE performance of 0.657 kcal mol^{-1} compared to the other splits which have RMSE values of approximately 0.690 to 0.700 kcal mol^{-1}. The R^2 show the same trend, with split 3 at 0.826 kcal mol^{-1}, and the rest at 0.803-0.810 kcal mol^{-1}. The bias values for all the splits are also negligible. The same type of performance is observed in the QPLS model, with split 3 having an RMSE value of 0.652 kcal mol^{-1}, and the other splits having a range of 0.670-0.682 kcal mol^{-1}. The R^2 values range from 0.813 to 0.827 kcal mol^{-1}, and the bias values are also negligible. In terms of model complexity, the PLS and QPLS methodologies are simply vector products between coefficients and descriptors, and are therefore simple linear models.

The basis sets used in the ALAMO models for the cross-validation study can be found in Table 3.13. These basis sets were previously used in the development of the original data-driven free energy of oslvation models. The RMSE values show the A, B, and C models generally have a consistent performance of around 0.650 to 0.700 kcal mol $^{-1}$, whereas the D, E, and F models have RMSE values that on average sit 0.100 kcal mol^{-1} higher than A, B, and C. Basis sets E and G achieve the lowest RMSE value of 0.470 kcal mol^{-1}, where basis set E has consistent performance across the splits, with the second-lowest RMSE values of 0.480 kcal mol^{-1} at split 15. In contrast, basis set G has more erratic performance with a significant decrease from 0.620 to 0.500 kcal mol^{-1} for 10 to 15 sets, and a substantial increase from 0.500 to 0.750 kcal mol^{-1}. The basis set trends are reflected in the R^2 plot as the highest value

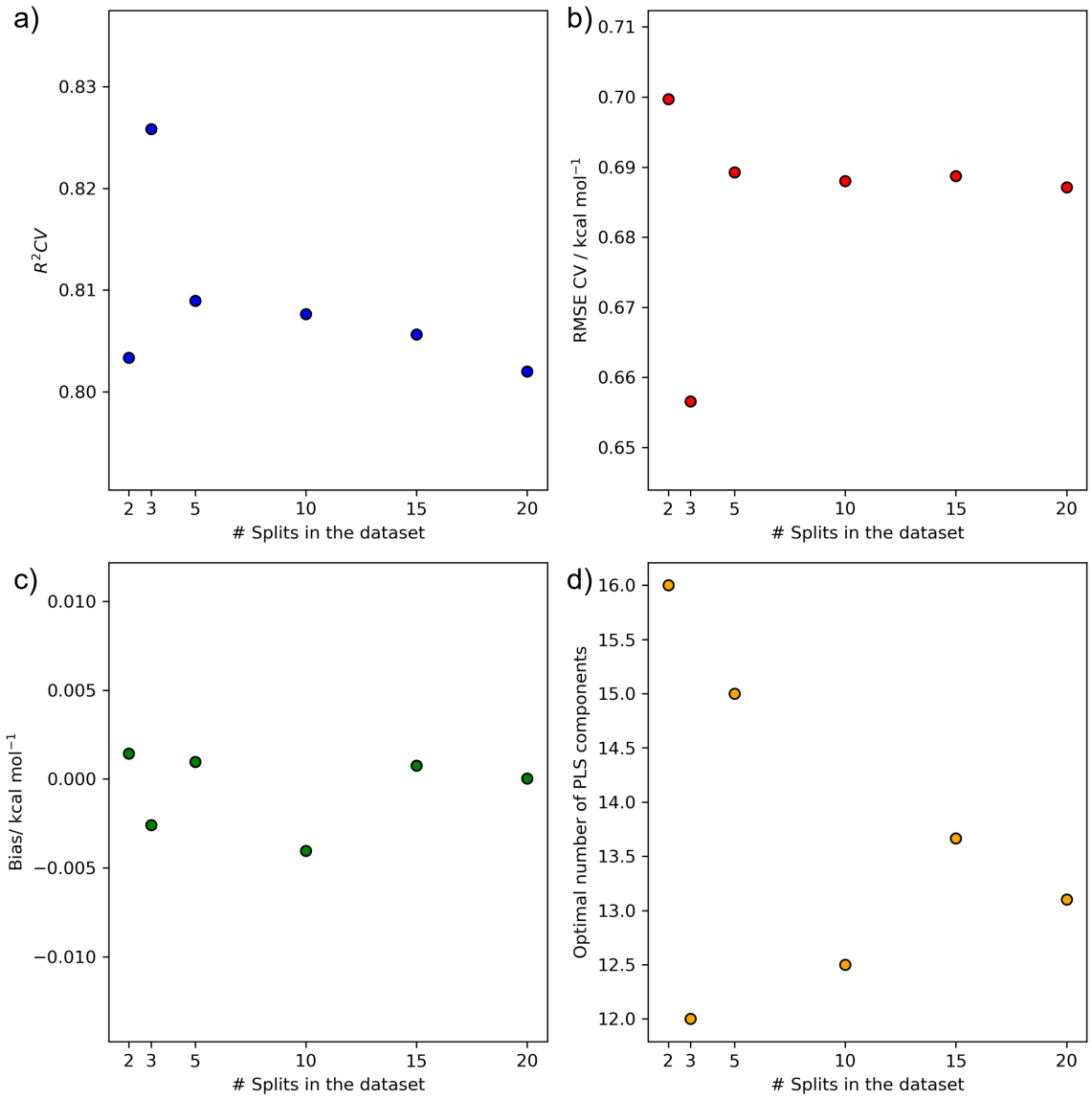

FIGURE 4.7: k-fold cross-validation results for the PLS methodology using the corresponding training/testing sets for each split from the $G_{i,j}^{CDS}$ data set with the performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d).

reaches 0.93 with a lower bound of 0.85, which shows excellent predictive capability. The bias values generally remain in the $\pm$ 0.01 kcal mol^{-1} in most cases, which is negligible. Further, split 3 has a significantly better performance compared to the other splits. Basis set E is chosen as the representative model over basis set G as the resulting models have consistent performance across the splits. Similarly to the cross-validation study carried out in 3.3.2, the PLS and QPLS methodologies, which are inherently linear methodologies, are outclassed by the ALAMO models whose basis sets contain both linear and bilinear terms.

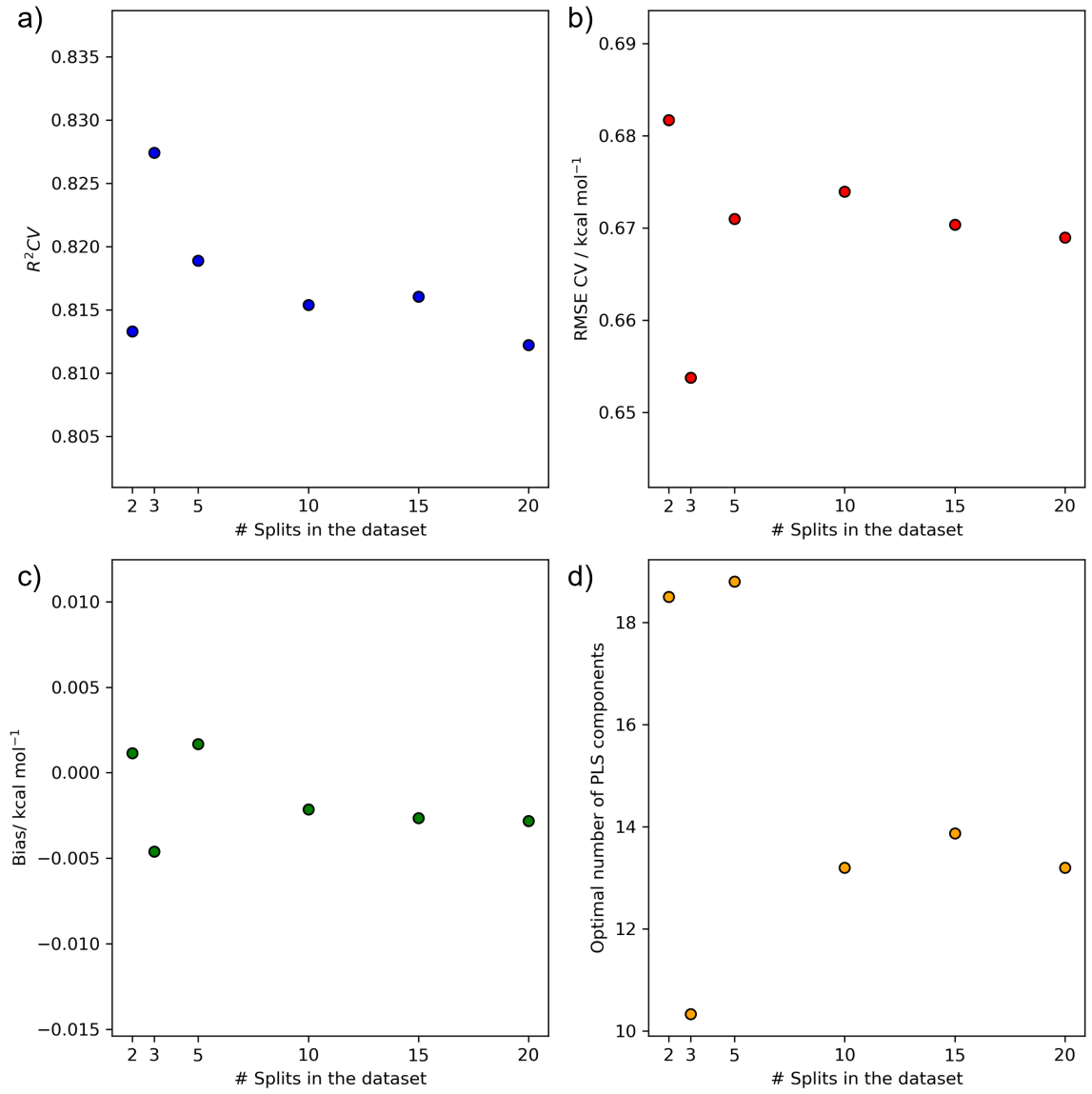

FIGURE 4.8: k-fold cross-validation results for the QPLS methodology using the corresponding training/testing sets for each split from the $G_{i,j}^{CDS}$ data set with the performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d).

Correlation between the target variable $\Delta G_{s,i,j}^{\circ,m}$ and the solute/solvent descriptors using the reduced $\Delta G_{s,i,j}^{\circ,m}$ database

In sections 3.3.3 and 4.1.4, correlation studies were carried out to investigate the relationships between the average number of solutes and solvents and the correlation coefficients of solute/solvent descriptor variables with the target variables $\Delta G_{s,i,j}^{\circ,m}$ and $G_{i,j}^{CDS}$. In this section, the same splits and training/testing data sets as the ones found in section 4.1.4. In comparison to the correlation study of the hybrid model, the reduced $\Delta G_{s,i,j}^{\circ,m}$ data set in this study will help investigate the effects of lowering the molecular diversity. The results of the

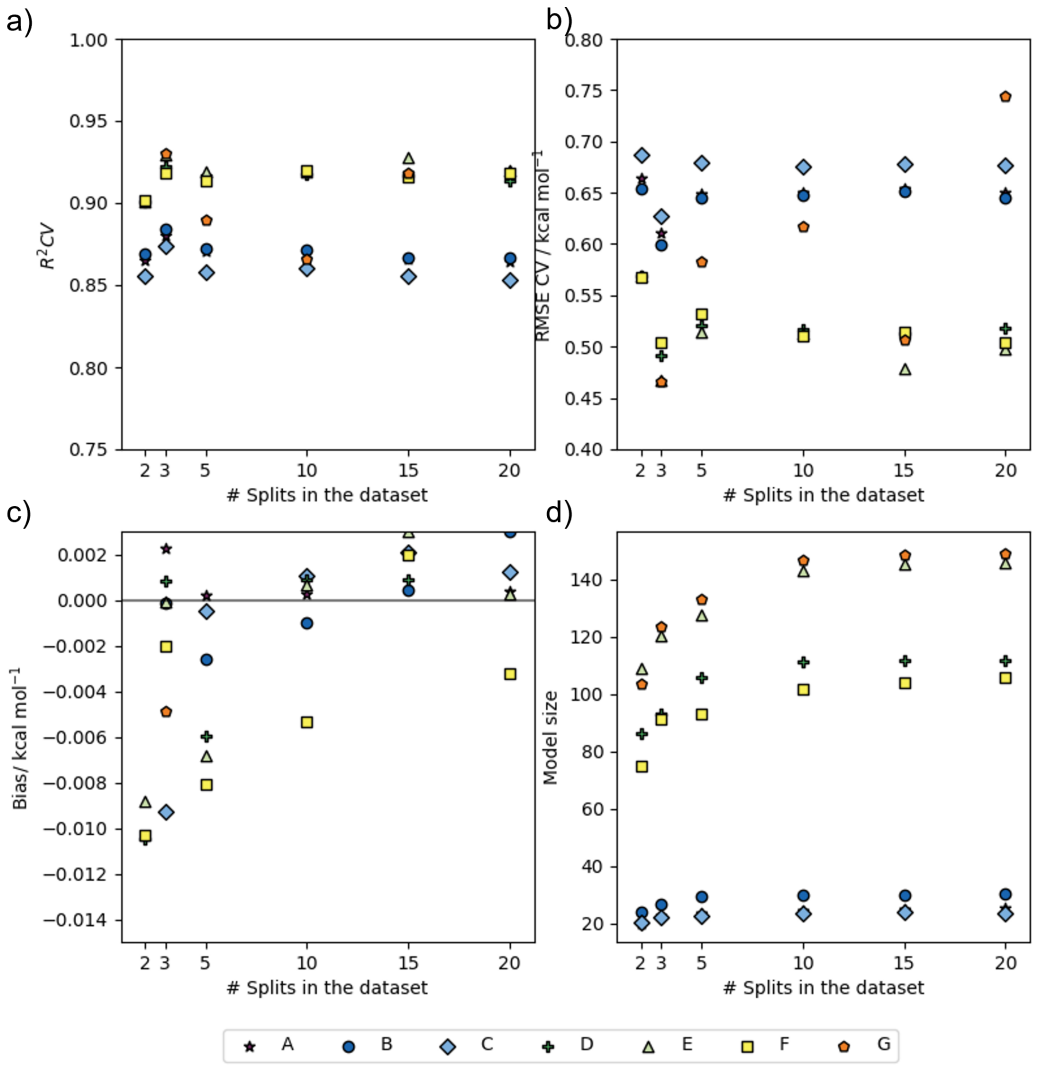

FIGURE 4.9: k-fold cross-validation results for the ALAMO methodology using the corresponding training/testing sets for each split from the $G_{i,j}^{CDS}$ data set with the performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d).

correlation study can be found in Figure 4.10.

In Figure 4.10, the same findings as the ones in section 3.3.3 are found. The square and triangle markers represent the training and testing sets, respectively. The average numbers of solutes and solvents in the training and testing data sets diverge as the number of splits increases. Similarly to Figure 3.3, the average correlation coefficients do not vary with the average numbers of solutes and solvents. The molecular volume, V_m, is the only descriptor to undergo a significant change in value in the training and testing sets. Further, the octanol/water partition coefficient, $\log K_{ow}$, and the solvent dipole moment, μ^j are the only solvent descriptors that have absolute coefficient values around 0.2 or larger. The rest of the

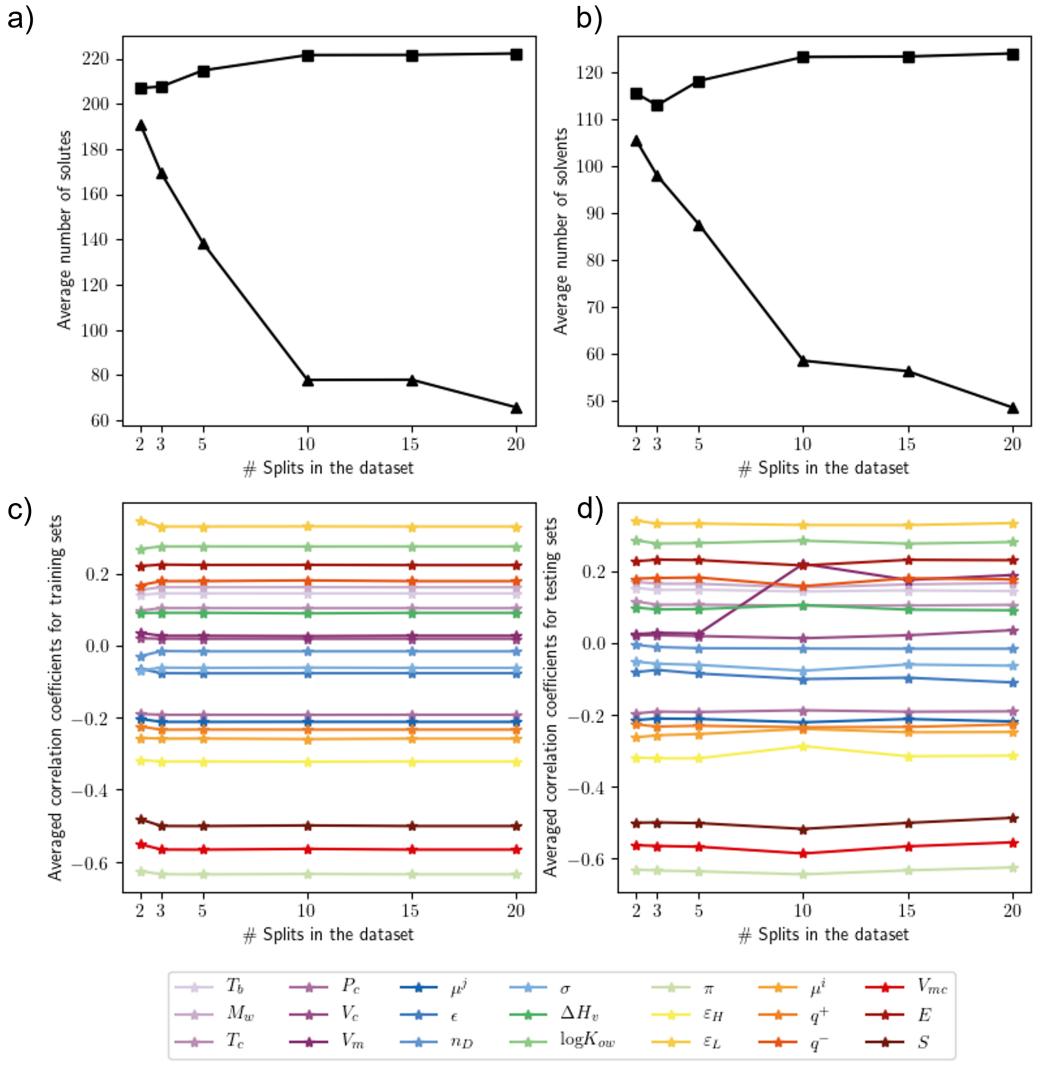

FIGURE 4.10: Plots of the average number of solutes (a) and solvents (b), and the averaged correlation coefficients of the 21 descriptor variables with the target variable $\Delta G^{\circ,m}_{s,i,j}$ for both training (c) and testing (d) sets across the number of splits. The square and triangle markers in the upper plots are the training and testing data, respectively.

solvent descriptor coefficients reside around -0.1 to 0.2, indicating low correlation with the free energy of solvation. The average coefficient values do not differ significantly from the values found in figure 3.3 and show that the distribution of solute and solvents do not affect the correlation between the descriptor and target variables.

Cross-validation using the reduced $\Delta G^{\circ,m}_{s,i,j}$ database

In Figures 4.11, 4.12 and 4.13, the cross-validation results for the PLS, QPLS and ALAMO methodologies using the reduced $\Delta G^{\circ,m}_{s,i,j}$ database can be found. The same data sets found

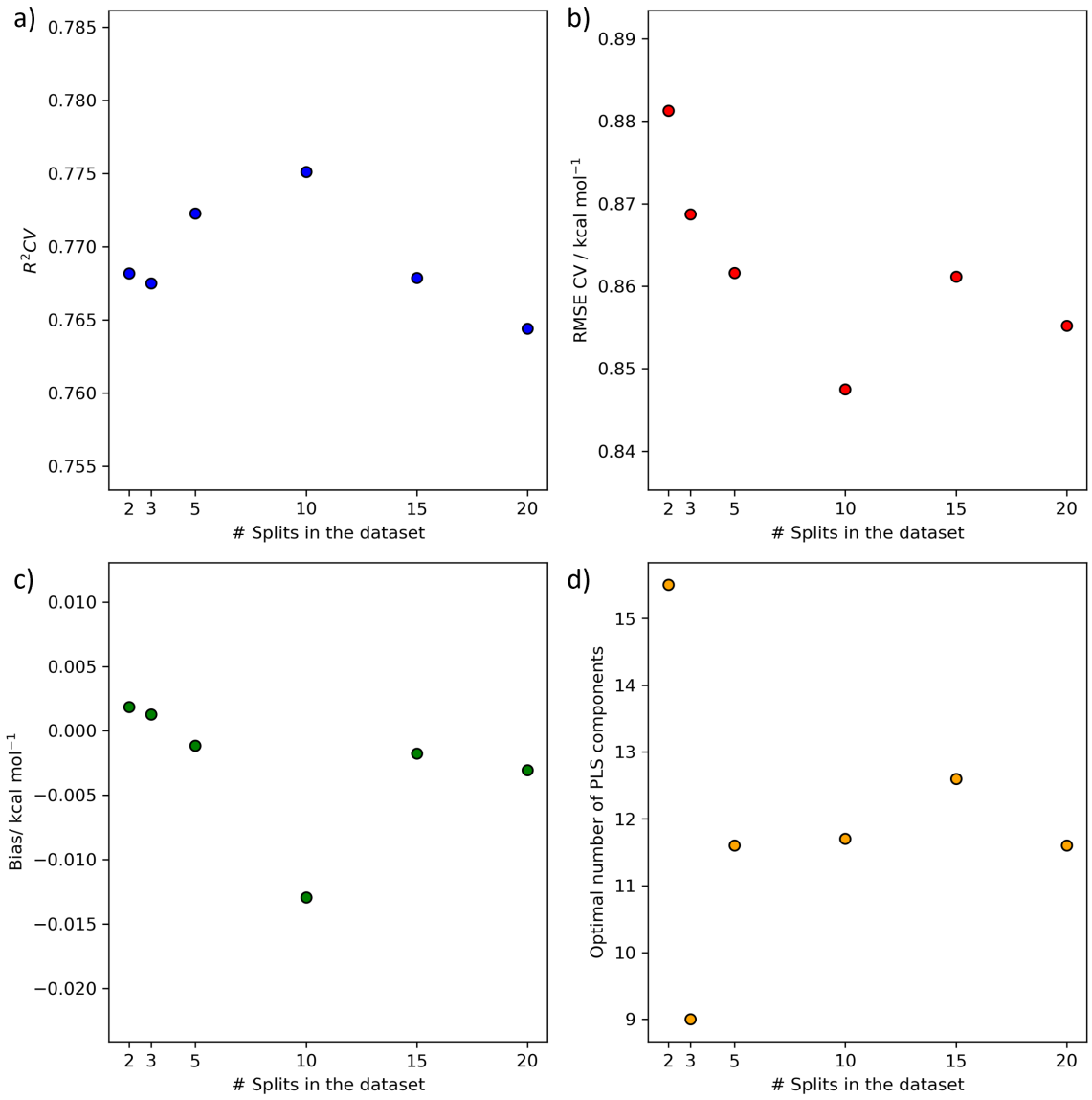

FIGURE 4.11: k-fold cross-validation results for the PLS methodology using training/testing sets for each split from the $\Delta G^{\circ,m}_{s,i,j}$ with performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d)

in the correlation study in section 4.1.4 are also utilised here. This means the same data sets have been used in the cross-validation of the hybrid models from section 4.1.4. The RMSE values for the PLS results range from 0.847-0.881 kcal mol^{-1} and QPLS from 0.825-0.864 kcal mol^{-1}. When compared to the results seen in section 3.3.2, the results are effectively the same. The R^2 values reach 0.781 for PLS and 0.791 for QPLS; the bias values remain in magnitudes of -3, indicating there is negligible bias in the models.

In figure 4.13, the ALAMO A, B, and C basis sets have similar performance to the PLS and QPLS models. This similarity is mainly due to the models being dominated by linear

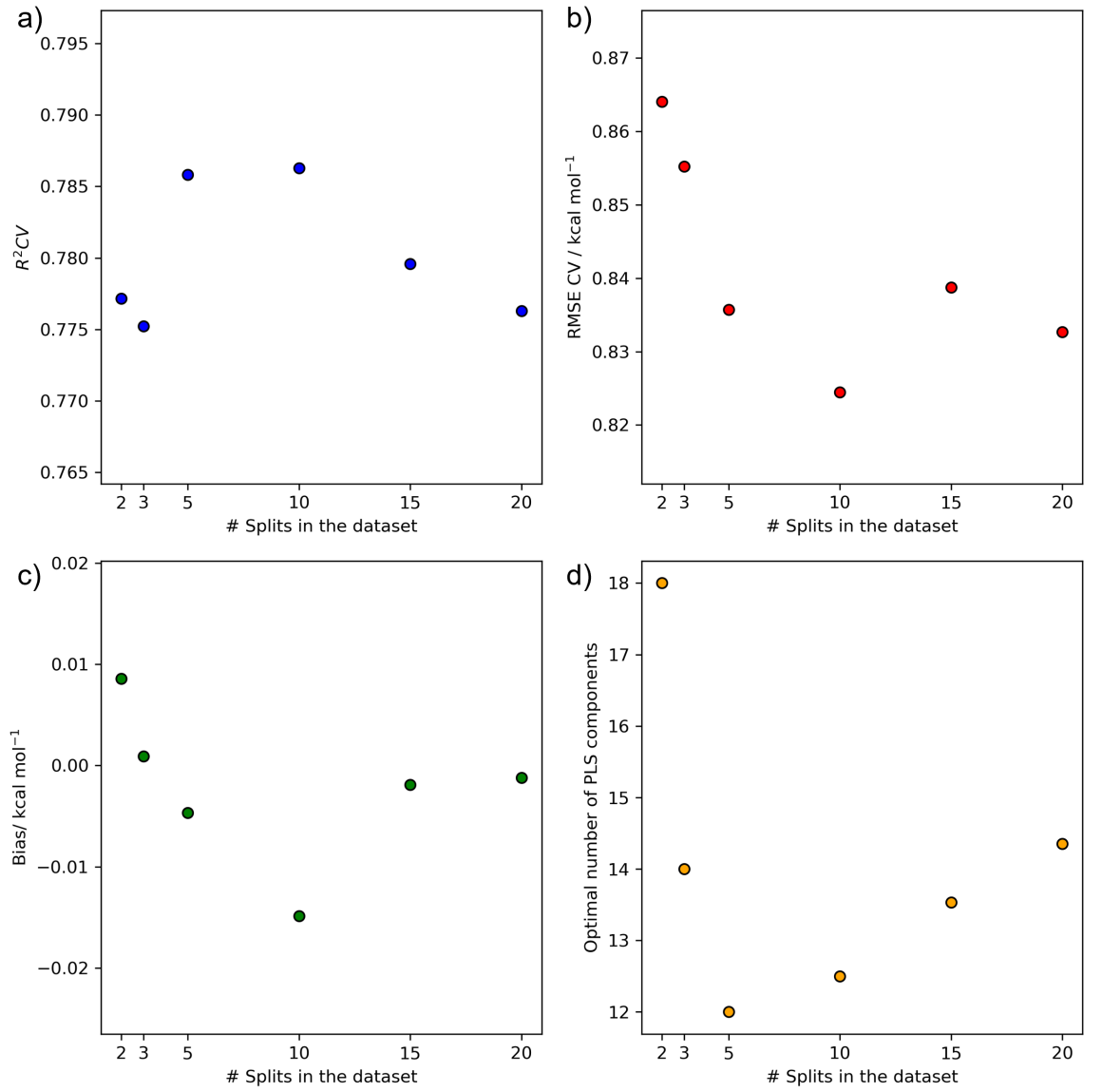

FIGURE 4.12: k-fold cross-validation results for the QPLS methodology using training/testing sets for each split from the $\Delta G^{\circ,m}_{s,i,j}$ with performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d)

effects, as the logarithmic, inverse, or quadratic terms do not contribute much to the model performance. However, these models consistently have better performance across the splits with an RMSE range (across the three models) of approximately 0.830 to 0.870 kcal mol^{-1}. For the E, D, and F basis sets, the performance improves as the number of splits increases from 2 to 20, with split 10 having the best performance for all the splits. For these basis sets, the RMSE values decrease steadily from a range of 0.640 to 0.700 kcal mol^{-1} to 0.600 to 0.680 kcal mol^{-1} for splits 2 to 5, with a low point of 0.46 kcal mol^{-1} for basis set E. In contrast, basis set G has erratic performance across the splits with RMSE values of 1.14 kcal mol^{-1}

for splits 2 and 5, with the best performance of all basis sets at 0.43 kcal mol^{-1} for 10 splits. The R^2 values do not include basis set G for splits 2 and 5 (as these values have significantly worse performance), but the performance of all models generally increases with the number of sets with a range of 0.85 to 0.94. Notably, the D, E, F, and G basis sets have the highest performance at split 10. This performance can be attributed to linear and bilinear terms in the basis sets, where D has inverse terms, E has inverse and inverse bilinear terms, and G has squared bilinear terms. The model size for all models seem to plateau after 10 splits with some the model size ranging from 20 to roughly 180, which can be interpreted as higher model complexity. However, the ALAMO model is a sum of linear and nonlinear terms. Therefore, while there are significantly more terms in the ALAMO model, the complexity of the model is on par with the PLS and QPLS models. The model bias values also range from -0.05 to 0.015 kcal mol^{-1}, which is negligible compared to the magnitudes of the experimental values.

4.2 Comparison between the HX and RX ALAMO models

The HX and RX series of ALAMO models will now be designated with a "HA" for hybrid ALAMO models and "RA" for reduced ALAMO models. Two splits in the reduced $\Delta G^{\circ,m}_{s,i,j}$ and $G^{CDS}_{i,j}$ cross-validation tests have been identified as having the best average values. However, only a single model from each of the splits can be chosen to represent the HA and RA set of models. Tables 4.1 and 4.2 contain the performance metrics for each set in the 10-fold and 3-fold cross-validation splits, respectively. For a given split, the "train", "test" and "overall" headers refer to the model performance concerning the training, testing and overall data sets. Therefore, a model can be chosen based on its overall performance or preference for the training and testing sets. In all of the models across both tables, the bias values are never larger than an absolute value of 0.05 kcal mol^{-1}, which is negligible in comparison to the experimental solvation free energies. Further, it can be seen in the overall RMSE values for the RA models that only one model had poor performance with an RMSE value of 0.955 kcal mol^{-1} whereas the rest of the models have RMSE values that range from 0.409 to 0.467 kcal mol^{-1}. The quality of the trends calculated by the RA models can be seen in the R^2 values with a range of 0.933 to 0.949 for the training data sets. The range of the R^2 values for the validation sets is slightly wider; however, it still shows excellent performance. The model with the lowest RMSE for the validation sets is selected in this work as it offers good

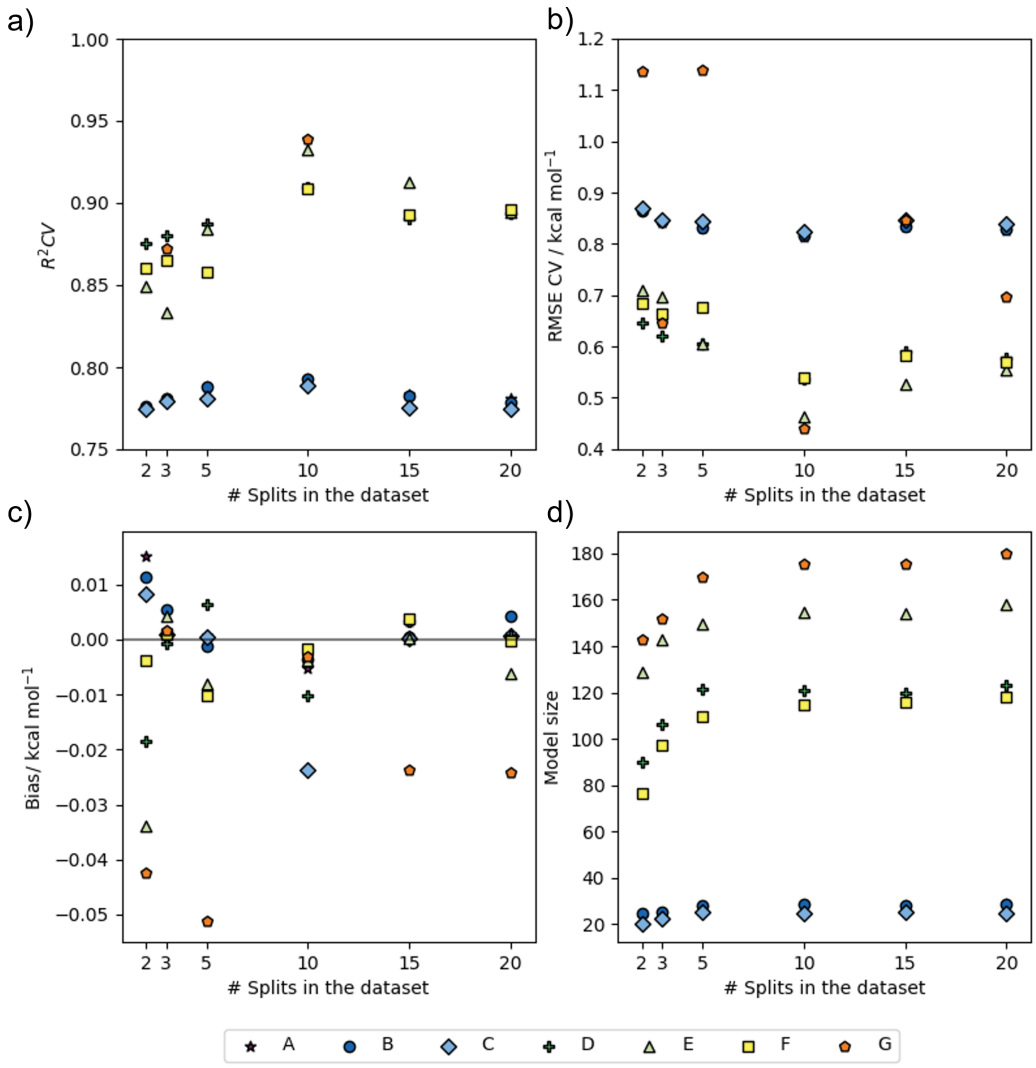

FIGURE 4.13: k-fold cross-validation results for the ALAMO methodology using training/testing sets for each split from the $\Delta G^{\circ,m}_{s,i,j}$ with performance metrics R^2CV (a), RMSE (b), Bias (c) and model size (d)

predictive capability. The model from set 5 has a significantly better RMSE values compared to the rest of the models for the validation set and is therefore the preferred choice. Further, the RMSE value of the training set is also notably low. Therefore, the model from set 5 is chosen to represents the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set and is denoted RA-G10-5 which signifies the 5th model that uses the G basis set from 10th split of the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set.

In contrast, the $G^{CDS}_{i,j}$ model values seen in table 4.2 show slightly worse performance for the overall RMSE values with values of 0.430, 0.442, and 0.445 kcal mol^{-1} obtained for the splits 1 to 3. There are opposing trends in the RMSE values of the training and testing sets where set 1 has a larger RMSE value of 0.435 kcal mol^{-1} for the training set and a significantly

small RMSE value of 0.362 kcal mol^{-1} for the testing set. For sets 2 and 3, their RMSE values of the training set are 0.401 kcal mol^{-1} with RMSE values of 0.516 and 0.523 kcal mol^{-1}, respectively, for the testing set. This contradicting trend is also seen in the R^2 values. The overall R^2 values are also high with values 0.939, 0.940 and 0.943 for splits 1 to 3. Here, set 1 from the $G_{i,j}^{CDS}$ data set has a significantly lower test RMSE value than set 2 and 3 with a slightly higher train RMSE value. Thus, the model of set 1 is chosen to represent the $G_{i,j}^{CDS}$ data set and is denoted HA-3E-1.

When comparing the performance of the RA-G10-5 and HA-3E-1 models, one may think they prefer to use the RA-G10-5 due to the better performance metrics. However, the difference in overall performance (compares the models against the whole set of 2047 data points, making it a fair comparison) is 0.021 kcal mol^{-1}, which is very small. Further, model HA-3E-1 is trained on only a third of the $G_{i,j}^{CDS}$ data set compared to model RA-G10-5, which is trained on 90% of the reduced $\Delta G_{s,i,j}^{\circ,m}$ data set. The model sizes of models HA-3E-1 and RA-G10-5 are 118 and 179, respectively. These points suggest that model HA-3E-1 is more robust in comparison to model RA-G10-5.

On the note of whether or not the hybrid QM/data-driven methodology provides a superior description of the solvation free energy, the PLS and QPLS models saw improved performance when the $G_{i,j}^{CDS}$ data set was used. For the ALAMO models, models basis sets A, B, and C saw a decrease in RMSE of about 0.15 kcal mol^{-1} whereas basis sets D, E, and F saw a general decline of about 0.05 kcal mol^{-1} for splits 10, 15, and 20 and a reduction of approximately 0.1 kcal mol^{-1} for splits 2, 3, and 5. From this, it can be concluded that the new formulation of the hybrid QM/data-driven model is an improvement over the standard $\Delta G_{s,i,j}^{\circ,m}$ data-driven approach. However, one must evaluate a more significant number of combinations of the data sets to determine if this finding is universally true.

4.2.1 Comparison of ALAMO-based models

Three ALAMO models have been identified as representative data-driven models, namely A-D10-10, RA-G10-5, and HA-E3-1; however, these models have been developed using different data sets with respect to the total number of data points, the proportion of data and the spread of the data points within the training and testing sets. Therefore, comparing these models is challenging as an independent set of $\Delta G_{s,i,j}^{\circ,m}$ data is required. Cross-validation indicates the predictive capability; however, if a new data set were to be formed from the

Table 4.1: Table of performance metrics for the each set in 10-fold cross-validation split of the reduced RX ALAMO models with basis set G.

Set	R^2			RMSE / kcal mol^{-1}			Bias / kcal mol^{-1}		
	Train	Test	Overall	Train	Test	Overall	Train	Test	Overall
RA-G10-1	0.945	0.950	0.945	0.425	0.384	0.422	-1.9E-03	-7.9E-03	-2.3E-03
RA-G10-2	0.948	0.955	0.948	0.412	0.410	0.412	3.7E-03	-4.9E-04	3.5E-03
RA-G10-3	0.931	0.966	0.933	0.472	0.370	0.467	-9.1E-03	3.0E-02	-7.1E-03
RA-G10-4	0.710	0.945	0.721	0.976	0.403	0.955	-1.9E-02	2.6E-02	-1.7E-02
RA-G10-5	0.948	0.955	0.948	0.414	0.350	0.411	7.2E-03	3.2E-03	7.0E-03
RA-G10-6	0.948	0.937	0.947	0.412	0.462	0.417	-1.7E-06	-2.6E-02	-2.6E-03
RA-G10-7	0.952	0.909	0.947	0.396	0.560	0.415	9.1E-07	-2.0E-02	-2.0E-03
RA-G10-8	0.950	0.912	0.947	0.404	0.525	0.418	5.2E-07	-5.6E-04	-5.5E-05
RA-G10-9	0.949	0.920	0.946	0.410	0.515	0.421	1.3E-05	1.3E-02	1.3E-03
RA-G10-10	0.949	0.937	0.949	0.408	0.428	0.409	-1.0E-03	-4.9E-02	-3.4E-03

Table 4.2: Table of performance metrics for the each sets in 3-fold cross-validation splits of the HX ALAMO model with basis set E.

Set	R^2			RMSE / kcal mol^{-1}			Bias / kcal mol^{-1}		
	Train	Test	Overall	Train	Test	Overall	Train	Test	Overall
HA-E3-1	0.943	0.955	0.943	0.435	0.362	0.430	-2.3E-03	-1.4E-02	-3.2E-03
HA-E3-2	0.953	0.912	0.940	0.401	0.516	0.442	1.1E-06	3.1E-02	1.0E-02
HA-E3-3	0.949	0.921	0.939	0.401	0.523	0.445	6.7E-06	-1.7E-02	-5.6E-03

reduced $\Delta G^{\circ,m}_{s,i,j}$ database, the models could have already been trained on some of the points, making the test pointless. Alternatively, an intersection of the testing sets between the three ALAMO models could be used as an indicator; however, the data sample could be biased. It may not contain certain molecule classes. Therefore, the only fair comparison is to compare against the overall performance of the models when compared against a data set. This section compares the overall model performance for each model and discusses the individual training and testing set performance.

Table 4.3 contains a comparison of the metrics between model A-D10-10 from section 3.3.6 and the models RA-G10-5 and HA-E3-1 from section 4.2 where the models are assessed against the reduced $\Delta G^{\circ,m}_{s,i,j}$ database. An important observation is an improved performance from the A-D10-10 model to the RA-G10-5 model. This is evidenced by an increase in the R^2 value from 0.920 to 0.945 and an RMSE value decrease from 0.513 to 0.422 kcal mol^{-1}. In terms of magnitude, the bias values are negligible. The training/testing splits for the models are 1951/216 to 1842/205 for A-D10-10 and RA-G10-5, respectively. These splits show the amount of data is roughly the same in both sets; however, the improved performance can also

be attributed to an increase in the model complexity as the model size increases from 132 to 179 terms. When contrasting the basis sets between A-D10-10 and RA-G10-5, the models share linear and bilinear terms; however, basis set G includes squared bilinear terms, whereas basis set D contains quadratic terms. The HA-E3-1 model was shown to be the better choice in section 4.2 as it had a more balanced performance and was trained on a smaller number of points, making it inherently more predictive. Further, the HA-E3-1 model size is only 118 terms, which is less than the A-D10-10 model. However, the E basis set contains additional inverse bilinear terms. Due to being trained against separate data sets, one cannot draw a direct comparison between A-D10-10, RA-G10-5 and HA-E3-1 models. However, two possible conclusions are clear from the comparison. First, a decrease in the number of points in the data set increased performance against the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set. This improvement can be attributed to a smaller range of molecule classes the model has to predict and also unique molecules that appear once. This is evidenced by the number of unique solute and solvent molecule classes in the original $\Delta G^{\circ,m}_{s,i,j}$ data set being 82 and 72, respectively, compared to 61 and 41 in the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set. If a potential user is interested in a model with a wider range of application, A-D10-10 is a better choice because it can be applied to 2167 data points. However, if a user is interested in a more precise model for the solute/solvent systems found in the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set (noted in appendix B), HA-E3-1 is the better choice. The model details for the A-D10-10, RA-G10-5, and HA-E3-1 models can be found in appendix C from tables C.5, C.6, C.7 and C.8.

TABLE 4.3: Comparison of the developed data-driven ALAMO models against the reduced $\Delta G^{\circ,m}_{s,i,j}$ database of 2047 data points.

Metric	A-D10-10	RA-G10-5	HA-E3-1
R^2	0.921	0.948	0.943
RMSE / kcal mol^{-1}	0.516	0.411	0.430
Bias	-1.0E-02	7.0E-03	-3.2E-03
Model Size	132	179	118

4.3 Conclusion

In this chapter, a modification was proposed to improve the performance of data-driven solvation models. The change involved incorporating a detailed description of a quantum-mechanically derived solute molecule and the partitioning of the solvation free energy into an

electrostatic and nonelectrostatic contribution. The electrostatic contribution was accounted for by calculating solute electronic energies in the vacuum and solvent phase, and the nonelectrostatic contribution is accounted for by regressing the 21 solute/solvent descriptors against the $G_{i,j}^{CDS}$ data points. The modification resulted in a new data set of 2047 data points for the solvation free energy $\Delta G_{s,i,j}^{\circ,m}$ and the nonelectrostatic contribution $G_{i,j}^{CDS}$.

In the latter part of this chapter, the new reduced $\Delta G_{s,i,j}^{\circ,m}$ and $G_{i,j}^{CDS}$ data sets were used in conjunction with the PLS, QPLS, and ALAMO frameworks to produce models for predicting solvation free energies. Since these data sets were analogous to one another as $\Delta G_{s,i,j}^{\circ,m}$ and $G_{i,j}^{CDS}$ were interchangeable, the pure data-driven and hybrid QM/data-driven approaches could be compared fairly. It was shown that there was a significant improvement in the RMSE value over the pure data-driven approach when using the $G_{i,j}^{CDS}$ data set with the PLS and QPLS frameworks of roughly 0.15-0.2 kcal mol^{-1}. In contrast, there was a similar improvement in RMSE value when using the $G_{i,j}^{CDS}$ with the ALAMO framework of roughly 0.15 kcal mol^{-1} for mostly linear models and a minor improvement of 0.05-0.1 kcal mol^{-1} for models with bilinear power terms.

From the cross validation study, it was seen that the RA series of models had the best average performance for a split of 10 while using basis G. In contrast, the HA series had the best average performance with a split of 3 when using basis set E. The models with the lowest RMSE values for the validation set were the 5th model for RA models and the 1st model for the HA models. These models were named RA-G10-5 and HA-E3-1. The RA-G10-5 model achieved an overall RMSE value of 0.411 kcal mol^{-1} and an R^2 value of 0.948. In contrast, the HA-E3-1 model achieved an overall RMSE value of 0.430 kcal mol^{-1} and an R^2 value of 0.943. While the RA-G10-5 model has better performance, it was also trained on 90% of the data in comparison to the HA-E3-1 model, which was trained on 66% of it. Therefore, the HA-E3-1 model is inherently more predictive and is chosen as the representative model for the data-driven models.

In conclusion, the ALAMO framework significantly improved the performance of a data-driven generalised solvation model. The incorporation of a detailed QM solute description further enhances the performance of the generalised solvation model.

Chapter 5

Systematic assessment of chosen predictive models for the free energy of solvation

5.1 Objective

In Chapter 2, it was stated that one of the main objectives of this thesis was to carry out a systematic comparison of categorically distinct models ranging from hybrid quantum-mechanical/activity coefficient models, data-driven models and statistical mechanical models. The comparison of these models will provide a benchmark while offering a guide to selecting the best model for a reader's intended purpose. In Chapters 3 and 4, several data-driven models were developed and shown to have excellent predictive capabilities. This systematic comparison will provide further context on how data-driven models compare to models that possess underlying physical theory.

This chapter will first introduce the methodology for carrying out the systematic comparison of the models, then introduce the data sets that will be used to compare the models and finally a thorough quantitative analysis of the models will be carried out which will highlight the strengths and weaknesses of each model.

5.2 Methodology of the comparative study

An important aspect of a systematic comparative study is having a common experimental data set that will allow for a fair comparison of all models involved. In Chapters 3 and 4,

the data-driven models were developed with the intention of being nonaqueous as water is a challenging molecule to model effectively. However, predictive tools such as COSMO-SAC, UNIFAC, or SAFT-γ Mie, are able to represent water through their physical theory and as such are able to model water. Therefore, to have a clear view of the capabilities of each model, two sets of experimental data are used in the systematic comparison. The first data set includes all of the models, which would inherently exclude any solute/solvent systems with water, and the second data set contains all of the solute/solvent systems that can be modelled by each predictive tool that excludes data-driven models. It is also important to discuss any and all computational details related to the calculation of the solvation free energies to ensure any reader can reproduce the calculations. Finally, the performances of each tool need to be quantified using some metrics for the benchmarking process.

5.2.1 Determining the common set of experimental solute and solvent data points

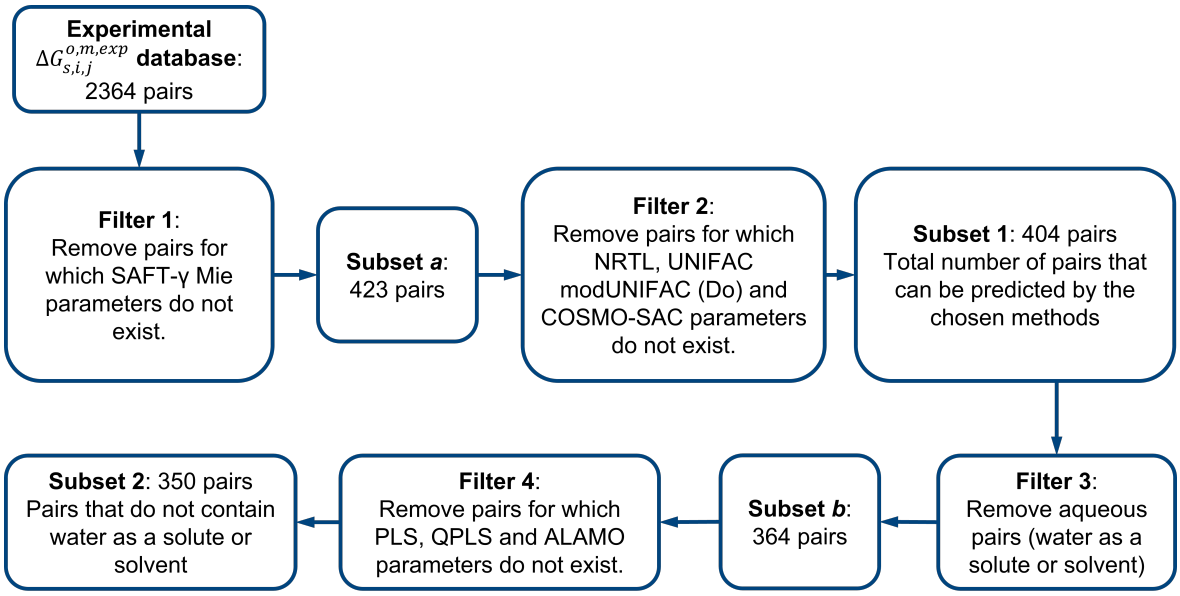

FIGURE 5.1: Flowchart for selecting the common set of solute and solvent pairs from the experimental free energy of solvation database.

A core part of the systematic assessment is selecting data that can be modelled by all of the predictive tools. Figure 5.1 features a flowchart of the logic for selecting this set of solute/solvent pairs, which and consists of three filters. The first filter is specific to the SAFT-γ Mie model as it is the most recent of the approaches considered and can thus treat the least number of solute/solvent pairs. In this filter, the solute/solvent pairs are screened

such that the pairs that do not have a designated parameter for the like and unlike group interactions using SAFT-γ Mie are discarded. Therefore, pairs that rely on combining rules for like and unlike group interactions are not considered. Screening the database through the first filter results in subset a which contains 423 viable pairs that can be modelled by using SAFT-γ Mie. In the second filter, there is a check to remove any pairs for which the parameters for NRTL, UNIFAC, modUNIFAC (Do) and COSMO-SAC are not available. In the cases of NRTL, UNIFAC, and modUNIFAC (Do), like and unlike parameters are required whereas the COSMO-SAC model requires a solute structure and the dielectric constant of the solvent. Therefore, by the end of the second filter, only pairs that can be modelled by each of the five models considered remain. Thus, starting from the complete experimental free energy of solvation database of 2364 data points, this results in 404 solute/solvent pairs in subset 1. In section 3.3.1, water was excluded as a potential solvent in the development data-driven models due to the the vastly different behaviour of water. Subset 1 is one of the two subsets that will be used in the comparison of models. Hence, a third filter is used to discard any pairs that contain water either as a solute or solvent. This resulted in subset b which contains 364 pairs. In the fourth filter, all pairs that do not have PLS, QPLS, or ALAMO parameters are excluded. These pairs include ones could not be modelled in Gaussian and are therefore excluded as the PLS, QPLS or ALAMO HX and RX models would be unable to model them. After the experimental data has been filtered, a resulting subset of 350 solute/solvent pairs is obtained as subset 2. Therefore, subsets 1 and 2 can be thought of the aqueous and nonaqueous sets of data. The solutes and solvents are then classified according to the scheme in section 2.2.4 for later analysis. A list of experimental data points used in this study is provided with their corresponding CAS numbers, molecule classes, types of bonding interactions, and their respective subsets in tables D.1 to D.12.

5.2.2 Computational details for the predictive models

The predictive tools considered in this study utilise different software packages. The gSAFT module in gPROMS ModelBuilder 5.1.0 is used for the calculations with the SAFT-γ Mie model, ASPEN Plus V8.4 is used for NRTL, UNIFAC, modUNIFAC (Do), and COSMO-SAC, an in-house Python code is used for PLS and QPLS, and the ALAMO software package is used for the ALAMO model (Cozad, Sahinidis, and Miller, 2014; Cozad, Sahinidis, and Miller, 2015; Wilson and Sahinidis, 2017). For the equation of state and activity coefficient

models, the Gibbs free energy of solvation, $\Delta G^{\circ,m}_{s,i,j}$, is calculated at 298 K, 1 atm, and at infinite dilution where a solute mole fraction of $x = 10^{-10}$ is used as a representative of infinite dilution. It was mentioned earlier in Chapter 2 that the equation of state and activity coefficient models use the transfer free energy version of the solvation free energy. Therefore, the outputs of these models are converted into the correct form using equation (2.2).

5.2.3 Model validation and error analysis

In Figure 4.4 of Chapter 4, the total number of pairs was referred to with the symbol P, where each pair was denoted with p. Since the common sets of experimental data used for the comparative study are subsets of P (as seen in Figure 5.1), the total number of each set will be designated with Q_v, where v is 1 or 2 depending on the subset. Therefore, the number of solute (i)/solvent (j) pairs belonging to a given subset is denoted as Q_v where $q = 1, 2, \ldots, Q_v$ denotes each pair in a given subset. The experimental and predicted Gibbs free energies of solvation for a given pair of solute i and solvent j in subset v are denoted as $\Delta G^{\circ,m,exp}_{s,q}$ and $\Delta G^{\circ,m,pred}_{s,q}$, respectively, when calculating metrics. The performance of the various models is assessed using several statistical criteria. These criteria can be found in Table 5.1.

TABLE 5.1: Performance criteria for the systematic assessment of predictive tools for the Gibbs free energy of solvation, $\Delta G^{\circ,m}_{s,q}$

Description	**Units**	**Expression**
Mean value of the $\Delta G^{\circ,m,exp}_{s,q}$ values ($\Delta G^{\circ,m,exp}_{s,mean}$)	kcal mol^{-1}	$\frac{1}{Q_v}\sum_{q=1}^{Q_v}\Delta G^{\circ,m,exp}_{s,q})$
Mean value of the $\Delta G^{\circ,m,pred}_{s,q}$ values ($\Delta G^{\circ,m,pred}_{s,mean}$	kcal mol^{-1}	$\frac{1}{Q_v}\sum_{q=1}^{Q_v}\Delta G^{\circ,m,pred}_{s,q}$
Coefficient of determination (R^2)	-	$R^2=\left(\frac{\sum_{q=1}^{Q_v}\left(\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,exp}_{s,mean}\right)\left(\Delta G^{\circ,m,pred}_{s,q}-\Delta G^{\circ,m,pred}_{s,mean}\right)}{\sqrt{\sum_{q=1}^{Q_v}\left(\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,exp}_{s,mean}\right)^2\sum_{q=1}^{Q_v}\left(\Delta G^{\circ,m,pred}_{s,q}-\Delta G^{\circ,m,pred}_{s,mean}\right)^2}}\right)^2$
Root mean square error (RMSE)	kcal mol^{-1}	$\sqrt{\frac{1}{Q_v}\sum_{q=1}^{Qv}\left(\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,pred}_{s,q}\right)^2}$
Mean signed error (MSE)	kcal mol^{-1}	$\frac{1}{Q_v}\sum_{q=1}^{Qv}\left(\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,pred}_{s,q}\right)$
Individual unsigned error (UE$_q$)	kcal mol^{-1}	$\left\lvert\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,pred}_{s,q}\right\rvert$
Mean unsigned error (MUE)	kcal mol^{-1}	$\frac{1}{Q_v}\sum_{q=1}^{Q_v}\left\lvert\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,pred}_{s,q}\right\rvert$
Individual unsigned relative error (URE$_q$)	kcal mol^{-1}	$\left\lvert\frac{\Delta G^{\circ,m,exp}_{s,q}-\Delta G^{\circ,m,pred}_{s,q}}{\Delta G^{\circ,m,exp}_{s,q}}\right\rvert$
Percentage mean unsigned relative error (MURE)	%	$100\times\frac{1}{Q_v}\sum_{q=1}^{Q_v}\mathrm{URE}_q$
Standard Deviation (SD)	kcal mol^{-1}	$\sqrt{\frac{1}{Q_v}\sum_{q=1}^{Q_v}\left(\Delta G^{\circ,m,pred}_{s,q}-\Delta G^{\circ,m,pred}_{s,mean}\right)^2}$

In Table 5.1, there are several performance metrics which are used for the benchmarking of the predictive solvation models. The averages of experimental and predicted solvation free energies are denoted as $\Delta G_{s,q}^{\circ,m,exp}$ and $\Delta G_{s,q}^{\circ,m,pred}$. The coefficient of determination, R^2, is useful for measuring the variance in the experimental data set captured by the predicted data set. In the accompanying expression for R^2, $\sum_q^{Q_v}$ is the sum of all pairs from $q = 1, 2, \ldots, Q_v$ belonging to subset v. The RMSE and MUe metrics are frequently-used metrics for assessing the error between the predicted and experimental data sets whereas the MSE and MRS metrics are useful for indicating (on average) whether the predicted data set is over-or under-predicted. Box plots are used to provide a more detailed representation of the data, where the various aspects of a box plot are identified in Figure 5.2. In box plots, the 25th and 75th percentiles, Q1 and Q3 respectively, are marked as the edges of the box. This indicates where 50% of the data is most commonly found (between Q1 and Q3), and is denoted the interquartile range (IQR). The mean value of the data set is indicated by the darker point and the median value is indicated by the darker horizontal line; both of which are usually within the IQR. In terms of errors, having a narrower IQR or a smaller distance between whiskers shows better agreement between the predicted and experimental data sets. The upper and lower whiskers mark the outer edges of the box plot at 1.5 IQR away from Q1 and Q3. Outliers are data points that lie outside the upper and lower whiskers. In this study, the unsigned error is used as the error metric to derive the box plots.

5.3 Results of the comparative study

In the results section, four tests have been carried out. For these tests, the predictive results from the models are compared to experimental data in non-aqueous solvents (subset 2) and also experimental data in non-aqueous solvents and in water (subset 1). In Figure 5.1, it was shown that subset 1 was meant to be used in the comparison of the SAFT-γ Mie, NRTL, UNIFAC, modUNIFAC (Do) and COSMO-SAC models. In contrast, subset 2 was meant to test the aforementioned models with the PLS, QPLS and ALAMO models as the latter were developed without water as a solvent. In the first test, the three pure activity coefficient models (ACMs), NRTL, UNIFAC, and modUNIFAC (Do) are compared against the experimental data in subset 1 to determine which of these models will be the representative of the ACM class. In the second test, the data-driven models A-D10-10 (Chapter 3), RA-G10-5

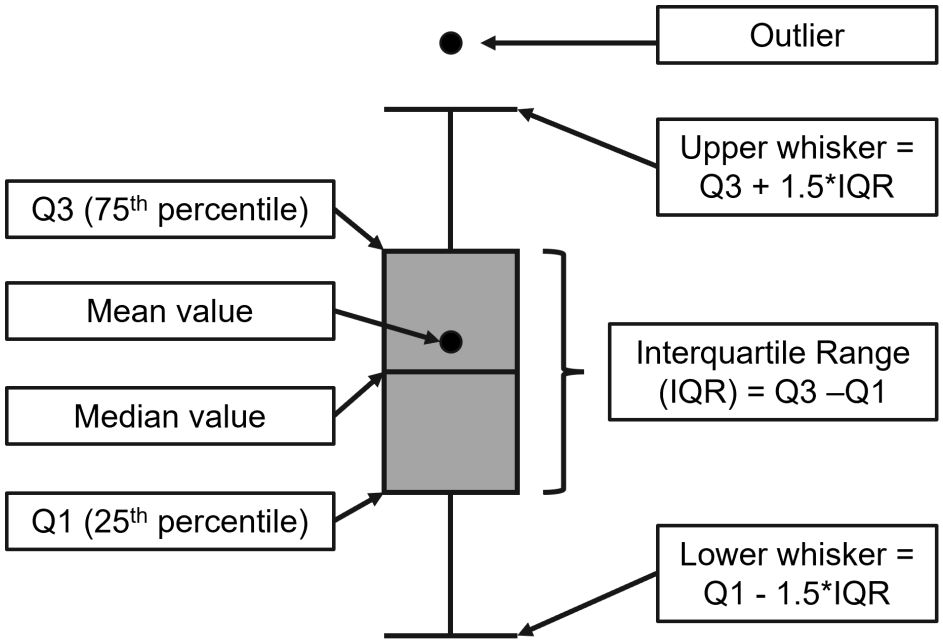

FIGURE 5.2: A figure denoting the various aspects of a box plot where Q1 and Q3 are the 25^{th} and 75^{th} percentiles. The mean is the average value over the original dataset of deviations, and the median is the most frequent value of the deviation data set. The outlier(s) are value(s) that exist outside of the upper and lower whiskers.

(Chapter 4) and HA-E3-1 (Chapter 4) are compared against the experimental data in subset 2 to determine which of these models performs the best with respect to this test. The last two tests will contain a test that focuses on a comparison of the SAFT-γ Mie, modUNIFAC (Do), COSMO-SAC, and HA-E3-1 for nonaqueous use (subset 2) whereas the final test focuses on a comparison of the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models for nonaqueous and water use (subset 1). Each tests consists of an overall comparison of the performance of each model through parity plots and a more detailed comparison according to molecule type in the form of box plots. In these tests, it is important to note that the performance of the models depends on both the physical theory supporting the models and the quality of the parameterisation. Thus, it is impossible to separate these effects and allocate errors to either theory or parameters. However, in models such as the data-driven or hybrid data-driven models found in Chapters 3 and 4, it was possible to compare the PLS, QPLS and ALAMO frameworks directly as the same experimental data was used to train the models. Otherwise, for models such as COSMO-SAC or SAFT-γ Mie which were developed outside this work, this is not possible. The next sections will focus on the four tests.

5.3.1 Comparison of activity coefficient models

In the comparison of the activity coefficient models, experimental data subset 1 was used to compare the NRTL, UNIFAC, and modUNIFAC (Do) models as it includes pairs that contain water as a solute or solvent. There are several figures and tables that are used to showcase the performance of the models. The overall performance of these models is showcased in Figure 5.3 in the form of three parity plots that compare the predicted output from the activity coefficient models and experimental data and Table 5.2 which contains outlines the quantitative performance using metrics found in Table 5.1. Box plots are used to plot the unsigned errors of the models according to the type of interaction between solute and solvent, solute class and solvent class in Figures 5.4, 5.5, and 5.6, respectively. The best performing models according to category are outlined in Tables 5.3, 5.4, and 5.5.

In all of the figures and tables introduced, the modUNIFAC (Do) model has the best overall performance with consistent performance across all plots with the smallest errors. The UNIFAC model achieves slightly worse agreement whereas, the NRTL model has poor performance. This is further evidenced in Table 5.2 by RMSE values of 0.624, 0.680, and 1.202 kcal mol^{-1} and R^2 values of 0.915, 0.900, 0.686, for the modUNIFAC (Do), UNIFAC, and NRTL models, respectively. From Figure 5.3, it can be seen that there is constant underprediction of the data points for all the models with accompanying MSE values of 0.345, 0.450 and 0.722 kcal mol^{-1} for the modUNIFAC (Do), UNIFAC, and NRTL models, respectively. The cluster of data points found in the top right corner of the three parity plots is a series of alkanes (carbon numbers 1 to 8) and isomeric alkanes in water. The isomeric alkanes include solutes such as 2,2,4-trimethylpentane, 2,2-dimethylpropane, 2,4 dimethylpentane, 2-methylpentane, and 2-methylpropane. Therefore, the ACM models shown in this study cannot represent the alkane/water dynamic effectively.

In Figure 5.4, it can be seen the UNIFAC and modUNIFAC (Do) models have similar performance with the latter having a better performance for the SA-NA and NA-NA pairs. These interaction types contain the majority of the points (128, and 146, respectively), resulting in the modUNIFAC (Do) model having consistently better performance. However, the UNIFAC model has slightly better performance for the SA-SA and E-NA pairs, with narrower box plots in both. The NRTL model has consistently worse performance a notably larger error for the SA-NA pair, and a significantly larger range for the NA-SA pairs. The numerous outliers

seen in the NA-SA type exceed 2 kcal mol^{-1} for all the models are the alkanes in water. In Table 5.3, the modUNIFAC (Do) model is seen to have the lowest MUE values for all pairs except SA-SA and E-SA. The minimum, maximum and median values used in Figure 5.4 can be found in tables E.1, E.2 and E.3 of appendix E.1.

In Figures 5.5 and 5.6, there are box plots that illustrate the spread of UEs for each activity model according to different solute classes and solvent classes, respectively. The trend of poor performance for alkanes in water can be seen as outliers in the alkane solute class and the water solvent class. However, for the water solvent class, the spread of the errors is wider with the IQR and whiskers of the box plots ranging from near 0 kcal mol^{-1} (rounded down), to near 3 kcal mol^{-1}. The MUE values of the water class are substantially higher at roughly 1 kcal mol^{-1} for all three models. This larger error suggests it is difficult to model solutes in water with the activity coefficient models. In contrast, the acid solute class is generally harder for the ACMs to model with errors ranging from near 0 kcal mol^{-1} to at least 2 kcal mol^{-1} and the MUE values of the three models are generally higher at 1.152, 0.772, and 0.803 kcal mol^{-1} for the NRTL, UNIFAC, and modUNIFAC (Do) models, respectively. It is also interesting to note the errors (MUE and median values) from the water solute class for the modUNIFAC (Do) model is substantially higher than the NRTL and UNIFAC models. The alkane solvent class has several outliers for the three models. Most other solute and solvent classes have errors that range from 0 kcal mol^{-1} to 1 kcal mol^{-1}, which is within the limit of experimental precision. The minimum, maximum and median values used Figure 5.5 can be found in tables E.4, E.5 and E.6 of appendix E.1 whereas the minimum, maximum and median values used in Figure 5.6 can be found in tables E.7, E.8 and E.9 of appendix E.1.

The alcohol, alkane and aromatic classes make up 340 out of the 404 data points in experimental data subset 1. Therefore, the unsigned errors are more reliable in those cases having been tested more than the acid, amino or ester classes. Thus, the modUNIFAC (Do) has the best performance amongst the activity coefficient models and is selected as the representative that will be tested against the HA-E3-1, SAFT-γ Mie and COSMO-SAC models.

5.3.2 Comparison of data-driven models

In Chapter 3, the ALAMO framework was used to produce three models that outperformed the PLS and QPLS models. It was determined that the A-D10-10, R-G10-5, and HA-E3-1

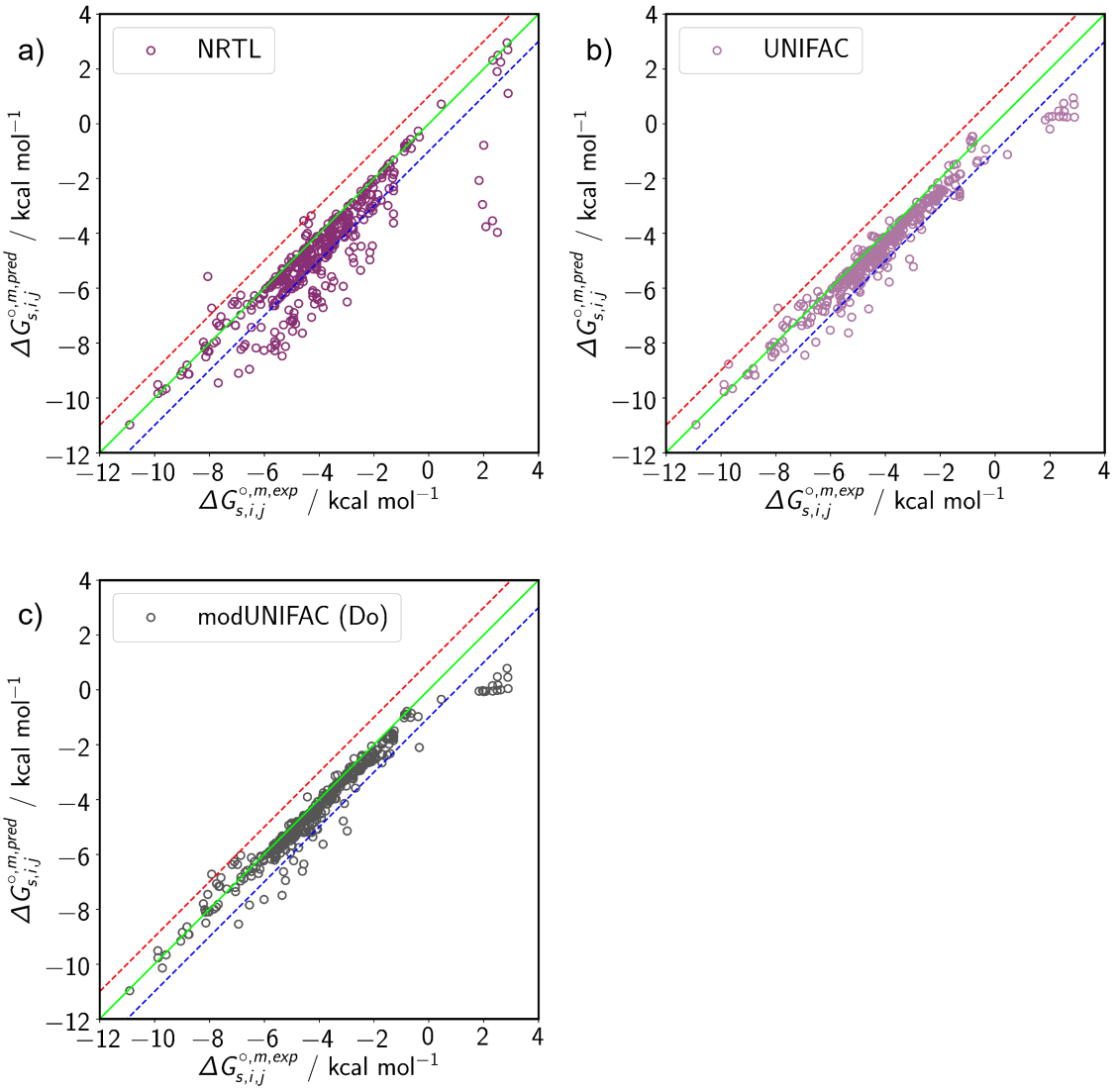

FIGURE 5.3: Parity plots of the NRTL (a), UNIFAC (b), and modUNIFAC (Do) (c) models where $\Delta G^{\circ,m,exp}_{s,i,j}$ and $\Delta G^{\circ,m,pred}_{s,i,j}$ are the experimental and predicted free energies of solvation respectively. The green line represents the equatorial line and the red and blue dotted lines represent a ± 1 kcal mol^{-1} deviation, respectively.

TABLE 5.2: Error analysis for the NRTL, UNIFAC, and modUNIFAC (Do) models when compared against the solute/solvent pairs in subset 1. The definitions of all the metrics can be found in Table 5.1.

Method	NRTL	UNIFAC	modUNIFAC (Do)
RMSE / kcal mol^{-1}	1.202	0.680	0.624
R^2	0.686	0.900	0.915
SD / kcal mol^{-1}	2.105	1.940	1.923
MSE / kcal mol^{-1}	0.722	0.450	0.345
MUE / kcal mol^{-1}	0.791	0.502	0.402
MURE / %	23.88	18.02	14.29

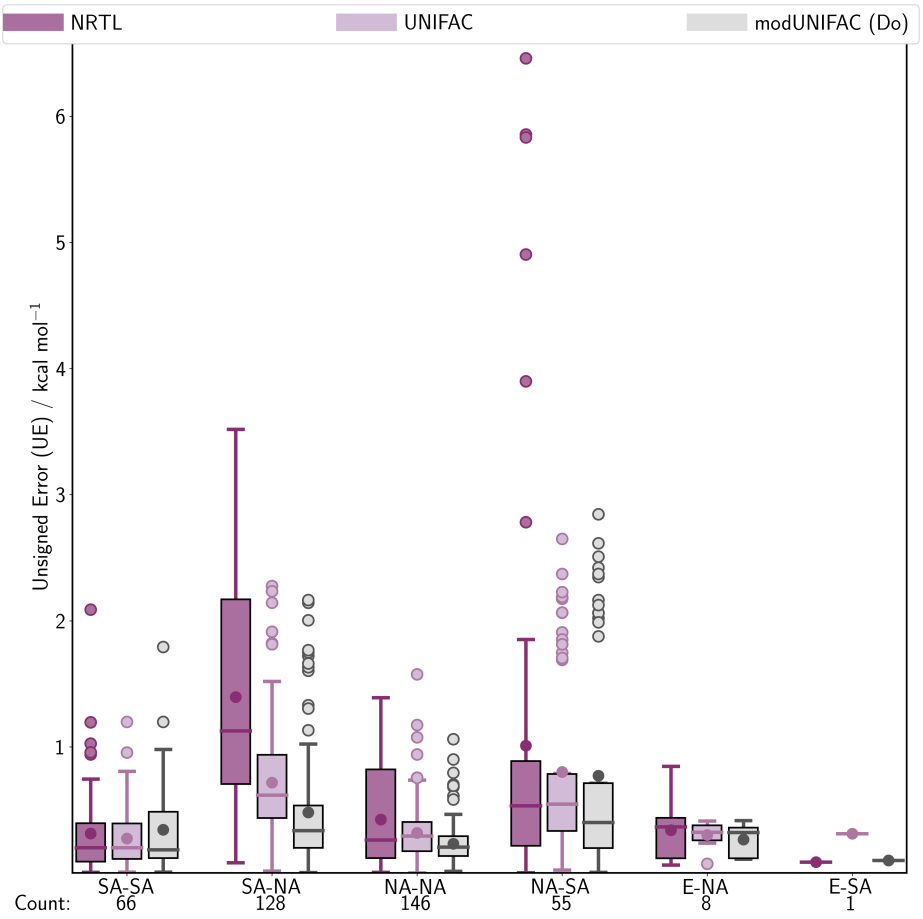

FIGURE 5.4: Box plots of unsigned errors (kcal mol^{-1}) for the NRTL, UNIFAC and modUNIFAC (Do) models per type of interaction between solute and solvent. The errors shown are a comparison of the predicted values and experimental values of the data in subset 1. The labels of the x-axis are written in the form of SOLUTE-SOLVENT, e.g. SA-SA represents a self-associating solute in a self-associating solvent. "Count" represents the number of pairs that exhibit that specific type of interaction. Figure 5.2 outlines the different aspects of a box plot.

TABLE 5.3: MUE per type of solvute/solvent interaction for the NRTL, UNIFAC, and modUNIFAC (Do) models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 1. The MUE value represents the mean point in Figure 5.4 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute, solvent	Count	MUE / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
SA-SA	66	0.313	0.273	0.342
SA-NA	128	1.396	0.716	0.480
NA-NA	146	0.423	0.318	0.231
NA-SA	55	1.009	0.802	0.772
E-NA	8	0.339	0.301	0.265
E-SA	1	0.086	0.311	0.099

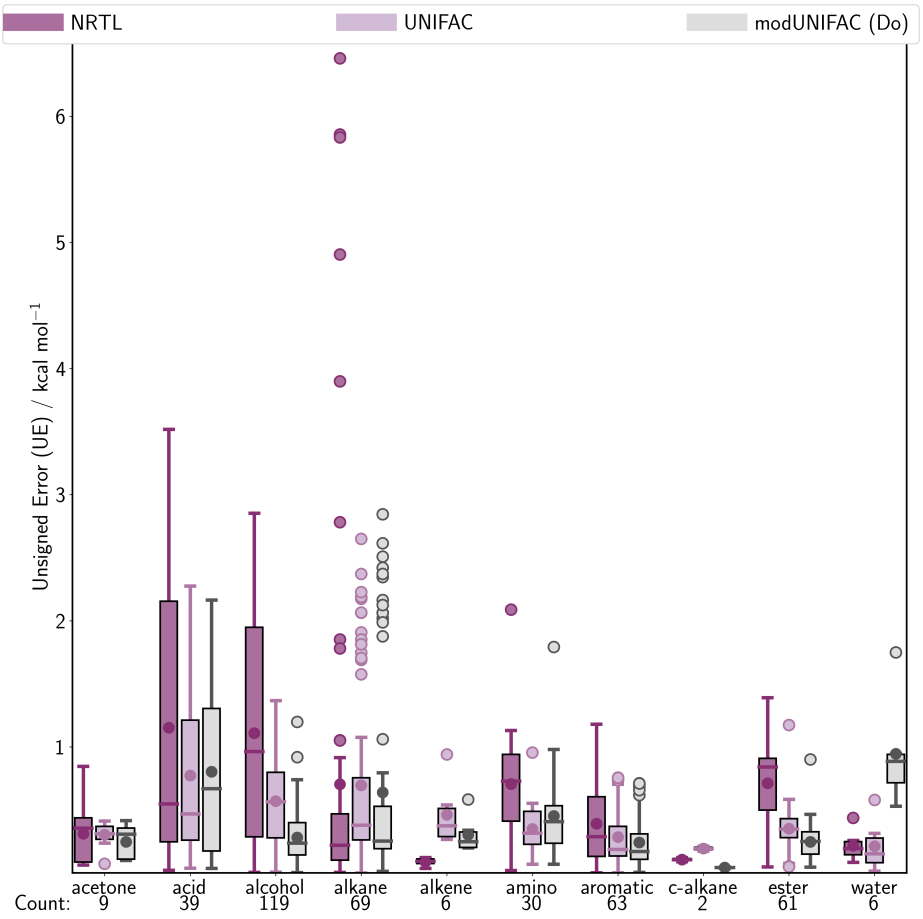

FIGURE 5.5: Box plots of unsigned errors for NRTL, UNIFAC, and modUNIFAC (Do) models per class of solute. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solute, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.4: MUE per class of solute for the NRTL, UNIFAC, and modUNIFAC (Do) models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 1. The MUE value represents the mean point in Figure 5.5 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute Class	Count	MUE / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acetone	9	0.311	0.302	0.246
acid	39	1.152	0.772	0.803
alcohol	119	1.109	0.570	0.282
alkane	69	0.704	0.696	0.638
alkene	6	0.088	0.462	0.303
amino	30	0.706	0.349	0.452
aromatic	63	0.391	0.284	0.241
c-alkane	2	0.105	0.193	0.042
ester	61	0.711	0.351	0.247
water	6	0.217	0.212	0.944

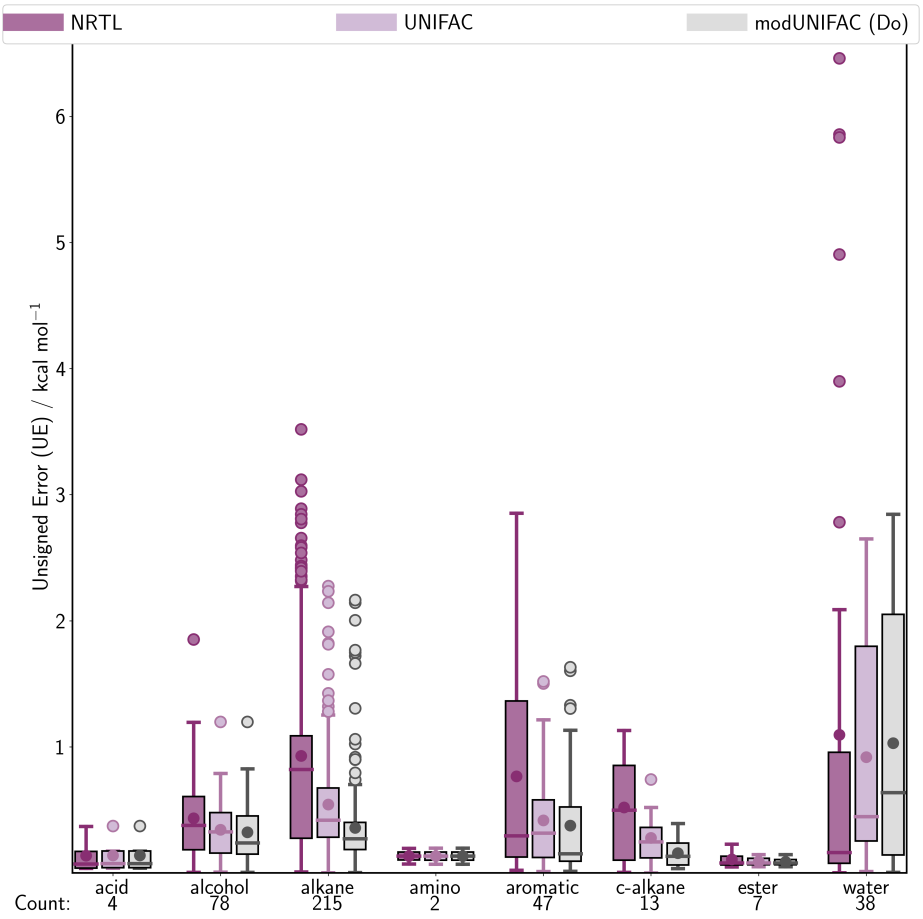

FIGURE 5.6: Box plots of unsigned errors for NRTL, UNIFAC, and modUNIFAC (Do) models per class of solvent. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solvent, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.5: MUE per class of solvent for the NRTL, UNIFAC, and modUNIFAC (Do) models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 1. The MUE value represents the mean point in Figure 5.6 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solvent Class	Count	MUE / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acid	4	0.137	0.140	0.140
alcohol	78	0.433	0.342	0.323
alkane	215	0.929	0.542	0.358
amino	2	0.133	0.133	0.134
aromatic	47	0.766	0.417	0.375
c-alkane	13	0.519	0.278	0.157
ester	7	0.107	0.091	0.088
water	38	1.095	0.919	1.030

models were the representative models of their development paths. The A-D10-10 model was shown to be the best choice for the broader range of 2167 data points versus 2047 data points of the HA-G10-5 and HA-E3-1 models, at the expense of roughly 0.1 kcal^{-1} in RMSE value. The RA-G10-5 model slightly outperformed the HA-E3-1 model, but the latter is inherently more predictive as it has been trained on fewer data and required a lesser number of terms to fit the data. In this section, the models are compared against the nonaqueous experimental data subset 2. The performance of each model is illustrated in Figures 5.7, 5.8, 5.9 and 5.10, accompanied by Tables 5.6, 5.7, 5.8, and 5.9. The minimum, maximum and median values for Figures 5.8, 5.9 and 5.10 can be found in Tables (E.10, E.11 and E.12), (E.13, E.14 and E.15), and (E.16, E.17 and E.18) of appendix E.2, respectively.

The performance of each model is only slightly different from one another; however, in Figure 5.7, on the lower end of the experimental $\Delta G_{s,i,j}^{\circ,m,exp}$ values of -6 to -10 kcal mol^{-1}, the data spread is similar, whereas, at the higher end of -1 to 1 kcal mol^{-1}, there are a few points that behave differently. Nearly all of the data points in each plot reside within the $\pm$ kcal mol^{-1} envelope. In comparison to the ACM models, these ALAMO models have excellent precision. This is also evidenced by RMSE values of 0.327, 0.298, and 0.300 kcal mol^{-1} and R^2 values of 0.964, 0.970, and 0.970 for the A-D10-10, RA-G10-5, and HA-E3-1 models, respectively. The MSE values are also close to zero; resulting in precise predictions with little over-or under-prediction.

In Figure 5.8, the performance of the models is seen to vary according to bonding type. Generally, the A-D10-10 model has larger boxes or larger IQRs compared to the RA-G10-5 and HA-E3-1 models. In contrast, there are marked differences in performance for the RA-G10-5 and HA-E3-1 models. For the SA-NA bonding type, the performance of both models are relatively similar; however, the spread of the outliers in the RA-G10-5 model is larger. In the NA-NA bonding type, while the box plot of the HA-E3-1 model is larger than the RA-G10-5 plot, the outliers do not exceed 0.8 kcal mol^{-1}. Another example can be seen for the NA-SA type where the HA-E3-1 model has both a lower median and mean than the other two models. However, the RA-G10-5 model has the best performance for the SA-SA, and the E-NA bonding types by around 0.3 kcal mol^{-1}. In Table 5.7, it can be seen that the RA-G10-5 model has the lowest MUE for the SA-SA, NA-NA and E-NA, whereas the HA-E3-1 model has the lowest MUE for the SA-NA and NA-SA models.

Figures 5.9 and 5.10 contain two sets of box plots that describe the spread of the UEs

according to the solute and solvent classes, respectively. For the solute classes, the performance of the three models is generally similar, with some exceptions. The RA-G10-5 model has lower errors for the acetone class, but higher errors for the aromatic class, the HA-E3-1 model has substantially larger errors for the alkene class, and the A-D10-10 has significantly larger errors for the c-alkane class. For the solvent classes, the HA-E3-1 model seems to have better performance for the acid, alcohol, aromatic classes but is outperformed in the alkane class. The alkane class consists of 213 of the 350 solute/solvent pairs, therefore, the RA-G10-5 model has marginally better performance overall in comparison to the A-D10-10 and HA-E3-1 models.

Overall, this shows that the ALAMO framework offers excellent precision with slightly better performance from the RA-G10-5 and HA-E3-1 models than the A-D10-10 model. Each model offers varied performance depending on the bonding type, solute class or solvent class; however, the performance is similar for specific groups or any differences are negligible. The HA-E3-1 model is chosen as the representative as it is inherently more predictive (trained on less experimental data and uses fewer terms) while practically having the same performance as the RA-G10-5 model.

5.3.3 Comparison of representative models HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, COSMO-SAC using the nonaqueous subset 2 of the experimental data.

In the previous tests, comparisons were carried out on models belonging to the same category of models (ACMs or data-driven models). In this test, the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models are compared against the nonaqueous experimental data of subset 2. This test will showcase the predictive capability of a statistical mechanical model and the newly developed hybrid data-driven model against established solvation models. The performance of these models are highlighted in Figures 5.11, 5.12, 5.13, and 5.14, with Tables 5.10, 5.11, 5.13 and 5.13 containing important metrics. The minimum, maximum and median values for 5.12, 5.13 and 5.14 can be found in Tables (E.19, E.20 and E.21), (E.22, E.23 and E.24) and (E.25, E.26 and E.27) of appendix E.3, respectively. Each of these models exhibit excellent performance with the majority of the errors residing 0.5 kcal mol^{-1}. These are also reflected in the MUE and median values of the box plots, with some exceptions.

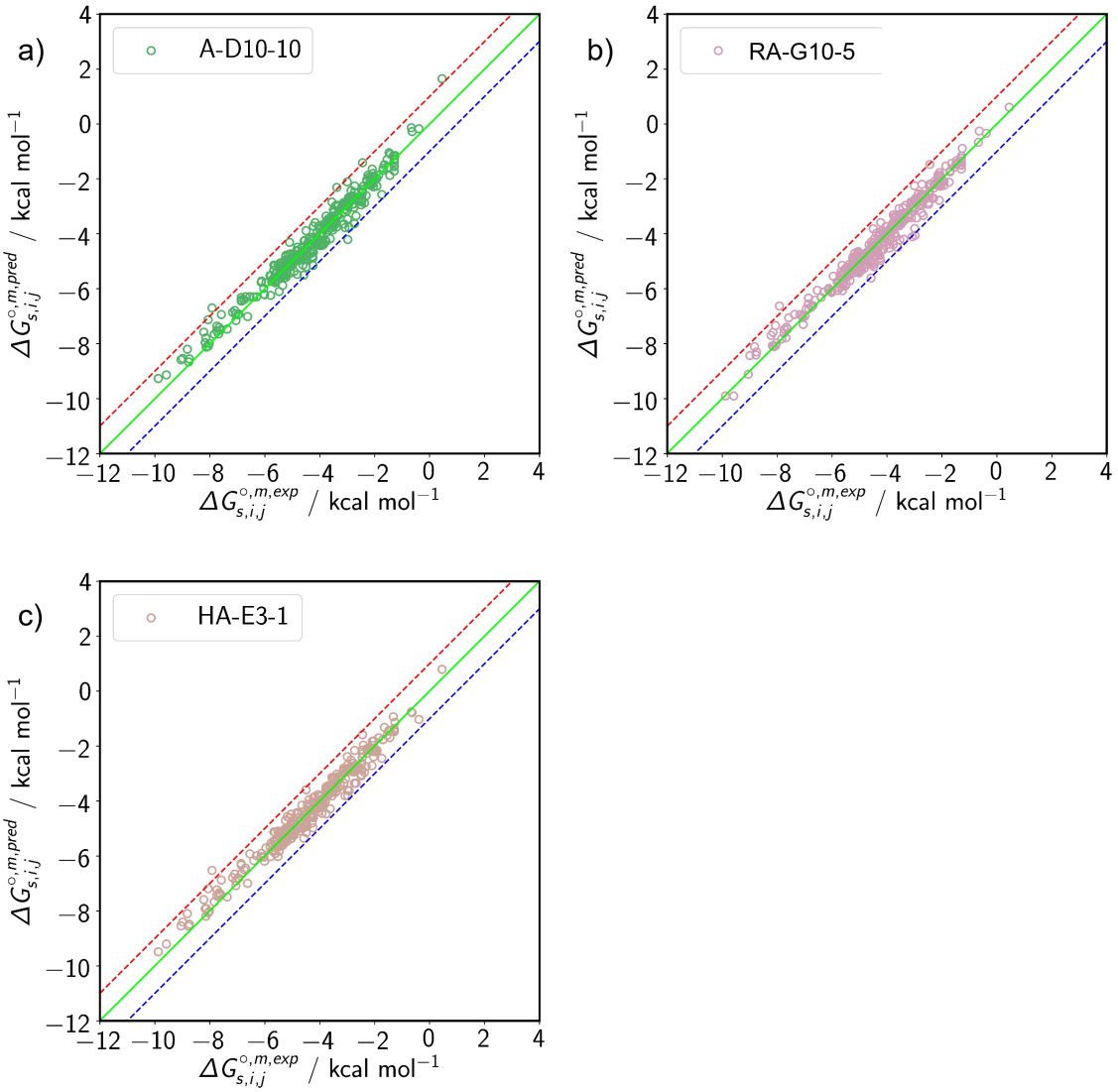

FIGURE 5.7: Parity plots of the A-D10-10 (a), RA-G10-5 (b), and HA-E3-1 (c) models where $\Delta G^{\circ,m,exp}_{s,i,j}$ and $\Delta G^{\circ,m,pred}_{s,i,j}$ are the experimental and predicted free energies of solvation respectively. The green line represents the equatorial line and the red and blue dotted lines represent a ± 1 kcal mol^{-1} deviation, respectively.

TABLE 5.6: Error analysis for the PLS, QPLS and ALAMO models developed in Chapter 3 when compared against the non-aqueous solute/solvent pairs in subset 2. The definitions of all the metrics can be found in Table 5.1.

Method	A-D10-10	RA-G10-5	HA-E3-1
RMSE / kcal mol^{-1}	0.327	0.298	0.300
R^2	0.964	0.970	0.970
SD / kcal mol^{-1}	1.688	1.690	1.650
MSE / kcal mol^{-1}	-0.0177	0.0004	-0.0135
MUE / kcal mol^{-1}	0.242	0.216	0.226
MURE / %	7.670	6.020	6.620

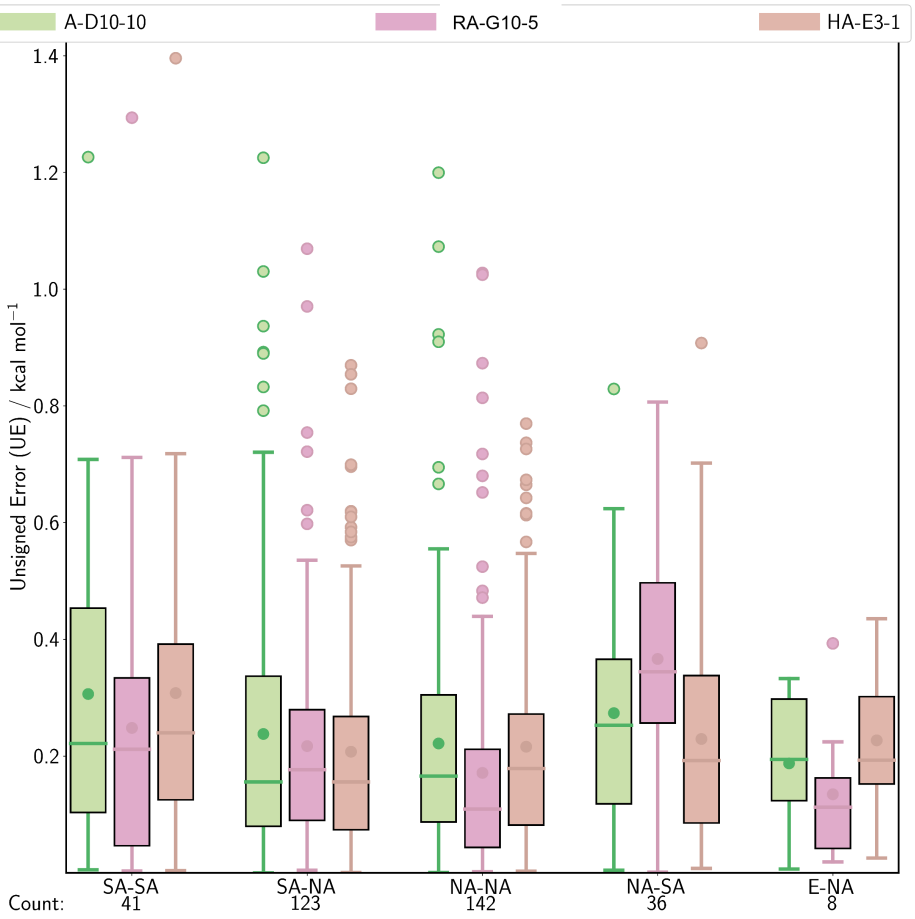

FIGURE 5.8: Box plots of unsigned errors (kcal mol^{-1}) for the A-D10-10, RA-G10-5 and HA-E3-1 models per type of interaction between solute and solvent. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The labels of the x-axis are written in the form of SOLUTE-SOLVENT, e.g. SA-SA represents a self-associating solute in a self-associating solvent. "Count" represents the number of pairs that exhibit that specific type of interaction. Figure 5.2 outlines the different aspects of a box plot.

TABLE 5.7: MUE per class of solute for the A-D10-10, RA-G10-5, and HA-E3-1 models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.8 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute, solvent	Count	MUE / kcal mol^{-1}		
		A-D10-10	RA-G10-5	HA-E3-1
SA-SA	41	0.306	0.248	0.308
SA-NA	123	0.238	0.217	0.208
NA-NA	142	0.222	0.171	0.216
NA-SA	36	0.274	0.366	0.229
E-NA	8	0.188	0.135	0.227

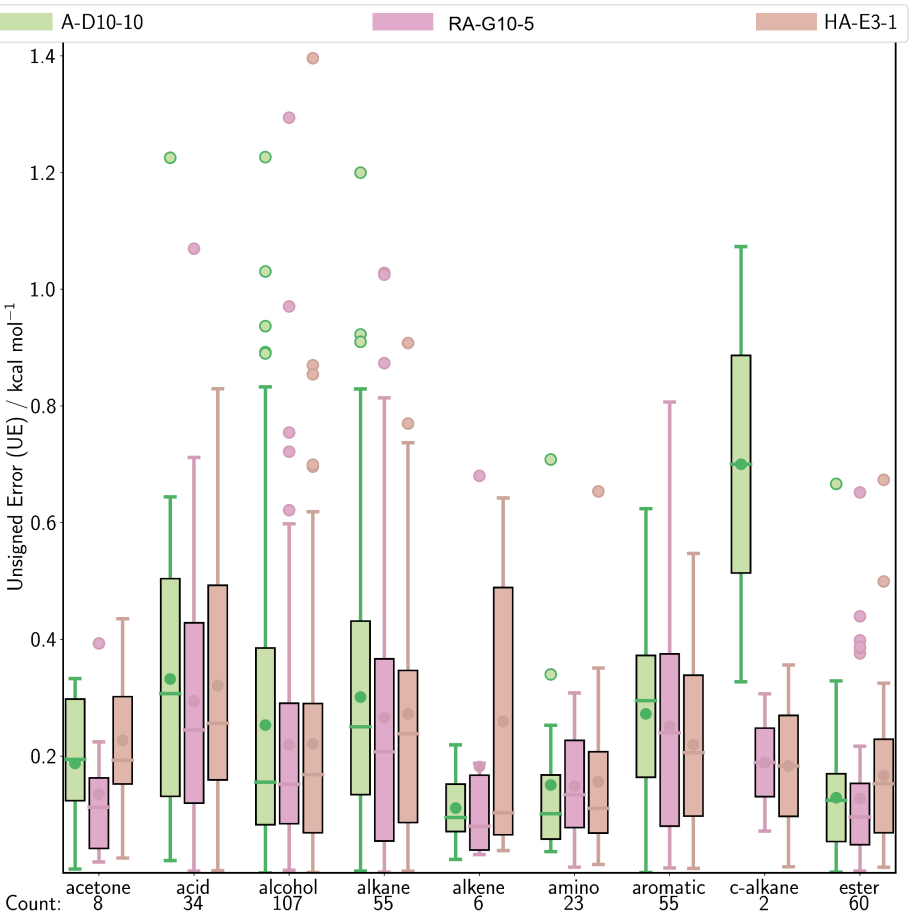

FIGURE 5.9: Box plots of unsigned errors for the A-D10-10, RA-G10-5 and HA-E3-1 models per class of solute. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solute, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.8: MUE per class of solute for the A-D10-10, RA-G10-5, and HA-E3-1 models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.9 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute Class	Count	MUE / kcal mol^{-1}		
		A-D10-10	RA-G10-5	HA-E3-1
acetone	8	0.188	0.135	0.227
acid	34	0.332	0.294	0.321
alcohol	107	0.253	0.219	0.221
alkane	55	0.301	0.266	0.273
alkene	6	0.111	0.182	0.259
amino	23	0.151	0.149	0.156
aromatic	55	0.272	0.251	0.219
c-alkane	2	0.700	0.189	0.183
ester	60	0.129	0.127	0.166

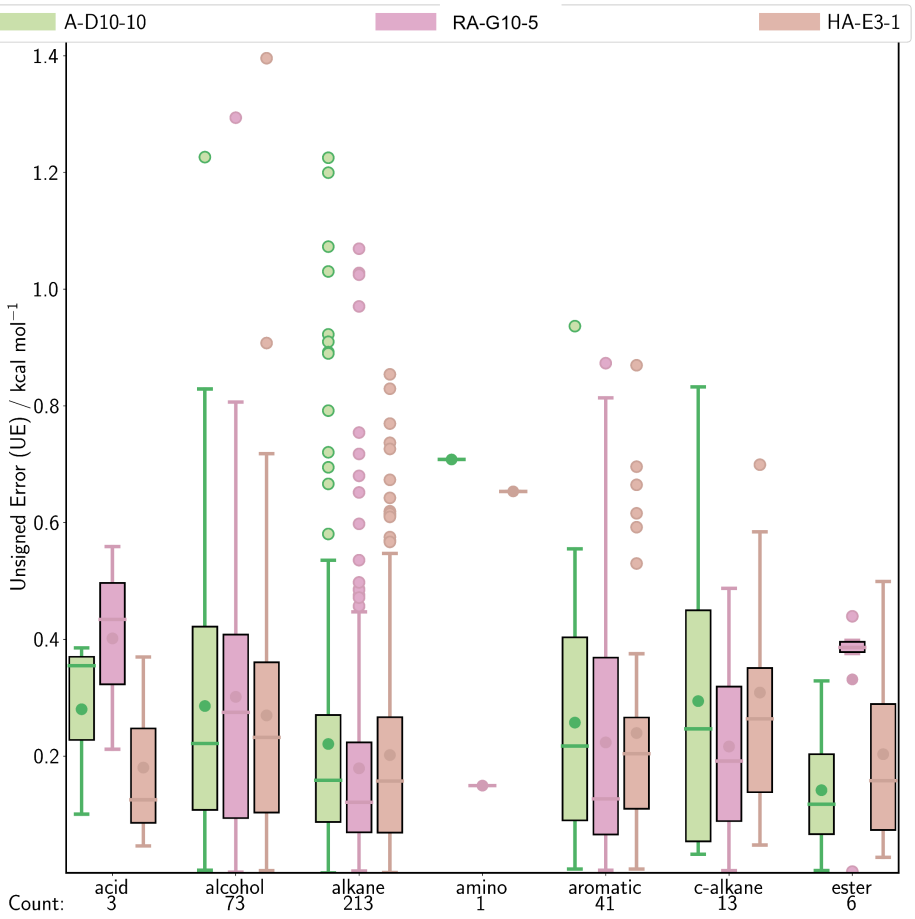

FIGURE 5.10: Box plots of unsigned errors for the A-D10-10, RA-G10-5 and HA-E3-1 models per class of solvent. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solvent, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.9: MUE per class of solvent for the A-D10-10, RA-G10-5, and HA-E3-1 models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.10 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solvent Class	Count	MUE / kcal mol^{-1}		
		A-D10-10	RA-G10-5	HA-E3-1
acid	3	0.280	0.402	0.180
alcohol	730	0.286	0.302	0.270
alkane	213	0.221	0.179	0.202
amino	1	0.708	0.149	0.653
aromatic	41	0.258	0.223	0.240
c-alkane	13	0.294	0.217	0.309
ester	6	0.142	0.331	0.203

The HA-E3-1 model has the best overall performance with a significantly lower RMSE value of 0.300 kcal mol^{-1} compared to 0.433, 0.463 and 0.418 kcal mol^{-1} for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models, respectively. The HA-E3-1 model generally had the best performance across the different bonding types, solute classes and solvent classes. Some exceptions include the SA-SA bonding type, where it was outperformed by the SAFT-γ Mie model. The performance of the HA-E3-1 model was the worst among the models for the alkene solute class and the acid, c-alkane, and ester solvent classes.

The SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models had similar overall performance but the strengths and weaknesses of each model for bonding type, solute class and the solvent class were more pronounced. For these three models, the MSE values show that the predicted values were underpredicted on average. In the case of the modUNIFAC (Do) and COSMO-SAC models, there is a constant underprediction with a few points residing above the equatorial line. In contrast, the distribution of data points also existed was more varied with some points residing near the +1 kcal mol^{-1} line. The acid solutes also proved difficult to model using the three models with the modUNIFAC (Do) mode having the worst performance. Further, outliers were observed for the alkane solvent class with all the models including HA-E3-1 having several data points outside the whiskers. The modUNIFAC (Do) model had the largest number of outliers with some errors ranging to >2 kcal mol^{-1}, resulting in poor performance. Further, the alkane solute class contained the majority of the data points with 213 of the 350 in subset 2.

The SAFT-γ Mie model had the scattered performance for the SA-NA and E-NA bonding types with ranges of 0 to 1.1 kcal mol^{-1} (excluding outliers) and 0.1 to 0.75 kcal mol^{-1}, respectively, for the boxes compared to the performance on other bonding types. Further, the SAFT-γ Mie had a broader range of errors for the acetone solute class and the aromatic solvent classes. However, the errors for the acetone solute class only reach up to 0.7 kcal mol^{-1}, in comparison to the aromatic solvent class. The performance of the SAFT-γ Mie model for the aromatic class is notable as the upper whisker of the box plot reaches errors of 1.4 kcal mol^{-1}. Otherwise, the performance across the solute and solvent classes was relatively consistent with most errors residing under 0.5 kcal mol^{-1}. The SAFT-γ Mie model has the best performance for the SA-SA bonding type, cycloalkane solute class and amino solvent classes with MUE values of 0.236, 0.031, and 0.186 kcal mol^{-1}, respectively.

The COSMO-SAC model has consistent performance across most categories of bonding

types, solute classes and solvent classes with the majority of the errors never exceeding 1 kcal mol^{-1}. Exceptions to this behaviour include the poor performance of the model in the NA-SA and E-NA bonding types, where the errors ranged from 0.25 to 1.1 kcal mol^{-1}. Further, the acid solute class contained errors that ranged from 0.05 kcal mol^{-1} to near 1.5 kcal mol^{-1}. The performance of the COSMO-SAC model for the aromatic solute class is also poor with errors reaching a value of 1.1 kcal mol^{-1}. In contrast, the other models have relatively similar performance for the aromatic solute class with the maximum error reaching 0.7 kcal mol^{-1}. The COSMO-SAC model was also unable to represent alcohol molecules as solvents with the maximum error reaching 1.5 kcal mol^{-1}.

In this section, the HA-E3-1 and, by extension, the A-D10-10 and RA-G10-5 models have the best overall performance compared to the physical models SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC. This difference in performance suggests that the ALAMO framework is excellent for developing data-driven models and has been applied successfully to the prediction of solvation free energies. However, there are some bonding types, solute classes, and solvent classes that the other models excel at predicting the free energy of solvation compared to the ALAMO-type models. Furthermore, a part of the data points used in the comparative study has been used to train all three ALAMO models, whereas the SAFT-γ Mie, modUNIFAC (Do), and COSMO-SAC models have not been trained against the experimental data. This is an important point that needs to be considered when assessing the true capabilities of these models. The physical theory supporting the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models would most likely allow for much more complex effects to be captured in comparison to the ALAMO-type models. Another important note is the fact that nearly all the molecules in this work (and which exist in the form of solvation free energies) are small molecules with low complexity in the form of functional groups. Finally, the ALAMO-type models are developed for binary use only whereas the physical models can consider mixed solvent systems.

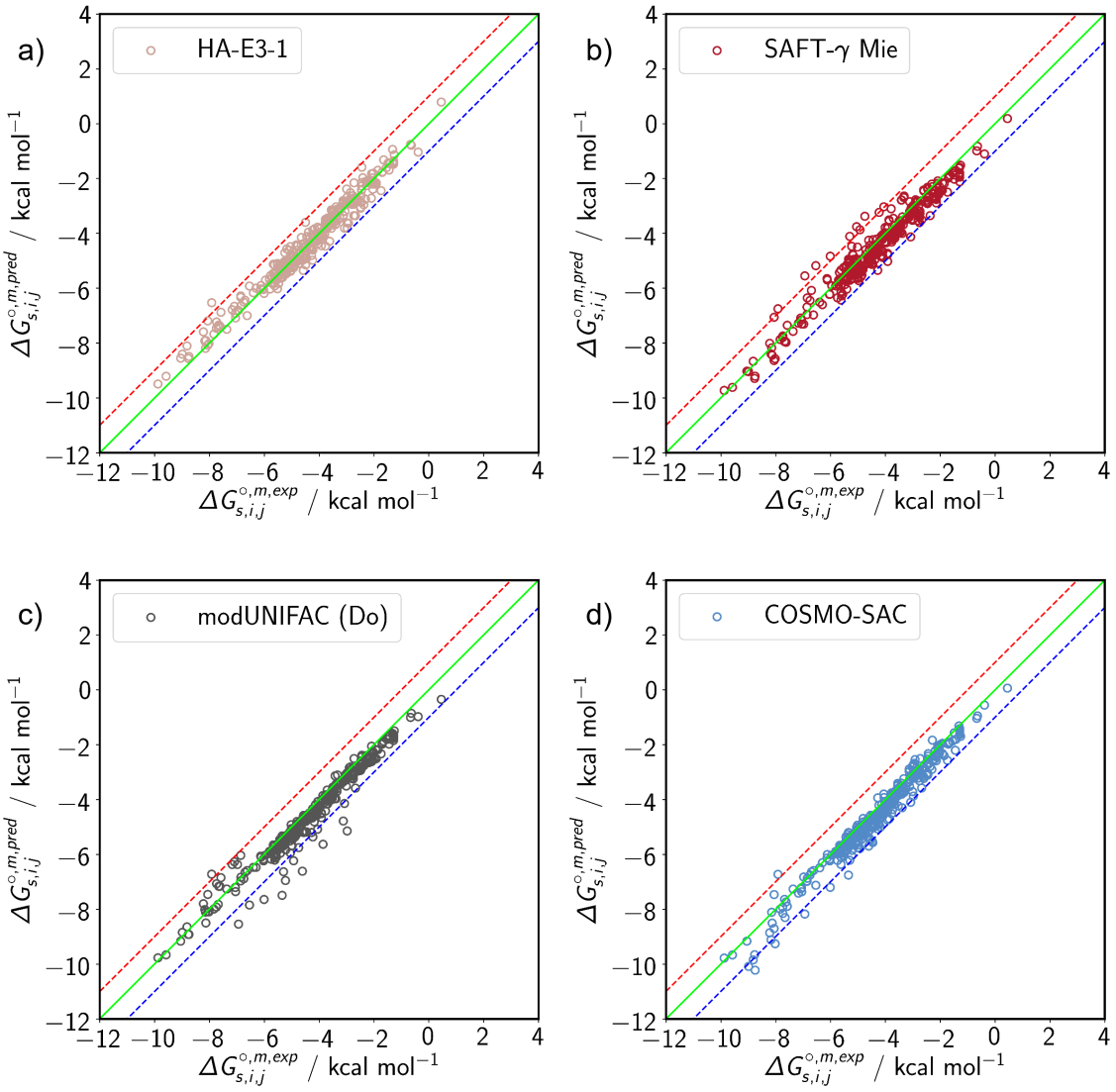

FIGURE 5.11: Parity plots of the HA-E3-1 (a), SAFT-γ Mie (b), modUNIFAC (Do) (c), and COSMO-SAC (d) models where $\Delta G^{\circ,m,exp}_{s,i,j}$ and $\Delta G^{\circ,m,pred}_{s,i,j}$ are the experimental and predicted free energies of solvation respectively. The green line represents the equatorial line and the red and blue dotted lines represent a $\pm$ 1 kcal mol^{-1} deviation, respectively.

TABLE 5.10: Error analysis for the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models when compared against the non-aqueous solute/solvent pairs in subset 2. The definitions of all the metrics can be found in Table 5.1.

Method	HA-E3-1	SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
RMSE / kcal mol^{-1}	0.300	0.433	0.463	0.418
R^2	0.970	0.937	0.927	0.941
SD / kcal mol^{-1}	1.650	1.650	1.670	1.772
MSE / kcal mol^{-1}	-0.014	0.179	0.276	0.283
MUE / kcal mol^{-1}	0.242	0.345	0.332	0.323
MURE / %	7.67	10.54	10.14	9.080

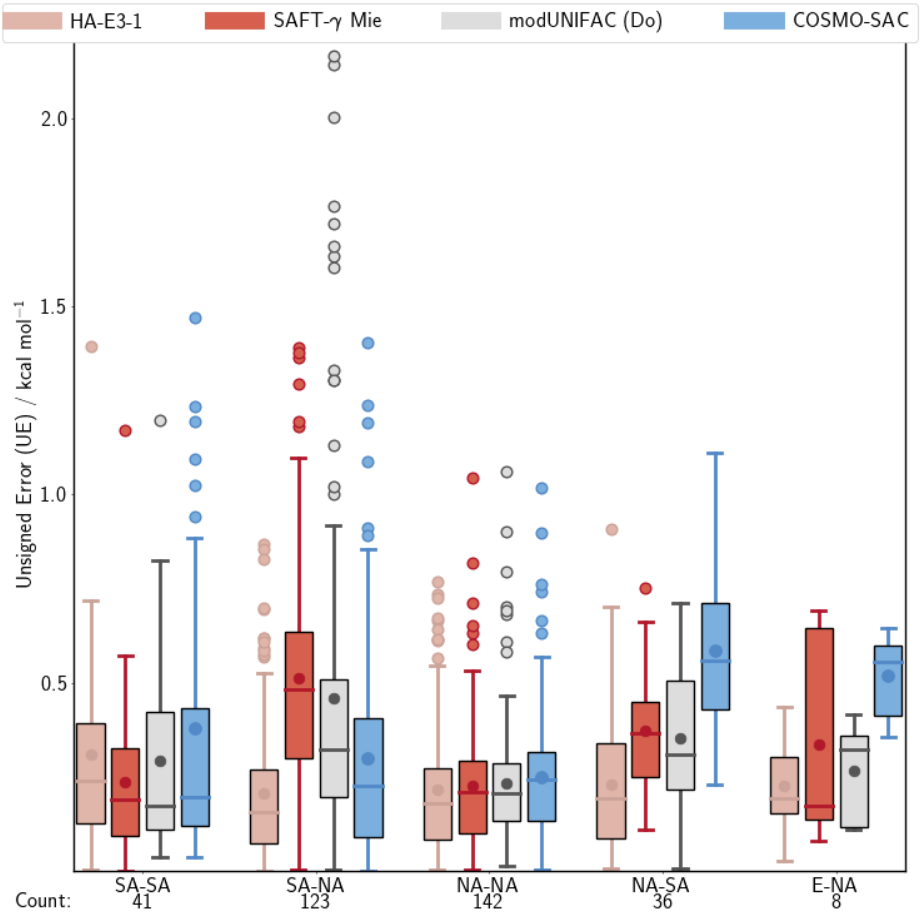

FIGURE 5.12: Box plots of unsigned errors (kcal mol^{-1}) for the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do), and COSMO-SAC models per type of interaction between solute and solvent. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The labels of the x-axis are written in the form of SOLUTE-SOLVENT, e.g. SA-SA represents a self-associating solute in a self-associating solvent. "Count" represents the number of pairs that exhibit that specific type of interaction. Figure 5.2 outlines the different aspects of a box plot.

TABLE 5.11: MUE per type of interaction between solute/solvent for the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.12 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute, solvent	Count	MUE / kcal mol^{-1}			
		HA-E3-1	SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
SA-SA	41	0.308	0.236	0.294	0.38
SA-NA	123	0.208	0.511	0.458	0.299
NA-NA	142	0.216	0.226	0.231	0.249
NA-SA	36	0.229	0.371	0.354	0.585
E-NA	8	0.227	0.336	0.265	0.518

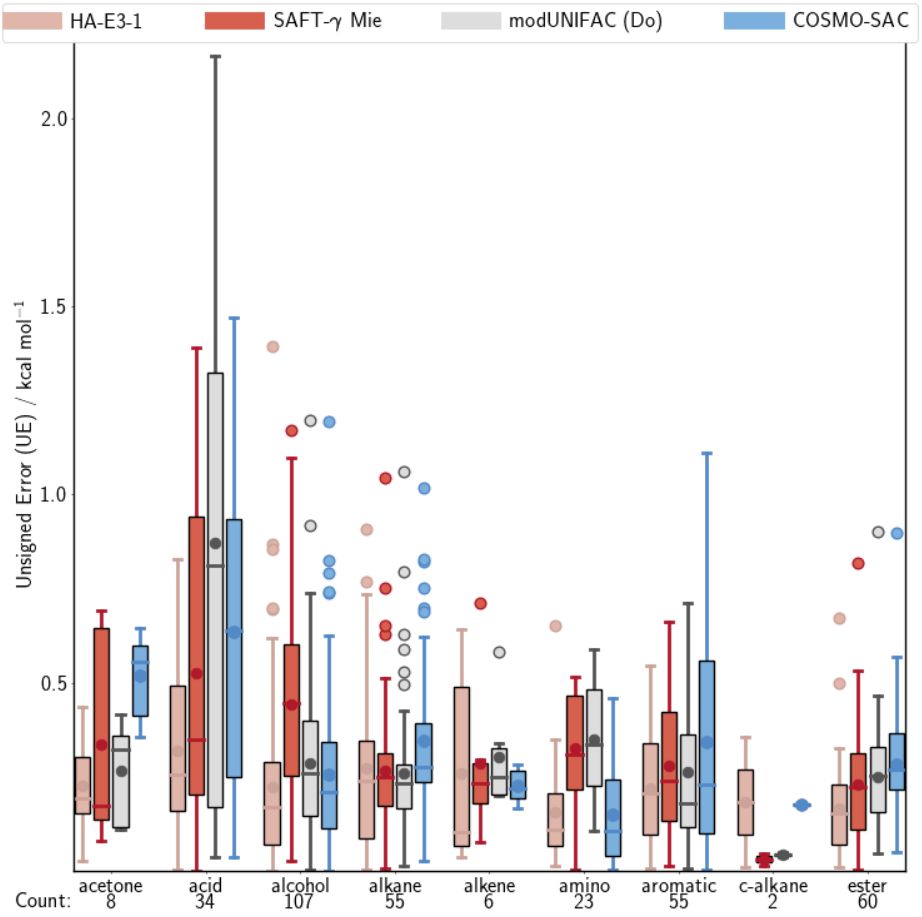

FIGURE 5.13: Box plots of unsigned errors for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models per class of solute. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solute, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.12: MUE per type of class of solute for the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.13 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solute class	Count	MUE / kcal mol^{-1}			
		HA-E3-1	SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
acetone	8	0.227	0.336	0.265	0.518
acid	34	0.321	0.526	0.872	0.635
alcohol	107	0.221	0.441	0.287	0.255
alkane	55	0.273	0.267	0.260	0.344
alkene	6	0.259	0.287	0.303	0.226
amino	23	0.156	0.325	0.348	0.150
aromatic	55	0.219	0.279	0.262	0.341
c-alkane	2	0.183	0.031	0.042	0.176
ester	60	0.166	0.228	0.249	0.284

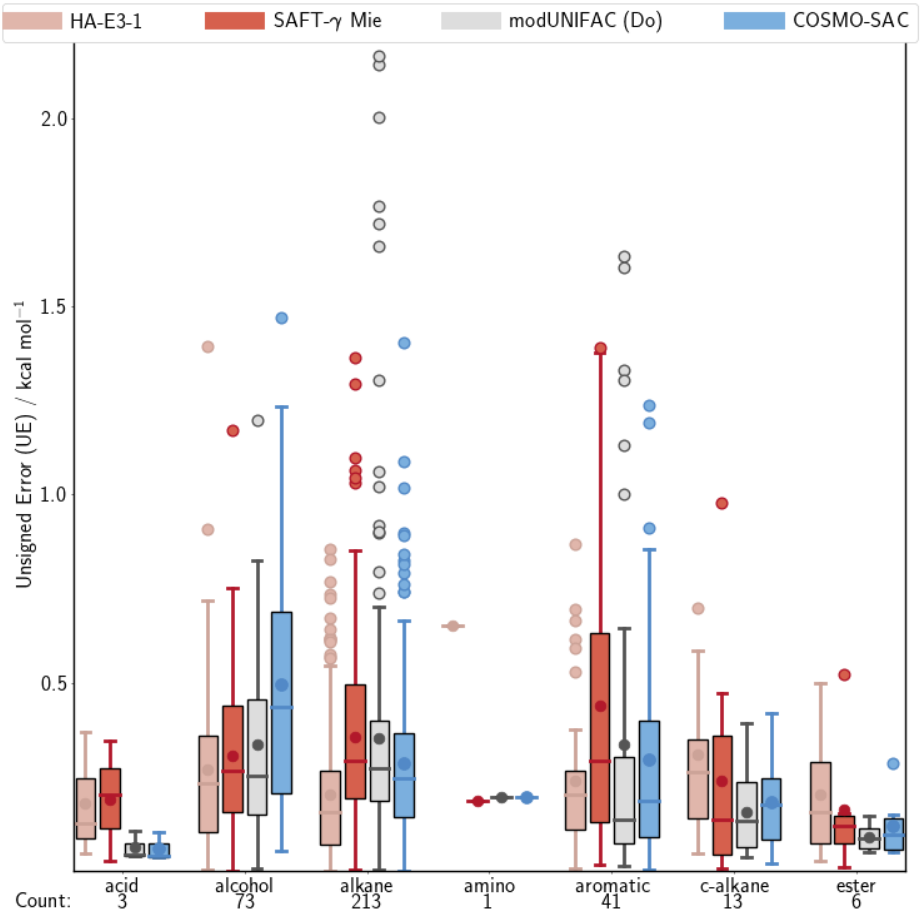

FIGURE 5.14: Box plots of unsigned errors for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models per class of solvent. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solvent, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.13: MUE per type of interaction between solute/solvent for the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models. The errors shown are a comparison of the predicted values and experimental values of the data in subset 2. The MUE value represents the mean point in Figure 5.14 and is defined by the MUE metric in Table 5.1. "Count" represents the number the number of pairs that exhibit that specific type of interaction.

Solvent class	Count	MUE / kcal mol^{-1}			
		HA-E3-1	SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
acid	3	0.180	0.191	0.062	0.06
alcohol	73	0.270	0.305	0.334	0.497
alkane	213	0.202	0.355	0.352	0.286
amino	1	0.653	0.186	0.197	0.196
aromatic	41	0.240	0.439	0.336	0.297
c-alkane	13	0.309	0.240	0.157	0.182
ester	6	0.203	0.164	0.090	0.121

5.3.4 Comparison of representative models SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC using the nonaqueous subset 1 of the experimental data.

In the previous test, the ALAMO model, HA-E3-1, was shown to outperform the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models. However, the HA-E3-1 model was developed for nonaqueous use and cannot model any solute/solvent systems with water effectively. Therefore, the final test excludes the HA-E3-1 models and only involves comparing the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models using the experimental data subset 1, which consists of nonaqueous and water solvents. The performance of each model is detailed in Figures 5.15, 5.16, 5.17 and 5.18, accompanied by Tables 5.14, 5.15, 5.16, and 5.17, respectively. The minimum, maximum, and median values used in Figures 5.16, 5.17 and 5.18 can be found in Tables E.28, E.29 and E.30, E.31, E.32, E.33, E.34, E.35 and E.36 of appendix E.4.

In section 5.3.3, the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models were shown to have comparable performance with slight differences in performance depending on the bonding type, solute class or solvent class. In contrast, the performance of the models is notably different when considering water as a solute and solvent. The cluster of data points seen in the top-right corner of each subplot in Figure 5.15 are alkanes and isomeric alkanes in water. The modUNIFAC (Do) and COSMO-SAC models exhibit poor performance for these data points, whereas the SAFT-γ Mie model has excellent performance. The RMSE and R^2 values of the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models are 0.429, 0.624, and 0.525 kcal mol^{-1}, and 0.960, 0.915, and 0.940, respectively. These results show that the SAFT-γ Mie model has the best overall performance when using experimental data subset 1. The models also have similar performance for the SA-SA and NA-NA pairs, the alkene and ester solute class, and the amino, aromatic, cycloalkane, and ester solvent classes. The aromatic solvent class contains errors that exceed 1 kcal mol^{-1}; however, the MUE and median values in the each box plot resides under 0.5 kcal mol^{-1}. For the other categories, there are marked differences in performance with each model excelling in a certain area. The models are shown to have poor performance when modelling the acid solute class with errors surpassing the 1 kcal mol^{-1} and reaching a maximum of 2.2 kcal mol^{-1}. For the acid solute class, the MUE and median values of the box plots are significantly larger than the predicted overall MUE values,

0.333, 0.402 and 0.379 kcal mol^{-1}. This suggests the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models cannot effectively the acid solute class. There is also generally poor performance when considering the alkane solvent class, which contains several outliers that exceed the 1 kcal mol^{-1} mark. However, for the models, the whiskers of the box plot reside within experimental precision.

The SAFT-γ Mie model has poor performance for SA-NA pairs, with an error range of 0 to 1.45 kcal mol^{-1}, with comparable performance (or the best performance) for other bonding types. Similarly, the SAFT-γ Mie model has the best performance or comparable performance for most solute and solvent classes; however, has the worst performance for the alcohol solute class. Despite having the worst performance, the maximum unsigned error is only 1.2 kcal mol^{-1} with the majority of the errors being less than 1 kcal mol^{-1}. The SAFT-γ Mie predictions for the water solute class have a small range of errors compared to the modUNIFAC (Do) and COSMO-SAC models. Previously, it was shown that the modUNIFAC (Do) and COSMO had a poor performance for alkanes in water as a solvent; however, it is seen that water is also challenging to model when considered as a solute.

The modUNIFAC (Do) model generally has consistent performance across the bonding types with worse performance than the SAFT-γ Mie and COSMO-SAC models. The model has low performance for NA-SA types where outliers reach a maximum of roughly 2 kcal mol^{-1}. These poor results can be attributed to the predictions of the alkane/water systems as evidenced by the results of the alkane solute class and water solvent class. The modUNIFAC (Do) also has several outliers that exceed 1.5 kcal mol^{-1} for SA-NA pairs, further reducing the precision of the model.

The COSMO-SAC model was shown to have slightly worse performance compared to the SAFT-γ model as evidenced by comparable error ranges for the SA-SA and NA-NA pairs. Further, it achieved the best performance for the SA-NA pair; however, has notably poor performance for the NA-SA and E-NA pairs. The results of the NA-SA pair, which reaches a maximum of 2 kcal mol^{-1} are attributed to the poor predictions of the alkane/water systems. Further, the E-NA type has a significantly higher MUE and median value compared to the other two models. The COSMO-SAC model has the best performance for the alcohol solute class, with poor performance for the alkane, aromatic and water solute classes. The range of errors for the aromatic solute class exceeds experimental precision. However, as the median value is relatively close to the other two models; this suggests the COSMO-SAC models cannot

represent some systems effectively. The water solute class shows that the COSMO-SAC model also cannot represent water as a solute. For the solvent classes, the COSMO-SAC model has the worst performance for the alcohol solvent class, with comparable performance for every other solvent class.

For the test between the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models, the SAFT-γ Mie model was shown to have the best overall performance. However, the analysis of the bonding types, solute classes and solvent classes has shown that each method has different strengths; so a potential user should select a model based on their desired solutes and solvents.

5.4 Conclusion

In this chapter, three activity coefficient models, an equation of state, a hybrid quantum-mechanical/activity coefficient model and three data-driven models were benchmarked in four tests using 2 data sets with and without water as a solvent. The first experimental data subset 1 was derived by filtering out solute/solvent pairs that could not be modelled by the SAFT-γ Mie, NRTL, UNIFAC, modUNIFAC (Do) and COSMO-SAC models, whereas the second experimental data subset 2 removes any solute/solvent pairs that contain water. Subset 1 is used when models are compared against nonaqueous and aqueous systems, whereas subset 2 is for nonaqueous systems only. The first test was a comparison of the three activity coefficient models, NRTL, UNIFAC, and modUNIFAC (Do) to select the best model to serve as a representative when compared against the other models. It was shown that the modUNIFAC (Do) model was the best activity coefficient model compared to the NRTL and UNIFAC models using experimental data subset 1. In the second test, the three ALAMO-type data-driven models, A-D10-10, RA-G10-5 and HA-E3-1, were compared to identify a suitable representative. The models were developed for nonaqueous use only and therefore the second test used the experimental data subset 2. It was seen that the data-driven models had very similar performance, with the A-D10-10 model being outperformed by the latter two models. The RA-G10-5 model achieved slightly better overall performance compared to the HA-E3-1 model with RMSE values of 0.298 kcal mol^{-1} to 0.300 kcal mol^{-1}. The difference in performance was deemed negligible and the HA-E3-1 model was chosen as the representative as the model was trained on less experimental data and is inherently more predictive. The third test

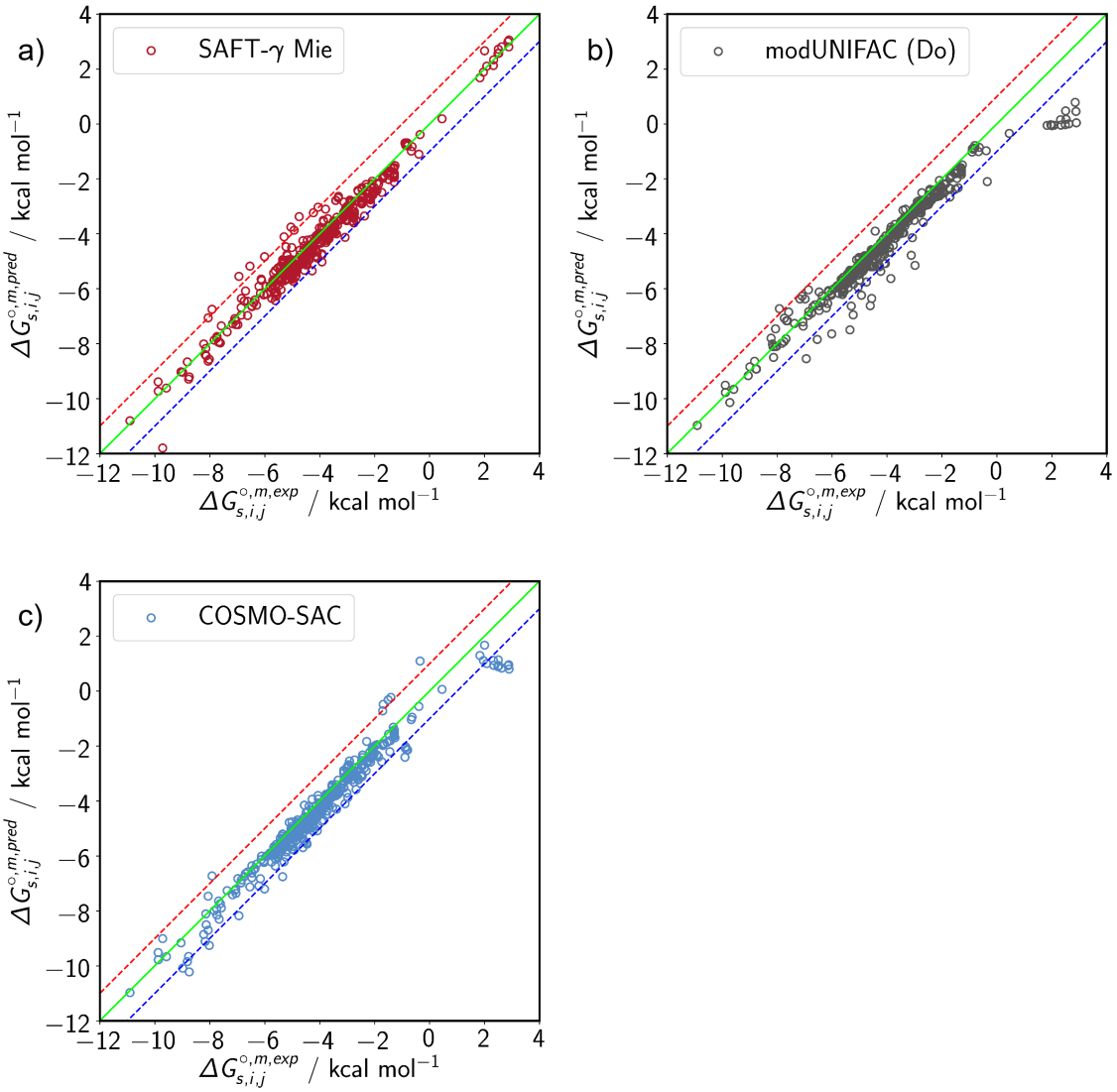

FIGURE 5.15: Parity plots of the SAFT-γ Mie (a), modUNIFAC (Do) (b), and COSMO-SAC models where $\Delta G^{\circ,m,exp}_{s,i,j}$ and $\Delta G^{\circ,m,pred}_{s,i,j}$ are the experimental and predicted free energies of solvation respectively. The green line represents the equatorial line and the red and blue dotted lines represent a ± 1 kcal mol^{-1} deviation, respectively.

TABLE 5.14: Error analysis for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models when compared against the solute/solvent pairs in subset 1. The definitions of all the metrics can be found in Table 5.1.

Method	SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
RMSE / kcal mol^{-1}	0.429	0.624	0.525
R^2	0.960	0.915	0.940
SD / kcal mol^{-1}	2.139	1.923	2.081
MSE / kcal mol^{-1}	0.172	0.345	0.302
MUE / kcal mol^{-1}	0.333	0.402	0.379
MURE / %	10.17	14.29	14.08

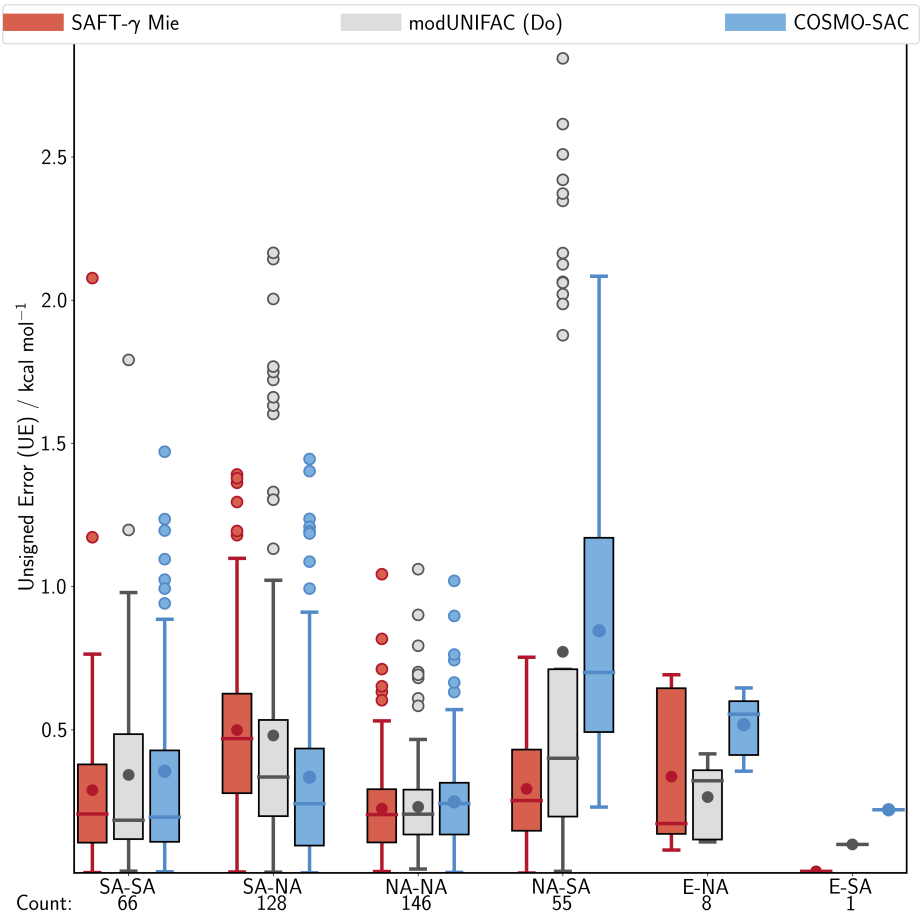

FIGURE 5.16: Box plots of unsigned errors (kcal mol^{-1}) for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models per type of interaction between solute and solvent. The errors shown are a comparison of the predicted values and experimental values of the data in subset 1. The labels of the x-axis are written in the form of SOLUTE-SOLVENT, e.g. SA-SA represents a self-associating solute in a self-associating solvent. "Count" represents the number of pairs that exhibit that specific type of interaction. Figure 5.2 outlines the different aspects of a box plot.

TABLE 5.15: Table showing the model with the lowest MUE per type of interaction. The MUE value represents the mean point in figure 5.16 and is defined by the MUE metric in Table 5.1.

Solute, solvent	Count	MUE / kcal mol^{-1}		
		SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
SA-SA	66	0.289	0.342	0.355
SA-NA	128	0.499	0.480	0.334
NA-NA	146	0.225	0.231	0.248
NA-SA	55	0.293	0.772	0.846
E-NA	8	0.336	0.265	0.518
E-SA	1	0.005	0.099	0.221

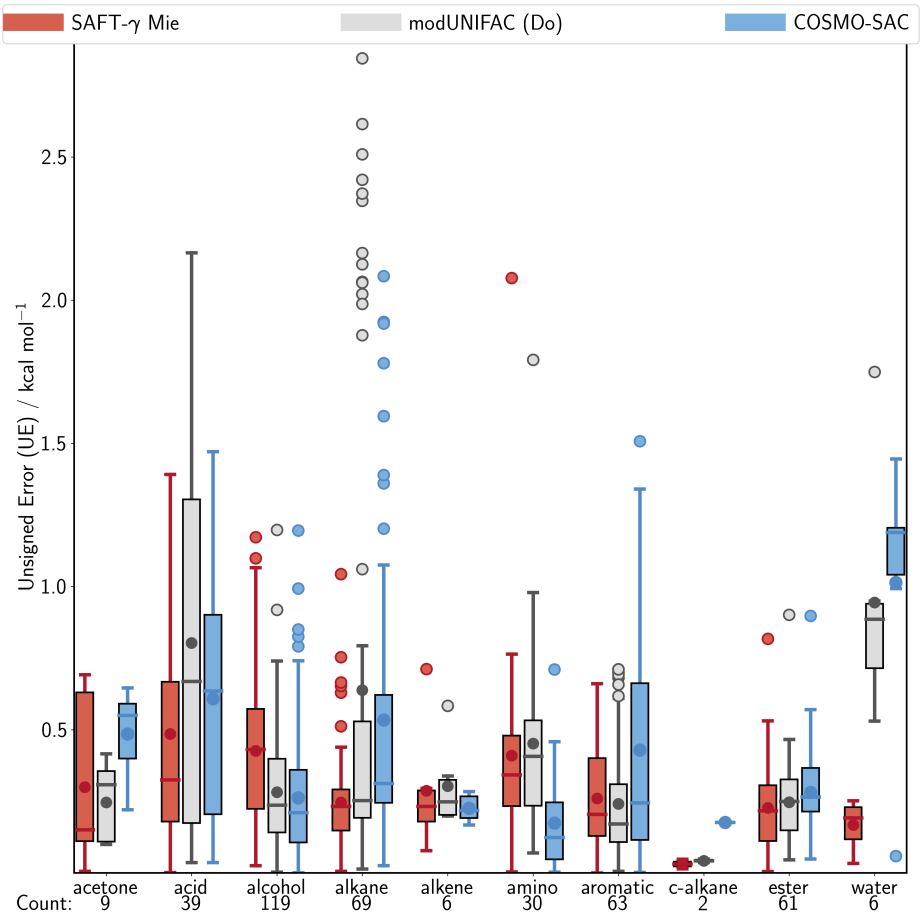

FIGURE 5.17: Box plots of unsigned errors for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models per class of solute. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solute, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.16: Table showing the model with the lowest MUE per solute class. The MUE value represents the mean point in figure 5.17 and is defined by the MUE metric in Table 5.1.

Solute class	Count	MUE / kcal mol^{-1}		
		SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
acetone	4	0.299	0.246	0.485
acid	78	0.485	0.803	0.609
alcohol	215	0.426	0.282	0.262
alkane	69	0.246	0.638	0.534
alkene	6	0.287	0.303	0.226
amino	30	0.409	0.452	0.174
aromatic	63	0.260	0.241	0.429
c-alkane	2	0.031	0.042	0.176
ester	61	0.227	0.247	0.281
water	6	0.167	0.944	1.014

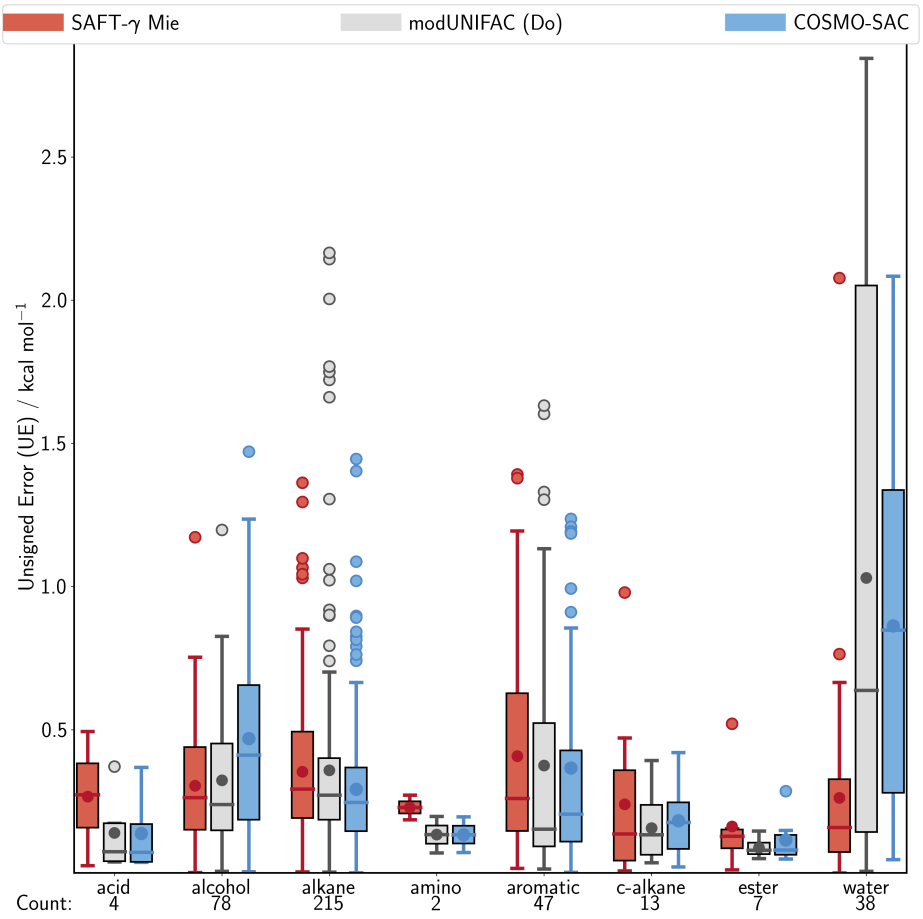

FIGURE 5.18: Box plots of unsigned errors for the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models per class of solvent. The errors shown are a comparison of the predicted and experimental values of the data in subset 1. The labels of the x-axis represent the classes of solvent, where "c-alkanes" denote cycloalkanes. "Count" represents the number of pairs that include a specific class of solvent. Figure 5.2 in the appendix outlines the different aspects of a box plot.

TABLE 5.17: Table showing the model with the lowest MUE per solute class. The MUE value represents the mean point in figure 5.18 and is defined by the MUE metric in Table 5.1.

Solvent class	Count	MUE / kcal mol^{-1}		
		SAFT-γ Mie	modUNIFAC (Do)	COSMO-SAC
acid	4	0.267	0.140	0.137
alcohol	78	0.304	0.323	0.469
alkane	218	0.353	0.358	0.292
amino	2	0.229	0.134	0.133
aromatic	47	0.408	0.375	0.367
c-alkane	13	0.240	0.157	0.182
ester	7	0.162	0.088	0.115
water	38	0.262	1.030	0.862

involved using the nonaqueous subset 2 to compare the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models. It was shown that the HA-E3-1 model achieved the best overall performance; however, some of the experimental data points were used to train the model directly. The model also had varied performance across different bonding types, solute classes and solvent classes. In the fourth test, the SAFT-γ Mie, modUNIFAC (Do), and COSMO-SAC models were compared against one another using the nonaqueous and aqueous subset 1. The SAFT-γ Mie model was shown to have the best overall performance; however, each model excels at different bonding types, solute classes and solvent classes.

An important note pertains to the diversity of molecules in the experimental subsets, as there are only 76 molecules, with the majority of them being alkane, alcohol and aromatic molecules. Furthermore, the combinations of solute and solvent molecules favour alcohols as solutes and alkanes as solvents. This study's findings may show that some models have the best overall performance; however, some molecule classes may only have data points in the single digits. Therefore, it is essential to select a model based on that knowledge. Finally, the HA-E3-1 model, modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models usually have medians or MUEs of 0.5 kcal mol^{-1}, which is significantly lower than the experimental precision, making these models excellent predictive tools for the prediction of solvation free energies. The performance of the the latter three models are notable as the models were not trained on the experimental free energy of solvation database.

Chapter 6

Conclusions and Future work

6.0.1 Summary

The free energy of solvation, $\Delta G^{\circ,m}_{s,i,j}$, was introduced as a fundamental thermodynamic quantity with a broad range of applications in Chapter 1. The property describes the difference in free energy of a solute i in the gas phase and a solute i in a given solvent j at infinite dilution. Given the importance of the free energy of solvation, there have been efforts to compile partition coefficients, Henry's constants, or solubilities in databases to be converted directly into the solvation free energy. However, some solute/solvent system data may not be available despite the amount of experimental data present. This lack of information is often because a solute of interest is not stable (i.e. reaction intermediates, transition states) or is dangerous to handle. In the case of mixed solvents, the potentially infinite number of concentrations and solvent combinations make the measurement of these systems infeasible. Therefore, predictive tools which rival experimental precision are highly desired. A wide range of predictive tools that span empirical data-driven models to *ab initio* quantum mechanical models have been applied to the prediction of solvation free energies. Most systematic assessments of these predictive tools usually apply to one category of models, with only a handful of studies covering multiple types of models. We proposed two objectives for this thesis: i) to develop a hybrid data-driven/quantum mechanical model to predict solvation free energies, and ii) to conduct a systematic assessment that spanned multiple categories of predictive tools.

In Chapter 2, we introduced key concepts such as relevant standard states or reference concentration scales. These standard states or concentration scales affect how the solvation free energy is expressed. Therefore, we present a guide on how to convert the solvation free energy between different forms. We compiled a database of solvation free energies using two sources and ensured all data points had the desired concentration scale and reference standard

state. Further, a series of 9 quantum-mechanical solute and 12 bulk solvent descriptors for a range of solutes and solvents were collated alongside the solvation free energies. In developing a data-driven model, the PLS framework has previously been used to develop a generalised data-driven solvation model that can be applied to any solute or solvent system provided the correct solute and solvent descriptors are available (Borhani et al., 2019). In this thesis, we also utilise the quadratic form of the PLS framework (QPLS) and the ALAMO framework in developing a data-driven model. The experimental database that contained the solvation free energies, solute descriptors and solvent descriptors were used in tandem with the PLS, QPLS and ALAMO frameworks to develop data-driven models in Chapters 3 and 4.

A review of the current state of systematic studies of predictive tools was also presented in Chapter 2. Most of the studies involved testing different iterations of models in a family of predictive tools (UNIFAC, COSMO-RS, COSMO-SAC); however, two studies compared various families of models. These studies include the comparison of two UNIFAC-type and two COSMO-SAC-type models and a comparison of two data-driven models and two *ab initio* quantum mechanical models. These studies show that there is a need for a systematic study that covers models from several families. Several popular models such as the SAFT-γ Mie equation of state, the modUNIFAC (Do) activity coefficient model, and the hybrid quantum mechanical/activity coefficient model, COSMO-SAC were chosen. We provided brief descriptions highlighting the crucial aspects of these models and the steps to calculate solvation free energies using them. In Chapter 5, we compare the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models against the newly-developed data-driven models from Chapters 3 and 4 in a systematic assessment using a common experimental data set.

In Chapter 3, we utilised the PLS, QPLS and ALAMO model frameworks to develop non-aqueous data-driven models using the experimental solvation free energies alongside the solute and solvent descriptions from the database compiled in Chapter 2. Firstly, we introduced the algorithms for each model and explained the important aspects of each model. We conducted cross-validation studies to determine if the distribution of data points in the training and validation sets would affect the predictive capabilities of the developed PLS, QPLS and ALAMO models. This cross-validation procedure involved averaging the performance metrics of models trained on different training and validation data sets. There were negligible effects for the PLS and QPLS models with both models also having similar performance. In the ALAMO

framework, a data-driven model is built on selecting the best subset of mathematical transformations of the descriptor variables from an initially broader set through optimisation. We identified several mathematical transformations such as linear, quadratic, inverse, and several powers of binomial terms that best represented the experimental data and created several basis sets composed of combinations of these terms. The results of the cross-validation study for the ALAMO framework showed more distinct effects when distributing the training and validation data sets for ALAMO models; however, we identified the best ratio of training to validation data and the best basis set which was composed of linear, inverse, and binomial terms. In the cross-validation study of the three frameworks, it was shown that the ALAMO framework offered excellent precision and high customisation and outperformed the PLS and QPLS models. While we identified the best ratio of training to validation data and the best basis set for the ALAMO framework, it was dependent on an average of 10 ALAMO models. Thus, among those models, the representative ALAMO model was chosen as the model with the lowest overall RMSE when compared with 2167 experimental data points. The chosen ALAMO model was named A-D10-10, which stands for an ALAMO model with basis set D (linear, inverse and a binomial term) that uses the 10th model from the selection of models with training to validation ratio of 10:1. The model achieved an RMSE value of 0.516 kcal mol^{-1} and a R^2 value of 0.921. We encourage a potential user to utilise this model to predict solvation free energies due to its wide range of applicability and predictive capability.

After the A-D10-10 model was developed, we sought to try to improve the model by incorporating a detailed description of a quantum-mechanically derived solute molecule in a solvent in Chapter 4. Doing so would provide information about how the solute i and solvent j molecules interact together. The free energy of solvation $\Delta G^{\circ,m}_{s,i,j}$ was partitioned into an electrostatic contribution, $\Delta E^{el}_{i,j}$, which is represented by the QM description of the solute molecule in a solvent, and nonelectrostatic contribution, $G^{CDS}_{i,j}$. The electrostatic contribution, $\Delta E^{el}_{i,j}$, is the difference between electronic energy of the solute molecule in the gas phase, $E^{el,IG}_{i,j}$, and the electronic energy of the solute molecule in the solvent phase, $E^{el,L}_{i,j}$. These electronic energies were calculated using the IEF-PCM model with the X3LYP/6-31G(d,p) level of theory in the Gaussian09 model suite (Frisch et al., 2016). Compared to the 2167 experimental data points of solute/solvent pairs used to develop the nonaqueous A-D10-10 model, we could only model 2047 of these solute/solvent pairs in Gaussian09. Once the 2047 pairs of gas-phase and solvent-phase electronic energies were calculated, the 2047

experimental $\Delta G^{\circ,m}_{s,i,j}$ values in the database were converted into $G^{CDS}_{i,j}$ values. This conversion into $G^{CDS}_{i,j}$ values resulted in two databases: i) a reduced database of $\Delta G^{\circ,m}_{s,i,j}$ and ii) and a database of $G^{CDS}_{i,j}$ values which contain the same solute/solvent pairs and solute and solvent descriptors. We used both databases to develop models with the PLS, QPLS and ALAMO frameworks to assess if incorporating the QM description of the solute molecule improved performance.

In the latter half of Chapter 4, the new reduced $\Delta G^{\circ,m}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets were used in conjunction with the PLS, QPLS and ALAMO frameworks to produce a series of data-driven (RX series of models) and hybrid QM/data-driven models (HX series of models), respectively, to predict solvation free energies. A cross-validation study using the same procedure as the one in Chapter 3 was carried out using both the $\Delta G^{\circ,m}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets to determine if the distribution of data in the training to validation sets affected model performance. In the context of the PLS and QPLS models, we found a significant improvement in the RMSE value of 0.15 to 0.2 kcal mol^{-1} for the HX series of models when compared to the RX series of models. For the ALAMO framework, we developed two families of models, the RA series which refers to ALAMO models developed using the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set and the HA series which refers to ALAMO models which were developed using the $G^{CDS}_{i,j}$ data set. For both families of models, we chose to use the same combined basis sets like the ones selected in Chapter 3 (Table 3.13). The results were similar to the ones in Chapter 3 for both families where the ALAMO models outperform the PLS and QPLS models with a varied performance across the different ratios of training to validation data. For the HA models, we determined there was a similar improvement (as seen in the HX PLS and QPLS models) of roughly 0.15 kcal mol^{-1} compared to the RA models when considering mostly linear models. However, only a minor improvement of 0.05 to 0.1 kcal mol^{-1} was observed when considering models with bilinear power terms.

From the cross-validation study, we were able to identify the best ratios of training to validation data and the best basis sets for the RA and HA series of models. The RA series achieved the best average performance with a ratio of 9:1 training to validation data and the basis set "G", which contains a linear, binomial and squared binomial term. In contrast, the HA series achieved its best average performance with a ratio of 2:1 training to validation data and the basis set "E", which contains linear, inverse, binomial, and inverse binomial terms. We selected representative models according to the lowest RMSE value of the validation

set. Therefore, the best models were the RA-G10-5 and HA-E3-1 models which achieved overall RMSE values of 0.411 and 0.430 kcal mol^{-1}, respectively. Based on these values, one can conclude that the RA-G10-5 model, which uses the reduced $\Delta G^{\circ,m}_{s,i,j}$ data set has superior performance. However, the model has been trained on 90% of the experimental in comparison to the HA-E3-1 model, which has only been trained on 66% of the data, making it inherently less predictive than the latter. Thus, we selected the HA-E3-1 model as the representative model for the data-driven models and have shown that the hybrid QM/data-driven approach enhances the performance of the data-driven model approach.

In Chapter 5, we carried out a systematic assessment of a range of predictive tools that included the SAFT-γ Mie equation of state; the NRTL, UNIFAC, and modUNIFAC (Do) activity coefficient models; the three data-driven ALAMO models, A-D10-10, RA-G10-5 and HA-E3-1; and the hybrid QM/activity coefficient model COSMO-SAC. A critical aspect of systematic studies is to determine a common validation data set to benchmark the models. We filtered the compiled database of 2364 data points from Chapter 2 to ensure each model could model each solute/solvent pair and obtained a nonaqueous and aqueous subset of 404 data points (subset 1) and a nonaqueous subset of 350 data points (subset 2). The nonaqueous subset is necessary as the data-driven models A-D10-10, RA-G10-5 and HA-E3-1, were developed for nonaqueous use only. Further, having two subsets of data allows us to assess how the performance of the non-data-driven models changes when adding aqueous pairs. In total, Chapter 5 contained four tests. The first two were designed to select a representative from the activity coefficient models using experimental data subset 1 and to select a representative from the data-driven models using experimental data subset 2. We selected the models based on their overall performance when compared to their corresponding subsets of data. The modUNIFAC (Do) model exhibited the best performance amongst the activity coefficient models. In contrast, the RA-G10-5 and HA-E3-1 models outperformed the A-D10-10 model and had very similar performances. We chose the HA-E3-1 model as the representative data-driven model as it was trained on less experimental data and is inherently more predictive.

The other two tests involved comparing the SAFT-γ Mie and COSMO-SAC models with the representative ACM and data-driven models. The nonaqueous test (subset 2) involved the HA-E3-1, SAFT-γ Mie, modUNIFAC (Do), and COSMO-SAC models. We observed superior performance from the HA-E3-1 with an overall RMSE value of 0.300 kcal mol^{-1} compared to 0.433, 463, and 418 kcal mol^{-1} for the SAFT-γ Mie, modUNIFAC (Do), and COSMO-SAC

models, respectively. We also categorised unsigned errors into box plots and found that the HA-E3-1 model has consistently low errors for nearly all categories. Thus, the HA-E3-1 model has been shown to severely outperform predictive tools that are supported by some physical theories.

In the test that compared the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models using the nonaqueous and aqueous subset of data (subset 1), we have shown that the SAFT-γ Mie model has the best overall performance. However, when the unsigned errors were categorised into box plots, we found that each of the three models excelled at different bonding pairs, solute classes or solvent classes. For example, the modUNIFAC (Do) and COSMO-SAC models exhibited poor performance for alkane/water systems with some errors exceeding 2 kcal mol^{-1}. In contrast, the modelling of aromatic molecules using the SAFT-γ Mie model was challenging with some errors and outliers exceeding 1 kcal mol^{-1}. Further, all three models had a poor performance for acids as solutes and alkanes as solvents with errors surpassing 1 kcal mol^{-1}. It is important for a reader to discern what solute/solvent systems they intend to model and select a model with the lowest errors. The HA-E3-1, modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models generally have median values or mean unsigned errors of 0.5 kcal mol^{-1}, which is substantially lower than the experimental precision making them excellent predictive tools for the prediction of solvation free energies.

This thesis uses several methodologies to develop, test and benchmark predictive tools to estimate the free energy of solvation. As such, this thesis serves as a guide for a potential user intending to estimate the free energy of solvation, assessing each tool depending on their intended applications. We compared four methodologies - data-driven, molecular, activity coefficient, and ab-initio - and found the HA-E3-1 data-driven model to be the best in terms of accuracy for nonaqueous systems. However, one important limitation of the HA-E3-1 hybrid QM/data-driven framework is its reliance on experimental data for the solvent properties. Pure solvent properties are readily available; however, mixed solvent information is not due to the infinite number of potential compositions, preventing the HA-E3-1 model from being applied to multiple solvent systems. Comparatively, the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models can be used to access the range of compositions due to their formulation. However, while the whole range is accessible, the lack of mixed solvent data means any predictions using the three models cannot be validated against experimental data. Further, only the HA-E3-1 and COSMO-SAC models can be used to access transition state

systems; the former requires the theoretical solute descriptors, whereas the latter requires a QM optimised structure of the transition state. Transition state properties cannot be experimentally determined; therefore, we cannot validate any predictions from the HA-E3-1 and COSMO-SAC models.

Once a potential user has assessed their needs depending on intended application, this thesis's predictive tools may be accessed using the following methods. MATLAB, Python or any equivalent software can be used to implement the HA-E3-1 model using its model parameters and coefficients that are available in Table C.8 in appendix C. For the SAFT-γ Mie predictive tool, the gPROMS ModelBuilder 5.1.0 (*gPROMS* 1997) modelling suite can be used to access the model where the functional groups are readily available. The process modelling software, ASPEN Plus V8.4 (*ASPEN Plus* 1994) can be used for the modUNIFAC (Do) and the COSMO-SAC models.

Solvation free energies may thus be applied in a variety of fields, including but not limited to, drug effectiveness, environmental studies, or the estimation of reaction rates in solution chemistry. However, as mentioned in Chapter 1, it is important to take into account other thermodynamic quantities in solution chemistry, such as the infinite dilution activity coefficient, solubility, Henry's constant, and the partition coefficient of the relevant species. A user may predict the solubility of an existing or new drug in any solvent provided the required model parameters and a value for the vapour pressure are present, given that the solubility is related to the effectiveness of a drug. A series of expressions to convert predicted solvation free energies into various thermodynamic quantities are available in the database guide document for the MNSol database (Marenich et al., 2012). Marenich et al. (2012) have presented an example of obtaining the aqueous solvation free energy from the vapour pressure and solubility, expressed as:

$$\Delta G_S^* = -2.303RT \log\left(\frac{S_{aq}/M^\circ}{P_{vapour}/P^\circ}\right) \tag{6.1}$$

where S_{aq} is the aqueous solubility of a solute with units mol/L, M° is the standard state of 1 mol/L, P_{vapour} is the vapour pressure of the solute and P° is the standard state pressure of 24.45 atm.

Similarly, in environmental studies, solvation free energy can be used to calculate pollutant content inside a water stream if converted into the Henry's Law constant. Comparable to the

solubility approach, Marenich et al. (2012) have demonstrated how to obtain the aqueous solvation free energy from the Henry's Law constant and is expressed as:

$$\Delta G_S^* = -2.303RT \log\left(K_H\right) \tag{6.2}$$

where K_H is the Henry's Law constant as a dimensionless quantity. The quantity is frequently reported with differing units. When handling the data, care must be taken either when converting experimental K_H values into ΔG_S^* energies or converting predicted ΔG_S^* values into K_H values.

The final potential application of solvation free energy in estimating reaction rates is illustrated in the work of Struebing et al. (2013), where the solvation free energy of reactants and transition states are functions of the computed rate constant. The solvation free energies are necessary in calculating the activation energy barrier for the solvation free energy and is written as:

$$\Delta^{\ddagger}\Delta G_{S,B}^{\circ} = \Delta G_{S,(CD)^{\ddagger},B}^{\circ} - \Delta G_{S,C,B}^{\circ} - \Delta G_{S,D,B}^{\circ} \tag{6.3}$$

where $\Delta^{\ddagger}$ indicates that an activation energy barrier is being calculated, ΔG_S° is the standard free energy of solvation, and the subscript $(CD)^{\ddagger}$ refers to the transition state structure formed from reactants C and D while in a solvent B. The solvation free energies for reactants C and D in any solvent can be calculated as long as the required parameters are available; however, the solvation free energy of the transition state structure cannot be calculated by the modUNIFAC (Do) and SAFT-γ Mie models due to a lack of functional groups.

In contrast, the HA-E3-1 model can model transition states as long as the electronic energy of the transition state in the solvent is calculated and the corresponding solute parameters are collected. If the user wants an estimate without having to carry out QM calculations, the A-D10-10 model can model transition states in the same way provided the solute parameters are available. The optimal option for calculating solvation free energies from the models in this work is the COSMO-SAC model. It is an *ab initio* model that can model transition states using sigma profiles using QM calculations. Through the examples shown, a user can obtain a series of thermodynamic properties as shown in the documentation for the Minnesota solvation database or through the expressions found in Chapter 2.

6.0.2 Main Contributions

- A new database of solvation free energies has been compiled. The database has an array of solvation free energies $\Delta G^{\circ,m}_{s,i,j}$, electrostatic contribution values, $\Delta E^{el}_{i,j}$ and nonelectrostatic contribution values, $G^{CDS}_{i,j}$. The database can be used with the 9 QM solute descriptors and 12 bulk solvent descriptors (which can be found in (Borhani et al., 2019)) to develop data-driven models using other mathematical frameworks. These energies in the database only correspond to solute/solvent pairs at infinite dilution, 1 atm and 298 K.

- The QPLS and ALAMO methodologies were applied successfully in predicting solvation free energies. The QPLS only achieved slightly better performance compared to the PLS framework. In contrast, the ALAMO framework exhibited excellent performance when developing the pure and reduced $\Delta G^{\circ,m}_{s,i,j}$ models as well as the hybrid QM/data-driven models. The resulting ALAMO models also outperformed the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC models when compared against nonaqueous data.

- A novel methodology for producing a generalised hybrid QM/data-driven models for predicting solvation free energies was proposed. In the methodology, the solvation free energy, $\Delta G^{\circ,m}_{s,i,j}$, was partitioned into electrostatic ($\Delta E^{el}_{i,j}$) and nonelectrostatic contributions ($G^{CDS}_{i,j}$). The electrostatic contribution, $\Delta E^{el}_{i,j}$, was expressed as the difference in electronic energies of a solute in the gas solvent phases. The electronic energies were calculated using the IEF-PCM model with X3LYP/6-31G(d,p) as the level of theory. Once the electronic energies were compiled, experimental solvation free energies were converted into nonelectrostatic contributions. After, a mathematical framework (PLS, QPLS, and ALAMO in this thesis) was used to relate the property to 9 QM solute descriptors and 12 bulk solvent descriptors. Doing so would provide the overall model with an inherent QM description of the solute molecule in a solvent while also having other properties to enhance the solute and solvent descriptions. The proposed methodology was a success as the hybrid QM/data-driven models based on $G^{CDS}_{i,j}$ values achieved better RMSE values compared to a data-driven model trained on $\Delta G^{\circ,m}_{s,i,j}$ values. Further, the hybrid QM/data-driven ALAMO models had better RMSE values than the SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC by about 0.1 kcal mol^{-1}.

- A systematic study spanning four predictive tool categories was carried out against a common set of experimental data to assess which models perform the best. These categories include activity coefficient models, data-driven models, an equation of state and a hybrid QM/activity coefficient model. Several metrics and parity plots were used to assess the overall performance of each model whereas box plots with categories such as interactions between solute and solvent, solute classes, and solvent classes were used to highlight the strengths and weaknesses of each model.

- The SAFT-γ Mie equation of state was utilised in the prediction of solvation free energies. The equation of state achieved good results when compared to the modUNIFAC (Do) and COSMO-SAC models. The latter models are popular in the prediction of infinite dilution activity coefficients and solvation free energies. When compared against the nonaqueous subset of data, the SAFT-γ Mie model performed similarly to the COSMO-SAC model. However, when compared against the nonaqueous and aqueous subset of data, the SAFT-γ Mie achieved the best overall performance with the majority of the errors remaining under 0.5 kcal mol^{-1}.

- A potential user interested in modelling the free energy of solvation for obtaining other thermodynamic quantities such as the partition coefficient, reaction rates, or for another purpose, they can consult this work to aid in selecting a predictive model for the free energy of solvation.

6.0.3 Future work

Expanding the database of solvation free energies and solute/solvent descriptors

Currently, the database used in this thesis is composed of experimental solvation free energies, electronic energies, nonelectrostatic contribution values, 9 QM solute descriptors and 12 bulk solvent descriptors for each data point. A simple way to improve data-driven models is to utilise more experimental data for model training and validation. There are sources of experimental solvation free energies available such as the free energy of self-solvation (the energy required to transfer a solute molecule i from the gas phase into a solvent i), which can be found in the CompSol database (Moine et al., 2017). Further, new solute or solvent descriptors can be compiled and added to the database. These descriptors can then be used in mathematical frameworks to produce higher quality models.

Improving the hybrid QM/data-driven methodology

The hybrid QM/data-driven methodology of incorporating a detailed QM description of the solute molecule in a solvent was shown to have excellent results for the ALAMO series of models, which consistently outperformed predictive tools such as SAFT-γ Mie, modUNIFAC (Do) and COSMO-SAC. The goal of the method was to expand the generalised form of data-driven solvation models developed by Borhani et al. (2019), which directly related the solvation free energy to the solute and solvent descriptors. A drawback of that formulation was that the solute and solvent descriptors were independent of one another. Therefore, the hybrid methodology improved the formulation by incorporating information on how the solute and solvent molecules interact together. It has been shown that this new formulation has achieved better results compared to the pure data-driven form. However, there are ways to improve the formulation. For example, the electronic energies can be calculated using a higher level of theory which would provide higher precision. Further, the $G_{i,j}^{CDS}$ variable can be related to the solute and solvent descriptors by using more complicated mathematical frameworks such as neural networks. However, the ALAMO framework is highly customisable, so the predicted data would fit the experimental data as closely as possible. Another avenue for improvement is using more descriptor variables or considering different combinations of mathematical functions to optimise the hybrid model. A substantial improvement involves incorporating the solute structure in the formulation to obtain a more detailed description. This approach is similar to the one proposed in the SMD model (Marenich, Cramer, and Truhlar, 2009).

Broader systematic studies

One of the main goals of this thesis was achieved when predictive tools spanning multiple categories were compared against one another using a nonaqueous subset of data and a non-aqueous and aqueous subset of data. The systematic study is meant to serve as a guide for potential users when selecting a predictive tool. However, the study can be expanded by incorporating models from other categories such as classical molecular dynamics or Monte Carlo models or QM models such as SMD or COSMO-RS. Further, for the newer SAFT-γ Mie equation of state, functional groups are still being developed and can be used to create

more solute/solvent systems. Otherwise, the validation data set can be expanded by including more experimental data.

Solvation models for mixed solvents

An ultimate goal for predictive tools for the free energy of solvation is to have models that can accurately predict the behaviour of mixed solvents over a broad range of compositions and combinations. This aspect would open multiple avenues for computer-aided molecular design or the design of drug molecules. For data-driven models like the ones developed in this thesis, it is substantially challenging to develop models using current methodologies as there is a severe lack of experimental data of mixed solvents due to the range of compositions and combinations. Thus, it is highly unlikely that pure data-driven models can ever model mixed solvents. In contrast, models with physical theory supporting them have been shown to model multicomponent systems. Equations of state such as the SAFT-γ Mie model are state-of-the-art predictive tools that can accurately model these systems with high precision. The final issue with mixed solvent solvation models is even if there was a high precision model, there would be almost no experimental solvation free energy data for validation. However, it would be an interesting avenue to explore nonetheless.

Bibliography

Abraham, Michael H. (1993). "Hydrogen bonding. 31. Construction of a scale of solute effective or summation hydrogen-bond basicity". In: *Journal of Physical Organic Chemistry* 6.12, pp. 660–684. ISSN: 0894-3230. DOI: 10.1002/poc.610061204. URL: https://onlinelibrary.wiley.com/doi/10.1002/poc.610061204.

Abraham, Michael H. et al. (1987a). "Linear solvation energy relationships. Part 37. An analysis of contributions of dipolarity-polarisability, nucleophilic assistance, electrophilic assistance, and cavity terms to solvent effects on t-butyl halide solvolysis rates". In: *Journal of the Chemical Society, Perkin Transactions 2* 7, p. 913. ISSN: 0300-9580. DOI: 10.1039/p29870000913. URL: http://xlink.rsc.org/?DOI=P29870000913.

— (1987b). "Linear solvation energy relationships. Part 38. An analysis of the use of solvent parameters in the correlation of rate constants, with special reference to the solvolysis of t-butyl chloride". In: *Journal of the Chemical Society, Perkin Transactions 2* 78.8, p. 1097. ISSN: 0300-9580. DOI: 10.1039/p29870001097. URL: http://xlink.rsc.org/?DOI=p29870001097.

— (1987c). "Reference t o the Solvolysis of t-Butyl Chloride". In: pp. 1097–1101.

Abraham, Michael H. et al. (1990). "Thermodynamics of solute transfer from water to hexadecane". In: *Journal of the Chemical Society, Perkin Transactions 2* 2, p. 291. ISSN: 0300-9580. DOI: 10.1039/p29900000291. URL: http://xlink.rsc.org/?DOI=p29900000291.

Ahlers, Jens and Jürgen Gmehling (2001). "Development of an universal group contribution equation of stateI. Prediction of liquid densities for pure compounds with a volume translated Peng-Robinson equation of state". In: *Fluid Phase Equilibria* 191.1-2, pp. 177–188. ISSN: 03783812. DOI: 10.1016/S0378-3812(01)00626-4.

— (2002). "Development of a Universal Group Contribution Equation of State. 2. Prediction of Vapor-Liquid Equilibria for Asymmetric Systems". In: *Industrial & Engineering Chemistry Research* 41.14, pp. 3489–3498. ISSN: 0888-5885. DOI: 10.1021/ie020047o. URL: https://pubs.acs.org/doi/10.1021/ie020047o.

Akaike, Hirotugu (1974). "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723. ISSN: 15582523. DOI: 10.1109/TAC.1974.1100705.

Ashcraft, Robert W., Sumathy Raman, and William H. Green (2007). "Ab initio aqueous thermochemistry: Application to the oxidation of hydroxylamine in nitric acid solution". In: *Journal of Physical Chemistry B* 111.41, pp. 11968–11983. ISSN: 15206106. DOI: 10.1021/jp073539t.

ASPEN Plus (1994). URL: https://www.aspentech.com/products/aspen-plus.aspx.

Baffi, G., E. B. Martin, and A. J. Morris (1999). "Non-linear projection to latent structures revisited: The quadratic PLS algorithm". In: *Computers and Chemical Engineering* 23.3, pp. 395–411. ISSN: 00981354. DOI: 10.1016/S0098-1354(98)00283-X.

Bannan, Caitlin C. et al. (2016). "Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water". In: *Journal of Chemical Theory and Computation* 12.8, pp. 4015–4024. ISSN: 15499626. DOI: 10.1021/acs.jctc.6b00449. arXiv: 15334406.

Bastos, J. C., M. E. Soares, and A. G. Medina (1988). "Infinite Dilution Activity Coefficients Predicted by UNIFAC Group Contribution". In: *Industrial and Engineering Chemistry Research* 27.7, pp. 1269–1277. ISSN: 15205045. DOI: 10.1021/ie00079a030.

Becke, A. D. (1988). "Density-functional exchange-energy approximation with correct asymptotic behavior". In: *Physical Review A* 38.6, pp. 3098–3100. ISSN: 0556-2791. DOI: 10.1103/PhysRevA.38.3098. URL: https://link.aps.org/doi/10.1103/PhysRevA.38.3098.

Ben-Naim, A (2006). *Molecular Theory of Solutions*. OUP Oxford. ISBN: 9780199299690. URL: https://books.google.co.uk/books?id=Q3cTDAAAQBAJ.

Borhani, Tohid N. et al. (2019). "Hybrid QSPR models for the prediction of the free energy of solvation of organic solute/solvent pairs". In: *Physical Chemistry Chemical Physics* 21.25, pp. 13706–13720. ISSN: 14639076. DOI: 10.1039/c8cp07562j.

Borhani, Tohid Nejad Ghaffar, Afsaneh Afzali, and Mehdi Bagheri (2016). "QSPR estimation of the auto-ignition temperature for pure hydrocarbons". In: *Process Safety and Environmental Protection* 103, pp. 115–125. ISSN: 09575820. DOI: 10.1016/j.psep.2016.07.004. URL: http://dx.doi.org/10.1016/j.psep.2016.07.004.

Borhani, Tohid Nejad Ghaffar, Mehdi Bagheri, and Zainuddin A. Manan (2013). "Molecular modeling of the ideal gas enthalpy of formation of hydrocarbons". In: *Fluid Phase Equilibria*

360, pp. 423–434. ISSN: 03783812. DOI: 10.1016/j.fluid.2013.09.066. URL: http://dx.doi.org/10.1016/j.fluid.2013.09.066.

Born, M. (1936). "On the Linearization of the Energy Density of the Electromagnetic Field". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 32.1, pp. 102–107. ISSN: 0305-0041. DOI: 10.1017/S0305004100018892. URL: https://www.cambridge.org/core/product/identifier/S0305004100018892/type/journal_article.

Born, Max (1920). "Volumen und Hydratationswärme der Ionen". In: *Zeitschrift für Physik* 1.1, pp. 45–48. ISSN: 1434-6001. DOI: 10.1007/BF01881023. URL: http://link.springer.com/10.1007/BF01881023.

Burger, Jakob et al. (2015). "A hierarchical method to integrated solvent and process design of physical CO2 absorption using the SAFT-γ Mie approach". In: *AIChE Journal* 61.10, pp. 3249–3269. ISSN: 15475905. DOI: 10.1002/aic.14838. arXiv: arXiv:1402.6991v1.

Buttery, Ron G., Louisa C. Ling, and D. G. Guadagni (1969). "Food Volatiles Volatilities of Aldehydes, Ketones, and Esters in Dilute Water Solution". In: *Journal of Agricultural and Food Chemistry* 17.2, pp. 385–389. ISSN: 15205118. DOI: 10.1021/jf60162a025.

Cancès, E., B. Mennucci, and J. Tomasi (1997). "A new integral equation formalism for the polarizable continuum model: Theoretical background and applications to Isotropic and anisotropic dielectrics". In: *Journal of Chemical Physics* 107.8, pp. 3032–3041. ISSN: 00219606. DOI: 10.1063/1.474659.

Chapman, W. G. et al. (1989). "SAFT: Equation-of-state solution model for associating fluids". In: *Fluid Phase Equilibria* 52.C, pp. 31–38. ISSN: 03783812. DOI: 10.1016/0378-3812(89)80308-5.

Chapman, Walter G. et al. (1990). "New reference equation of state for associating liquids". In: *Industrial & Engineering Chemistry Research* 29.8, pp. 1709–1721. ISSN: 0888-5885. DOI: 10.1021/ie00104a021. URL: http://pubs.acs.org/doi/abs/10.1021/ie00104a021.

Cherkassky, Vladimir, Don Gehring, and Filip Mulier (1996). "Comparison of adaptive methods for function estimation from samples". In: *IEEE Transactions on Neural Networks* 7.4, pp. 969–984. ISSN: 10459227. DOI: 10.1109/72.508939.

Cozad, Alison, Nikolaos V. Sahinidis, and David C. Miller (2014). "Learning surrogate models for simulation-based optimization". In: *AIChE Journal* 60.6, pp. 2211–2227. ISSN: 00011541. DOI: 10.1002/aic.14418. arXiv: 0201037v1 [arXiv:physics]. URL: http://doi.wiley.com/10.1002/aic.14418.

Cozad, Alison, Nikolaos V. Sahinidis, and David C. Miller (2015). "A combined first-principles and data-driven approach to model building". In: *Computers and Chemical Engineering* 73, pp. 116–127. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2014.11.010. URL: http://dx.doi.org/10.1016/j.compchemeng.2014.11.010.

Cramer, Christopher J. (2004). *Essentials of Computational Chemistry (Second Edition) Theories and Models*. Wiley. ISBN: 0-470-09181-9.

Cramer, Christopher J., George R. Famini, and Alfred H. Lowrey (1993). "Use of Calculated Quantum Chemical Properties as Surrogates for Solvatochromic Parameters in Structure-Activity Relationships". In: *Accounts of Chemical Research* 26.11, pp. 599–605. ISSN: 15204898. DOI: 10.1021/ar00035a006.

Cramer, Christopher J. and Donald G. Truhlar (1991). "General Parameterized SCF Model for Free Energies of Solvation in Aqueous Solution". In: *Journal of the American Chemical Society* 113.22, pp. 8305–8311. ISSN: 15205126. DOI: 10.1021/ja00022a017.

— (1999). "Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics". In: *Chemical Reviews* 99.8, pp. 2161–2200. ISSN: 00092665. DOI: 10.1021/cr960149m.

— (2006). "SMx continuum models for condensed phases". In: *Trends and Perspectives in Modern Computational Science, Maroulis, G.; Simos, TE, Eds.; Lecture Series on Computer and Computational Sciences* 6.1, pp. 112–140. DOI: doi:10.1201/b12251-8. URL: http://static.msi.umn.edu/rreports/2006/158.pdf.

— (2008). "A Universal Approach to Solvation Modeling". In: *Acc. Chem. Res.* 41.6, pp. 760–768. ISSN: 1520-4898. DOI: 10.1021/ar800019z.

DDBST GmbH (2018). *Dortmund Data Bank*.

Delgado, Eduardo J and Gonzalo a Jaña (2009). "Quantitative Prediction of Solvation Free Energy in Octanol of Organic Compounds". In: *Internet Electronic Journal of Molecular Design* 10, pp. 1031–1044. DOI: 10.3390/ijms10031031.

Deublein, Stephan et al. (2011). "Ms2: A molecular simulation tool for thermodynamic properties". In: *Computer Physics Communications* 182.11, pp. 2350–2367. ISSN: 00104655. DOI: 10.1016/j.cpc.2011.04.026. URL: http://dx.doi.org/10.1016/j.cpc.2011.04.026.

Duarte Ramos Matos, Guilherme et al. (2017). "Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database". In: *Journal of Chemical and Engineering Data* 62.5, pp. 1559–1569. ISSN: 15205134. DOI: 10.1021/acs.jced.7b00104.

Dufal, Simon et al. (2014). "Prediction of Thermodynamic Properties and Phase Behavior of Fluids and Mixtures with the SAFT-γ Mie Group-Contribution Equation of State". In: *Journal of Chemical & Engineering Data* 59.10, pp. 3272–3288. DOI: 10.1021/je500248h.

Dunning, Thom H. (1989). "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen". In: *The Journal of Chemical Physics* 90.2, pp. 1007–1023. ISSN: 0021-9606. DOI: 10.1063/1.456153. URL: http://aip.scitation.org/doi/10.1063/1.456153.

Easton, R. Evan et al. (1996). "The MIDI! basis set for quantum mechanical calculations of molecular geometries and partial charges". In: *Theoretical Chemistry Accounts* 93.5, pp. 281–301. ISSN: 1432881X. DOI: 10.1007/s002140050153.

Famini, George R. and Leland Y. Wilson (1993). "Using theoretical descriptors in structure-activity relationships: Solubility in supercritical CO2". In: *Journal of Physical Chemistry* 6.10, pp. 539–544. DOI: 10.1002/poc.610061002.

— (1999). "Using theoretical descriptors in linear free energy relationships: Characterizing several polarity, acid and basicity scales". In: *Journal of Physical Organic Chemistry* 12.8, pp. 645–653. ISSN: 08943230. DOI: 10.1002/(SICI)1099-1395(199908)12:8<645::AID-POC165>3.0.CO;2-S.

Fingerhut, Robin et al. (2017). "Comprehensive Assessment of COSMO-SAC Models for Predictions of Fluid-Phase Equilibria". In: *Industrial and Engineering Chemistry Research* 56.35, pp. 9868–9884. ISSN: 15205045. DOI: 10.1021/acs.iecr.7b01360.

Floris, F. and J. Tomasi (1989). "Evaluation of the dispersion contribution to the solvation energy. A simple computational model in the continuum approximation". In: *Journal of Computational Chemistry* 10.5, pp. 616–627. ISSN: 0192-8651. DOI: 10.1002/jcc.540100504. URL: http://doi.wiley.com/10.1002/jcc.540100504.

Foster, Dean P. and Edward I. George (1994). "The Risk Inflation Criterion for Multiple Regression". In: *The Annals of Statistics* 22.4, pp. 1947–1975.

Fredenslund, Aage, Russell L. Jones, and John M. Prausnitz (1975). "Group-contribution estimation of activity coefficients in nonideal liquid mixtures". In: *AIChE Journal* 21.6, pp. 1086–1099. ISSN: 15475905. DOI: 10.1002/aic.690210607.

Fredenslund, Aage et al. (1977). "Computerized Design of Multicomponent Distillation Columns Using the UNIFAC Group Contribution Method for Calculation of Activity Coefficients". In: *Industrial & Engineering Chemistry Process Design and Development* 16.4, pp. 450–

462. ISSN: 0196-4305. DOI: 10.1021/i260064a004. URL: http://pubs.acs.org/doi/abs/10.1021/i260064a004.

Frisch, M. J. et al. (2004). *Gaussian 03, Revision C.02*. Wallingford CT.

Frisch, M. J. et al. (2016). *Gaussian 09, Revision D.01*. Wallingford CT.

García-Muñoz, Salvador (2020). *Personal Communication*.

Geladi, Paul and Bruce R. Kowalski (1986). "Partial least-squares regression: a tutorial". In: *Analytica Chimica Acta* 185.9, pp. 1–17. ISSN: 00032670. DOI: 10.1016/0003-2670(86)80028-9. URL: https://linkinghub.elsevier.com/retrieve/pii/0003267086800289.

Glass, Colin W. et al. (2014). "Ms 2: A molecular simulation tool for thermodynamic properties, new version release". In: *Computer Physics Communications* 185.12, pp. 3302–3306. ISSN: 00104655. DOI: 10.1016/j.cpc.2014.07.012. arXiv: 1507.07548. URL: http://dx.doi.org/10.1016/j.cpc.2014.07.012.

Gmehling, Jürgen, Jiding Li, and Martin Schiller (1993). "A Modified UNIFAC Model. 2. Present Parameter Matrix and Results for Different Thermodynamic Properties". In: *Industrial and Engineering Chemistry Research* 32.1, pp. 178–193. ISSN: 15205045. DOI: 10.1021/ie00013a024.

Gmehling, Jürgen et al. (1998). "A Modified UNIFAC (Dortmund) Model. 3. Revision and Extension". In: *Industrial & Engineering Chemistry Research* 37.12, pp. 4876–4882. ISSN: 0888-5885. DOI: 10.1021/ie980347z. URL: http://pubs.acs.org/doi/abs/10.1021/ie980347z.

Gmehling, Jürgen et al. (2002). "A modified UNIFAC (Dortmund) model. 4. Revision and extension". In: *Industrial and Engineering Chemistry Research* 41.6, pp. 1678–1688. ISSN: 08885885. DOI: 10.1021/ie0108043.

gPROMS (1997). URL: www.psenterprise.com/products/gproms.

Grant, Eliana (2019). "Optimisan of SNAr reaction; mechanism, kinetics, solvent design". PhD thesis. Imperial College London.

Gros, H. P., S. Bottini, and E. A. Brignole (1996). "A group contribution equation of state for associating mixtures". In: *Fluid Phase Equilibria* 116.1-2, pp. 537–544. ISSN: 03783812. DOI: 10.1016/0378-3812(95)02928-1.

Gros, H. P., S. B. Bottini, and E. A. Brignole (1997). "High pressure phase equilibrium modeling of mixtures containing associating compounds and gases". In: *Fluid Phase Equilibria* 139.1-2, pp. 75–87. ISSN: 03783812. DOI: 10.1016/s0378-3812(97)00099-x.

Guggenheim, Edward Armand (1952). *Mixtures: the theory of the equilibrium properties of some simple classes of mixtures and alloys*. Clarendon Press.

Guthrie, J. Peter (2009). "A blind challenge for computational solvation free energies: Introduction and overview". In: *Journal of Physical Chemistry B* 113.14, pp. 4501–4507. ISSN: 15206106. DOI: 10.1021/jp806724u.

Guthrie, J. Peter and Igor Povar (2009). "A test of various computational solvation models on a set of "difficult" organic compounds". In: *Canadian Journal of Chemistry* 87.8, pp. 1154–1162. ISSN: 0008-4042. DOI: 10.1139/v09-071.

Hannan, E. J. and B. G. Quinn (1979). "The Determination of the Order of an Autoregression". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2, pp. 190–195. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1979.tb01072.x. URL: http://doi.wiley.com/10.1111/j.2517-6161.1979.tb01072.x.

Hansen, Henrik K. et al. (1991). "Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension". In: *Industrial and Engineering Chemistry Research* 30.10, pp. 2352–2355. ISSN: 15205045. DOI: 10.1021/ie00058a017.

Haslam, Andrew J. et al. (2020). "Expanding the Applications of the SAFT-γMie Group-Contribution Equation of State: Prediction of Thermodynamic Properties and Phase Behavior of Mixtures". In: *Journal of Chemical and Engineering Data* 65.12, pp. 5862–5890. ISSN: 15205134. DOI: 10.1021/acs.jced.0c00746.

Hehre, W.J. et al. (1986). *Ab Initio Molecular Orbital Theory*. Vol. 7. 3. New York: Wiley, pp. 379–379. DOI: 10.1002/jcc.540070314. URL: http://doi.wiley.com/10.1002/jcc.540070314.

Hine, Jack and Pradip K. Mookerjee (1975). "The Intrinsic Hydrophilic Character of Organic Compounds. Correlations in Terms of Structural Contributions". In: *Journal of Organic Chemistry* 40.3, pp. 292–298. ISSN: 15206904. DOI: 10.1021/jo00891a006.

Ho, Junming, Andreas Klamt, and Michelle L. Coote (2010). "Comment on the correct use of continuum solvent models". In: *Journal of Physical Chemistry A* 114.51, pp. 13442–13444. ISSN: 10895639. DOI: 10.1021/jp107136j. arXiv: arXiv:1011.1669v3.

Holderbaum, T. and J. Gmehling (1991). "PSRK: A Group Contribution Equation of State Based on UNIFAC". In: *Fluid Phase Equilibria* 70.2-3, pp. 251–265. ISSN: 03783812. DOI: 10.1016/0378-3812(91)85038-V.

Hooper, Herbert H., Stefan Michel, and John M. Prausnitz (1988). "Correlation of liquid-liquid equilibria for some water-organic liquid systems in the region 20-250.degree.C". In: *Industrial & Engineering Chemistry Research* 27.11, pp. 2182–2187. ISSN: 0888-5885. DOI: 10.1021/ie00083a039. URL: https://pubs.acs.org/doi/abs/10.1021/ie00083a039.

Hsieh, Chieh Ming and Shiang Tai Lin (2008). "Determination of cubic equation of state parameters for pure and mixture fluids from first principle solvation calculations". In: 54. AIChE Journal, p. 2174.

— (2009). "First-principles predictions of vapor-liquid equilibria for pure and mixture fluids from the combined use of cubic equations of state and solvation calculations". In: *Industrial and Engineering Chemistry Research* 48.6, pp. 3197–3205. ISSN: 08885885. DOI: 10.1021/ie801118a.

— (2012). "First-principles prediction of phase equilibria using the PR+COSMOSAC equation of state". In: *Asia-Pacific Journal of Chemical Engineering* 7.March, S1–S10. ISSN: 19322135. DOI: 10.1002/apj.608. URL: http://doi.wiley.com/10.1002/apj.608.

Hsieh, Chieh Ming, Shiang Tai Lin, and Jadran Vrabec (2014). "Considering the dispersive interactions in the COSMO-SAC model for more accurate predictions of fluid phase behavior". In: *Fluid Phase Equilibria* 367.April, pp. 109–116. ISSN: 03783812. DOI: 10.1016/j.fluid.2014.01.032. URL: http://dx.doi.org/10.1016/j.fluid.2014.01.032.

Hsieh, Chieh Ming, Stanley I. Sandler, and Shiang Tai Lin (2010). "Improvements of COSMO-SAC for vapor-liquid and liquid-liquid equilibrium predictions". In: *Fluid Phase Equilibria* 297.1, pp. 90–97. ISSN: 03783812. DOI: 10.1016/j.fluid.2010.06.011.

Hsieh, Chieh Ming et al. (2011). "A predictive model for the solubility and octanol-water partition coefficient of pharmaceuticals". In: *Journal of Chemical and Engineering Data* 56.4, pp. 936–945. ISSN: 00219568. DOI: 10.1021/je1008872.

Huron, Marie J. and Pierre Claverie (1974a). "Calculation of the interaction energy of one molecule with its whole surrounding. II. Method of calculating electrostatic energy". In: *The Journal of Physical Chemistry* 78.18, pp. 1853–1861. ISSN: 0022-3654. DOI: 10.1021/j100611a018. URL: https://pubs.acs.org/doi/10.1021/j100611a018.

— (1974b). "Calculation of the interaction energy of one molecule with its whole surrounding. III. Application to pure polar compounds". In: *The Journal of Physical Chemistry* 78.18, pp. 1862–1867. ISSN: 0022-3654. DOI: 10.1021/j100611a019. URL: https://pubs.acs.org/doi/10.1021/j100611a019.

Huron, MarieJose and Pierre Claverie (1972). “Calculation of the Interaction Energy of”. In: *J. Phys. Chem.* 76.15, pp. 2123–2133.

Hutacharoen, P. (2017). “Prediction of Partition Coefficients and Solubilities of Active Pharmaceutical Ingredients with the SAFT-γ Mie Group-contribution Approach”. In: *Department of Chemical Engineering* May.

Jakob, Antje et al. (2006). “Further Development of Modified UNIFAC (Dortmund): Revision and Extension 5”. In: *Industrial & Engineering Chemistry Research* 45.23, pp. 7924–7933. ISSN: 0888-5885. DOI: 10.1021/ie060355c. URL: http://pubs.acs.org/doi/abs/10.1021/ie060355c.

Jalan, Amrit et al. (2010). “Predicting solvation energies for kinetic modeling”. In: *Annual Reports Section "C" (Physical Chemistry)* 106, p. 211. ISSN: 0260-1826. DOI: 10.1039/b811056p. URL: http://xlink.rsc.org/?DOI=b811056p.

Jaubert, Jean Noël and Fabrice Mutelet (2004). “VLE predictions with the Peng-Robinson equation of state and temperature dependent kij calculated through a group contribution method”. In: *Fluid Phase Equilibria* 224.2, pp. 285–304. ISSN: 03783812. DOI: 10.1016/j.fluid.2004.06.059.

Jensen, Frank (2007). *Introduction to Computational Chemistry*. 2nd ed. Wiley. ISBN: 9780470011867.

Karelson, Mati, Victor S. Lobanov, and Alan R. Katritzky (1996). “Quantum-Chemical Descriptors in QSAR/QSPR Studies”. In: *Chemical Reviews* 96.3, pp. 1027–1044. ISSN: 0009-2665. DOI: 10.1021/cr950202r. URL: http://pubs.acs.org/doi/abs/10.1021/cr950202r.

Katritzky, Alan R. et al. (2003a). “A General Treatment of Solubility. 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series of Solvents”. In: *Journal of Chemical Information and Computer Sciences* 43.6, pp. 1794–1805. ISSN: 00952338. DOI: 10.1021/ci034120c.

— (2003b). “A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents”. In: *Journal of Chemical Information and Computer Sciences* 43.6, pp. 1806–1814. ISSN: 00952338. DOI: 10.1021/ci034122x.

Katritzky, Alan R. et al. (2010). “Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction”. In: *Chemical reviews* 110.10, pp. 5714–5789. URL: http://pubs.acs.org/doi/pdf/10.1021/cr900238d.

Kirkwood, John G. (1934). "Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions". In: *The Journal of Chemical Physics* 2.7, pp. 351–361. ISSN: 0021-9606. DOI: 10.1063/1.1749489. URL: http://aip.scitation.org/doi/10.1063/1.1749489.

Klamt, A. and G. Schüürmann (1993). "COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient". In: *Journal of the Chemical Society, Perkin Transactions 2* 5, pp. 799–805. ISSN: 1472779X. DOI: 10.1039/P29930000799.

Klamt, Andreas (2011). "The COSMO and COSMO-RS solvation models". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5, pp. 699–709. ISSN: 17590884. DOI: 10.1002/wcms.56.

— (2016). "COSMO-RS for aqueous solvation and interfaces". In: *Fluid Phase Equilibria* 407, pp. 152–158. ISSN: 03783812. DOI: 10.1016/j.fluid.2015.05.027. URL: http://dx.doi.org/10.1016/j.fluid.2015.05.027.

— (2018). "The COSMO and COSMO-RS solvation models". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 8.1, e1338. ISSN: 17590876. DOI: 10.1002/wcms.1338. URL: http://doi.wiley.com/10.1002/wcms.1338.

Klamt, Andreas and Michael Diedenhofen (2015). "Calculation of solvation free energies with DCOSMO-RS". In: *Journal of Physical Chemistry A* 119.21, pp. 5439–5445. ISSN: 15205215. DOI: 10.1021/jp511158y.

Lafitte, Thomas et al. (2013). "Accurate statistical associating fluid theory for chain molecules formed from Mie segments". In: *J Chem Phys* 139.15, p. 154504. DOI: 10.1063/1.4819786. URL: https://www.ncbi.nlm.nih.gov/pubmed/24160524.

Larsen, Bent L., Peter Rasmussen, and Aage Fredenslund (1987). "A Modified UNIFAC Group-Contribution Model for Prediction of Phase Equilibria and Heats of Mixing". In: *Industrial and Engineering Chemistry Research* 26.11, pp. 2274–2286. ISSN: 15205045. DOI: 10.1021/ie00071a018.

Lee, B. and F. M. Richards (1971). "The interpretation of protein structures: Estimation of static accessibility". In: *Journal of Molecular Biology* 55.3. ISSN: 00222836. DOI: 10.1016/0022-2836(71)90324-X.

Lee, Chengteh, Weitao Yang, and Robert G. Parr (1988). "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density". In: *Physical Review*

B 37.2, pp. 785–789. ISSN: 0163-1829. DOI: 10.1103/PhysRevB.37.785. URL: https://link.aps.org/doi/10.1103/PhysRevB.37.785.

Li, Jiabo, Christopher J. Cramer, and Donald G. Truhlar (1998). "MIDI! basis set for silicon, bromine, and iodine". In: *Theoretical Chemistry Accounts* 99.3, pp. 192–196. ISSN: 1432881X. DOI: 10.1007/s002140050323.

Lin, Shiang Tai and Stanley I. Sandler (1999). "Infinite dilution activity coefficients from Ab initio solvation calculations". In: *AIChE Journal* 45.12, pp. 2606–2618. ISSN: 00011541. DOI: 10.1002/aic.690451217.

— (2002). "A Priori Phase Equilibrium Prediction from a Segment Contribution Solvation Model". In: *Industrial & Engineering Chemistry Research* 41.5, pp. 899–913. ISSN: 0888-5885. DOI: 10.1021/ie001047w. URL: http://pubs.acs.org/doi/abs/10.1021/ie001047w.

Lin, Shiang Tai et al. (2004). "Prediction of vapor pressures and enthalpies of vaporization using a COSMO solvation model". In: *Journal of Physical Chemistry A* 108.36, pp. 7429–7439. ISSN: 10895639. DOI: 10.1021/jp048813n.

Lin, Shiang Tai et al. (2011). "Prediction of miscibility gaps in water/ether mixtures using COSMO-SAC model". In: *Fluid Phase Equilibria* 310.1-2, pp. 19–24. ISSN: 03783812. DOI: 10.1016/j.fluid.2011.06.015. URL: http://dx.doi.org/10.1016/j.fluid.2011.06.015.

Lipinski, Christopher A. et al. (1997). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". In: *Advanced Drug Delivery Reviews* 23.SUPPL. Pp. 3–25. ISSN: 0169409X. DOI: 10.1016/j.addr.2012.09.019.

Lowrey, Alfred H. et al. (1995). "Quantum chemical descriptors for linear solvation energy relationships". In: *Computers & Chemistry* 19.3, pp. 209–215. ISSN: 00978485. DOI: 10.1016/0097-8485(94)00058-M.

Lymperiadis, Alexandros et al. (2007). "A group contribution method for associating chain molecules based on the statistical associating fluid theory (SAFT- γ)". In: *Journal of Chemical Physics* 127.23. ISSN: 00219606. DOI: 10.1063/1.2813894.

Lymperiadis, Alexandros et al. (2008). "A generalisation of the SAFT-γ group contribution method for groups comprising multiple spherical segments". In: *Fluid Phase Equilibria* 274.1-2, pp. 85–104. ISSN: 03783812. DOI: 10.1016/j.fluid.2008.08.005.

Mackay, Donald and Wan Ying Shiu (1981). "A critical review of Henry's law constants for chemicals of environmental interest". In: *Journal of Physical and Chemical Reference Data* 10.4, pp. 1175–1199. ISSN: 15297845. DOI: 10.1063/1.555654.

Magnussen, Thomas, Peter Rasmussen, and Aage Fredenslund (1981). "Unifac Parameter Table for Prediction of Liquid-Liquid Equilibria". In: *Industrial and Engineering Chemistry Process Design and Development* 20.2, pp. 331–339. ISSN: 01964305. DOI: 10.1021/i200013a024.

Mallows, C. L. (1973). "Some comments on Cp". In: *Technometrics* 15.4, pp. 661–675. ISSN: 15372723. DOI: 10.1080/00401706.1973.10489103.

Marenich, Aleksandr V., Christopher J. Cramer, and Donald G. Truhlar (2009). "Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions". In: *The Journal of Physical Chemistry B* 113.18, pp. 6378–6396. ISSN: 1520-6106. DOI: 10.1021/jp810292n. URL: http://pubs.acs.org/doi/abs/10.1021/jp810292n.

Marenich, Aleksandr V. et al. (2012). *Minnesota Solvation Database*.

McCabe, Clare, Amparo Galindo, and Peter T. Cummings (2003). "Anomalies in the Solubility of Alkanes in Near-Critical Water". In: *The Journal of Physical Chemistry B* 107, pp. 12307–12314. ISSN: 1520-6106. DOI: 10.1021/jp0352332. URL: http://pubs.acs.org/doi/abs/10.1021/jp0352332.

Mennucci, B., E. Cancès, and J. Tomasi (1997). "Evaluation of solvent effects in isotropic and anisotropic dielectrics and in ionic solutions with a unified integral equation method: Theoretical bases, computational implementation, and numerical applications". In: *Journal of Physical Chemistry B* 101.49, pp. 10506–10517. ISSN: 15206106. DOI: 10.1021/jp971959k.

Mennucci, Benedetta and Jacopo Tomasi (1997). "Continuum solvation models: A new approach to the problem of solute's charge distribution and cavity boundaries". In: *Journal of Chemical Physics* 106.12, pp. 5151–5158. ISSN: 00219606. DOI: 10.1063/1.473558.

Michielan, Lisa et al. (2008). "Prediction of the aqueous solvation free energy of organic compounds by using autocorrelation of molecular electrostatic potential surface properties combined with response surface analysis". In: *Bioorganic and Medicinal Chemistry* 16.10, pp. 5733–5742. ISSN: 09680896. DOI: 10.1016/j.bmc.2008.03.064.

Mie, Gustav (1903). "Zur kinetischen Theorie der einatomigen Körper". In: *Annalen der Physik* 316.8, pp. 657–697. DOI: 10.1002/andp.19033160802.

Miertuš, S., E. Scrocco, and J. Tomasi (1981). "Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects". In: *Chemical Physics* 55.1, pp. 117–129. ISSN: 03010104. DOI: 10.1016/0301-0104(81)85090-2.

Moine, Edouard et al. (2017). "Estimation of Solvation Quantities from Experimental Thermodynamic Data: Development of the Comprehensive CompSol Databank for Pure and Mixed Solutes". In: *Journal of Physical and Chemical Reference Data* 46.3, p. 033102. ISSN: 0047-2689. DOI: 10.1063/1.5000910. URL: http://aip.scitation.org/doi/10.1063/1.5000910.

Mullins, Eric et al. (2006). "Sigma-profile database for using COSMO-based thermodynamic methods". In: *Industrial and Engineering Chemistry Research* 45.12, pp. 4389–4415. ISSN: 08885885. DOI: 10.1021/ie060370h.

Mullins, Eric et al. (2008). "Sigma profile database for predicting solid solubility in pure and mixed solvent mixtures for organic pharmacological compounds with COSMO-based thermodynamic methods". In: *Industrial and Engineering Chemistry Research* 47.5, pp. 1707–1725. ISSN: 08885885. DOI: 10.1021/ie0711022.

Muzenda, Edison (2013). "From UNIQUAC to Modified UNIFAC Dortmund : A Discussion". In: *3rd International Conference on Medical Sciences and Chemical Engineering (ICMSCE'2013)* 5, pp. 32–41. URL: http://psrcentre.org/images/extraimages/81213836.pdfhttp://psrcentre.org/images/extraimages/81213836.pdf.

Nait Saidi, Chourouk, Detlev Conrad Mielczarek, and Patrice Paricaud (2020). "Predictions of solvation Gibbs free energies with COSMO-SAC approaches". In: *Fluid Phase Equilibria* 517, p. 112614. ISSN: 03783812. DOI: 10.1016/j.fluid.2020.112614. URL: https://doi.org/10.1016/j.fluid.2020.112614.

Neath, Andrew A. and Joseph E. Cavanaugh (2012). "The Bayesian information criterion: Background, derivation, and applications". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2, pp. 199–203. ISSN: 19395108. DOI: 10.1002/wics.199.

Nicholls, Anthony et al. (2008). "Predicting small-molecule solvation free energies: An informal blind test for computational chemistry". In: *Journal of Medicinal Chemistry* 51.4, pp. 769–779. ISSN: 00222623. DOI: 10.1021/jm070549+.

Orozco, Modesto and F. Javier Luque (2000). "Theoretical methods for the description of the solvent effect in biomolecular systems". In: *Chemical Reviews* 100.11, pp. 4187–4225. ISSN: 00092665. DOI: 10.1021/cr990052a.

Papaioannou, Vasileios et al. (2011). "Simultaneous prediction of vapour-liquid and liquid-liquid equilibria (VLE and LLE) of aqueous mixtures with the SAFT-γ group contribution approach". In: *Fluid Phase Equilibria* 306.1, pp. 82–96. ISSN: 03783812. DOI: 10.1016/j.fluid.2011.02.016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378381211001002.

Papaioannou, Vasileios et al. (2014). "Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from Mie segments". In: *The Journal of Chemical Physics* 140.5, p. 54107. DOI: 10.1063/1.4851455. URL: https://www.ncbi.nlm.nih.gov/pubmed/24511922.

Papaioannou, Vasileios et al. (2016). "Application of the SAFT-γ Mie group contribution equation of state to fluids of relevance to the oil and gas industry". In: *Fluid Phase Equilibria* 416, pp. 104–119. DOI: 10.1016/j.fluid.2015.12.041.

Park, Young Woong and Diego Klabjan (2013). "Subset selection for multiple linear regression via optimization". PhD thesis. Northwestern University.

Peng, Yun et al. (2009). "Developing a predictive group-contribution-based SAFT-VR equation of state". In: *Fluid Phase Equilibria* 277.2, pp. 131–144. ISSN: 03783812. DOI: 10.1016/j.fluid.2008.11.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S037838120800397X.

Pierotti, G. J., C. H. Deal, and E. L. Derr (1959). "Activity Coefficients and Molecular Structure". In: *Industrial and Engineering Chemistry* 51.1, pp. 95–102. ISSN: 00197866. DOI: 10.1021/ie50589a048.

Pierotti, Robert A. (1963). "The solubility of gases in liquids". In: *Journal of Physical Chemistry* 67.9, pp. 1840–1845. ISSN: 00223654. DOI: 10.1021/j100803a024.

— (1976). "A Scaled Particle Theory of Aqueous and Nonaqueous Solutions". In: *Chemical Reviews* 76.6, pp. 717–726. ISSN: 15206890. DOI: 10.1021/cr60304a002.

Plyasunov, Andrey V. and Everett L. Shock (2000). "Thermodynamic functions of hydration of hydrocarbons at 298.15 K and 0.1 MPa". In: *Geochimica et Cosmochimica Acta* 64.3, pp. 439–468. ISSN: 00167037. DOI: 10.1016/S0016-7037(99)00330-0.

Pye, Cory C. et al. (2009). "An implementation of the conductor-like screening model of solvation within the amsterdam density functional package - Part II. COSMO for real solvents". In: *Canadian Journal of Chemistry* 87.7, pp. 790–797. ISSN: 00084042. DOI: 10.1139/V09-008.

Ramos, M. Carolina dos et al. (2011). "Extending the GC-SAFT-VR approach to associating functional groups: Alcohols, aldehydes, amines and carboxylic acids". In: *Fluid Phase Equilibria* 306.1, pp. 97–111. ISSN: 03783812. DOI: 10.1016/j.fluid.2011.03.026. URL: https://linkinghub.elsevier.com/retrieve/pii/S037838121100166X.

Reichardt, C. and T. Welton (2014). *Solvents and Solvent Effects in Organic Chemistry*, pp. 1573–1586. ISBN: 978-0-08-043408-7. DOI: 10.1016/B978-0-12-416677-6.00029-9. URL: http://linkinghub.elsevier.com/retrieve/pii/B9780124166776000299.

Renon, H and J M Prausnitz (1968). "Local compositions in thermodynamics excess functions for liquids mixtures". In: *AIChE J.* 14.1, pp. 116–128.

Ribeiro, Raphael F. et al. (2010). "Prediction of SAMPL2 aqueous solvation free energies and tautomeric ratios using the SM8, SM8AD, and SMD solvation models". In: *Journal of Computer-Aided Molecular Design* 24.4, pp. 317–333. ISSN: 0920654X. DOI: 10.1007/s10822-010-9333-9.

Sadeqzadeh, Majid et al. (2016). "The development of unlike induced association-site models to study the phase behaviour of aqueous mixtures comprising acetone, alkanes and alkyl carboxylic acids with the SAFT-γ Mie group contribution methodology". In: *Fluid Phase Equilibria* 407, pp. 39–57. ISSN: 03783812. DOI: 10.1016/j.fluid.2015.07.047. URL: http://dx.doi.org/10.1016/j.fluid.2015.07.047.

Sinnecker, Sebastian et al. (2006). "Calculation of solvent shifts on electronic g-tensors with the conductor-like screening model (COSMO) and its self-consistent generalization to real solvents (direct COSMO-RS)". In: *Journal of Physical Chemistry A* 110.6, pp. 2235–2245. ISSN: 10895639. DOI: 10.1021/jp056016z.

Siougkrou, Eirini (2014). "Systematic methods for solvent design : Towards better reactive processes". In: February.

Skjold-Jørgensen, Steen (1984). "Gas solubility calculations. II. Application of a new group-contribution equation of state". In: *Fluid Phase Equilibria* 16.3, pp. 317–351. ISSN: 03783812. DOI: 10.1016/0378-3812(84)80005-9. URL: http://www.scopus.com/inward/record.

url?eid=2-s2.0-84944731113&partnerID=tZOtx3y1https://linkinghub.elsevier.com/retrieve/pii/0378381284800059.

Skjold-Jørgensen, Steen (1988). "Group Contribution Equation of State (GC-EOS): A Predictive Method for Phase Equilibrium Computations Over Wide Ranges of Temperature and Pressures up to 30 MPa". In: *Industrial and Engineering Chemistry Research* 27.1, pp. 110–118. ISSN: 15205045. DOI: 10.1021/ie00073a021. URL: https://linkinghub.elsevier.com/retrieve/pii/0378381284800059https://pubs.acs.org/doi/abs/10.1021/ie00073a021.

Stanescu, Ioana and Luke E. K. Achenie (2006). "A theoretical study of solvent effects on Kolbe–Schmitt reaction kinetics". In: *Chemical Engineering Science* 61.18, pp. 6199–6212. ISSN: 00092509. DOI: 10.1016/j.ces.2006.05.025. URL: http://linkinghub.elsevier.com/retrieve/pii/S0009250906003356.

Staverman, A. J. (1950). "The entropy of high polymer solutions. Generalization of formulae". In: *Recueil des Travaux Chimiques des Pays-Bas*. ISSN: 01650513. DOI: 10.1002/recl.19500690203.

Stephen, P J et al. (1994). "Ab Initio Calculation of Vibrational Absorption". In: *The Journal of Physical Chemistry* 98.45, pp. 11623–11627.

Struebing, Heiko (2011). "Identifying optimal solvents for reactions using quantum mechanics and computer-aided molecular design". PhD thesis. Imperial College London.

Struebing, Heiko et al. (2013). "Computer-aided molecular design of solvents for accelerated reaction kinetics". In: *Nature Chemistry* 5.11, pp. 952–957. ISSN: 1755-4330. DOI: 10.1038/nchem.1755. URL: https://www.ncbi.nlm.nih.gov/pubmed/24153374http://www.nature.com/articles/nchem.1755.

Thompson, Jason D., Christopher J. Cramer, and Donald G. Truhlar (2004). "New universal solvation model and comparison of the accuracy of the SM5.42R, SM5.43R, C-PCM, D-PCM, and IEF-PCM continuum solvation models for aqueous and organic solvation free energies and for vapor pressures". In: *Journal of Physical Chemistry A* 108.31, pp. 6532–6542. ISSN: 10895639. DOI: 10.1021/jp0496295.

Tomasi, J, B Mennucci, and E. Cancès (1999). "The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level". In: *Journal of Molecular Structure: THEOCHEM* 464.1-3, pp. 211–226. ISSN: 01661280. DOI: 10.1016/S0166-1280(98)00553-3. URL: https://www.sciencedirect.

com/science/article/pii/S0166128098005533http://linkinghub.elsevier.com/retrieve/pii/S0166128098005533.

Tomasi, Jacopo, Benedetta Mennucci, and Roberto Cammi (2005a). "Quantum mechanical continuum solvation models". In: *Chemical Reviews* 105.8, pp. 2999–3093. ISSN: 00092665. DOI: 10.1021/cr9904009.

— (2005b). "Quantum mechanical continuum solvation models". In: *Chemical Reviews* 105.8, pp. 2999–3093. ISSN: 00092665. DOI: 10.1021/cr9904009. arXiv: cr9904009 [10.1021]. URL: http://pubs.acs.org/doi/abs/10.1021/cr9904009.

Voutsas, Epaminondas C. and Dimitrios P. Tassios (1996). "Prediction of Infinite-Dilution Activity Coefficients in Binary Mixtures with UNIFAC. A Critical Evaluation". In: *Industrial and Engineering Chemistry Research* 35.4, pp. 1438–1445. ISSN: 08885885. DOI: 10.1021/ie9503555.

Wang, Li Hsin, Chieh Ming Hsieh, and Shiang Tai Lin (2015). "Improved Prediction of Vapor Pressure for Pure Liquids and Solids from the PR+COSMOSAC Equation of State". In: *Industrial and Engineering Chemistry Research* 54.41, pp. 10115–10125. ISSN: 15205045. DOI: 10.1021/acs.iecr.5b01750.

Wang, Shu, Stanley I. Sandler, and Chau Chyun Chen (2007). "Refinement of COSMO-SAC and the applications". In: *Industrial and Engineering Chemistry Research* 46.22, pp. 7275–7288. ISSN: 08885885. DOI: 10.1021/ie070465z.

Wang, Shu, Yuhua Song, and Chau Chyun Chen (2011). "Extension of COSMO-SAC solvation model for electrolytes". In: *Industrial and Engineering Chemistry Research* 50.1, pp. 176–187. ISSN: 08885885. DOI: 10.1021/ie100689g.

Weidlich, Ulrich and Juergen Gmehling (1987). "A modified UNIFAC model. 1. Prediction of VLE, hE, and .gamma..infin." In: *Industrial & Engineering Chemistry Research* 26.7, pp. 1372–1381. ISSN: 0888-5885. DOI: 10.1021/ie00067a018. URL: https://pubs.acs.org/doi/abs/10.1021/ie00067a018.

Wertheim, M. S. (1984a). "Fluids with highly directional attractive forces. I. Statistical thermodynamics". In: *Journal of Statistical Physics* 35, pp. 19–34. ISSN: 00224715. DOI: 10.1007/BF01017362.

— (1984b). "Fluids with highly directional attractive forces. II. Thermodynamic perturbation theory and integral equations". In: *Journal of Statistical Physics* 35, pp. 35–47. ISSN: 00224715. DOI: 10.1007/BF01017363.

Wertheim, M. S. (1986a). "Fluids with highly directional attractive forces. III. Multiple attraction sites". In: *Journal of Statistical Physics* 42, pp. 459–476. ISSN: 00224715. DOI: 10.1007/BF01127721.

— (1986b). "Fluids with highly directional attractive forces. IV. Equilibrium polymerization". In: *Journal of Statistical Physics* 42, pp. 477–492. ISSN: 00224715. DOI: 10.1007/BF01127722.

Wilson, Zachary T. and Nikolaos V. Sahinidis (2017). "The ALAMO approach to machine learning". In: *Computers and Chemical Engineering* 106, pp. 785–795. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2017.02.010. arXiv: 1705.10918. URL: http://dx.doi.org/10.1016/j.compchemeng.2017.02.010.

Wittig, Roland, Jürgen Lohmann, and Jürgen Gmehling (2003). "Vapor-liquid equilibria by UNIFAC group contribution. 6. Revision and extension". In: *Industrial and Engineering Chemistry Research* 42.1, pp. 183–188. ISSN: 08885885. DOI: 10.1021/ie0205061.

Wold, Herman (1973). "Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments". In: *Multivariate Analysis-III*. P. R. Kris. Academic Press, pp. 383–407. DOI: 10.1016/b978-0-12-426653-7.50032-6.

Wold, Svante, Nouna Kettaneh-Wold, and Bert Skagerberg (1989). "Nonlinear PLS modeling". In: *Chemometrics and Intelligent Laboratory Systems* 7.1-2, pp. 53–65. ISSN: 01697439. DOI: 10.1016/0169-7439(89)80111-X. URL: https://linkinghub.elsevier.com/retrieve/pii/016974398980111X.

Xu, Xin et al. (2005). "An extended hybrid density functional (X3LYP) with improved descriptions of nonbond interactions and thermodynamic properties of molecular systems". In: *Journal of Chemical Physics* 122.1. ISSN: 00219606. DOI: 10.1063/1.1812257.

Zanith, Caroline C. and Josefredo R. Pliego (2015). "Performance of the SMD and SM8 models for predicting solvation free energy of neutral solutes in methanol, dimethyl sulfoxide and acetonitrile". In: *Journal of computer-aided molecular design* 29.3, pp. 217–224. ISSN: 15734951. DOI: 10.1007/s10822-014-9814-3.

Zhang, Jin, Badamkhatan Tuguldur, and David Van Der Spoel (2015). "Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation". In: *Journal of Chemical Information and Modeling* 55.6, pp. 1192–1201. ISSN: 15205142. DOI: 10.1021/acs.jcim.5b00106.

— (2016). "Correction: Force Field Benchmark of Organic Liquids. 2. Gibbs Energy of Solvation (J. Chem. Inf. Model. (2015) 55:6 (1192-1201) DOI: 10.1021/acs.jcim.5b00106)". In:

Journal of Chemical Information and Modeling 56.4, pp. 819–820. ISSN: 15205142. DOI: 10.1021/acs.jcim.6b00081.

Zhao, Yan, Nathan E. Schultz, and Donald G. Truhlar (2006). "Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions". In: *Journal of Chemical Theory and Computation* 2.2, pp. 364–382. ISSN: 15499618. DOI: 10.1021/ct0502763.

Appendix A

List of molecules used in the study and in the experimental subsets

This section contains lists of molecules in this study, and tables containing the molecules in each experimental data subset found in chapter 5. The list of the molecules used in this study can be found in tables A.1 and A.2 with their corresponding CAS numbers, molecule class, and molecule type. The tables containing the experimental data subset 1 can be found in tables XX and the tables containing experimental data subset 2 can be found in tables YY.

TABLE A.1: Table showing the molecules used in this study, with their corresponding CAS numbers, molecule classes and molecule interaction types as classified in section 2.2.4

Molecule name	CAS Number	Molecule Class	Molecule Type
acetone	67-64-1	acetone	E
n-pentanoic acid	109-52-4	acid	SA
n-propanoic acid	79-09-4	acid	SA
n-hexanoic acid	142-62-1	acid	SA
n-butanoic acid	107-92-6	acid	SA
n-heptanoic acid	111-14-8	acid	SA
1-nonanol	143-08-8	alcohol	SA
1-octanol	111-87-5	alcohol	SA
1-butanol	71-36-3	alcohol	SA
1-heptanol	111-70-6	alcohol	SA
1-hexanol	111-27-3	alcohol	SA
2-methyl-1-propanol	78-83-1	alcohol	SA
ethanol	64-17-5	alcohol	SA
2-propanol	67-63-0	alcohol	SA
1-pentanol	71-41-0	alcohol	SA
1-propanol	71-23-8	alcohol	SA
methanol	67-56-1	alcohol	SA
1-decanol	112-30-1	alcohol	SA
2-butanol	78-92-2	alcohol	SA
t-butanol	75-65-0	alcohol	SA
n-dodecanol	112-53-8	alcohol	SA
3-methyl-1-butanol	123-51-3	alcohol	SA
t-amyl alcohol	75-85-4	alcohol	SA
n-heptane	142-82-5	alkane	NA
n-pentane	109-66-0	alkane	NA
n-octane	111-65-9	alkane	NA
n-hexadecane	544-76-3	alkane	NA
n-nonane	111-84-2	alkane	NA
2,2,4-trimethylpentane	540-84-1	alkane	NA
n-hexane	110-54-3	alkane	NA
n-decane	124-18-5	alkane	NA

Table A.2: Table showing the molecules used in this study, with their corresponding CAS numbers, molecule classes and molecule interaction types as classified in section 2.2.4

Molecule name	CAS Number	Molecule Class	Molecule Type
n-pentadecane	629-62-9	alkane	NA
n-dodecane	112-40-3	alkane	NA
n-undecane	1120-21-4	alkane	NA
2,4-dimethylpentane	108-08-7	alkane	NA
2-methylpentane	107-83-5	alkane	NA
2,2-dimethylpropane	463-82-1	alkane	NA
2-methylpropane	75-28-5	alkane	NA
n-butane	106-97-8	alkane	NA
n-propane	74-98-6	alkane	NA
ethane	74-84-0	alkane	NA
methane	74-82-8	alkane	NA
2-methylbutane	78-78-4	alkane	NA
propene	115-07-1	alkene	NA
1-hexene	592-41-6	alkene	NA
1-pentene	109-67-1	alkene	NA
s-trans-1,3-butadiene	106-99-0	alkene	NA
1-butene	106-98-9	alkene	NA
ethene	74-85-1	alkene	NA
ethylamine	75-04-7	amino	SA
butylamine	109-73-9	amino	SA
pentylamine	110-58-7	amino	SA
propylamine	107-10-8	amino	SA
1,2-diaminoethane	107-15-3	amino	SA
methylamine	74-89-5	amino	SA
benzene	71-43-2	aromatic	NA
m-xylene	108-38-3	aromatic	NA
ethylbenzene	100-41-4	aromatic	NA
toluene	108-88-3	aromatic	NA
o-xylene	95-47-6	aromatic	NA
p-xylene	106-42-3	aromatic	NA
butylbenzene	104-51-8	aromatic	NA
isopropylbenzene	98-82-8	aromatic	NA
1,2,3-trimethylbenzene (hemellitene)	526-73-8	aromatic	NA
trimethylbenzene (hemellitene)	95-63-6	aromatic	NA
cyclohexane	110-82-7	c-alkane	NA
cyclopentane	287-92-3	c-alkane	NA
pentyl acetate	628-63-7	ester	NA
n-butyl acetate	123-86-4	ester	NA
ethyl acetate	141-78-6	ester	NA
n-propyl acetate	109-60-4	ester	NA
methyl acetate	79-20-9	ester	NA
methyl propanoate	554-12-1	ester	NA
methyl butanoate	623-42-7	ester	NA
ethyl propanoate	105-37-3	ester	NA
water	7732-18-5	water	SA

Appendix B

List of molecules used in the experimental subsets found in chapter 5

This section contains the experimental database.

TABLE B.1: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1	x	x	fenchlorphos	n-octanol	299-84-3	111-87-5	-11.69	-5.00	-6.69
2	x		triethyleneglycol	triethyleneglycol	112-27-6	112-27-6	-10.93	-	-
3	x		n-dodecanol	n-dodecanol	112-53-8	112-53-8	-10.91	-	-
4	x		triethanolamine	triethanolamine	102-71-6	102-71-6	-10.33	-	-
5	x	x	1-nonanol	1-nonanol	143-08-8	143-08-8	-9.05	-2.44	-6.61
6	x	x	4-methylphenol	1-decanol	106-44-5	112-30-1	-8.91	-3.07	-5.84
7	x	x	m-cresol	4-methyl-2-pentanone	108-39-4	108-10-1	-8.79	-3.42	-5.37
8	x	x	o-cresol	1-heptanol	95-48-7	111-70-6	-8.78	-3.19	-5.59
9	x	x	benzylalcohol	benzylalcohol	100-51-6	100-51-6	-8.61	-3.62	-4.99
10	x	x	o-cresol	1-decanol	95-48-7	112-30-1	-8.58	-2.93	-5.65
11	x	x	m-cresol	butylethanoate	108-39-4	123-86-4	-8.44	-2.72	-5.72
12	x	x	n-octanol	ethylethanoate	111-87-5	141-78-6	-8.41	-2.02	-6.39
13	x	x	n-octanol	n-octanol	111-87-5	111-87-5	-8.13	-2.29	-5.84
14	x	x	hexanoicacid	2-butanol	142-62-1	78-92-2	-8.11	-3.52	-4.59
15	x	x	n-octanol	benzene	111-87-5	71-43-2	-8.06	-1.13	-6.93
16	x	x	pentanoicacid	1-butanol	109-52-4	71-36-3	-8.02	-3.75	-4.27
17	x	x	4-methylaniline	chloroform	106-49-0	67-66-3	-8.01	-2.76	-5.25
18	x	x	1-heptanol	1-heptanol	111-70-6	111-70-6	-7.84	-2.55	-5.29
19	x	x	acetophenone	dichloroethane	98-86-2	107-06-2	-7.83	-3.44	-4.39
20	x	x	methylbenzoate	chloroform	93-58-3	67-66-3	-7.81	-2.59	-5.22
21	x		phenol	phenol	108-95-2	108-95-2	-7.79	-	-
22	x	x	nitrobenzene	chloroform	98-95-3	67-66-3	-7.78	-3.07	-4.71
23	x		tributylamine	tributylamine	102-82-9	102-82-9	-7.78	-	-
24	x	x	benzamide	n-hexane	55-21-0	110-54-3	-7.77	-2.28	-5.49
25	x	x	4-methylphenol	dichloromethane	106-44-5	75-09-2	-7.71	-3.20	-4.51
26	x	x	4-methylaniline	ethylethanoate	106-49-0	141-78-6	-7.63	-3.02	-4.61
27	x	x	acetophenone	acetophenone	98-86-2	98-86-2	-7.59	-3.70	-3.89
28	x	x	4-methylaniline	bromobenzene	106-49-0	108-86-1	-7.59	-2.91	-4.68
29	x	x	4-methylphenol	toluene	106-44-5	108-88-3	-7.56	-1.70	-5.86

TABLE B.2: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
30	x	x	o-cresol	chloroform	95-48-7	67-66-3	-7.55	-2.51	-5.04
31	x	x	naphthalene	carbontetrachloride	91-20-3	56-23-5	-7.55	-0.88	-6.67
32	x	x	hexanoicacid	chloroform	142-62-1	67-66-3	-7.51	-2.72	-4.79
33	x	x	phenol	dichloromethane	108-95-2	75-09-2	-7.50	-3.24	-4.26
34	x	x	o-nitrotoluene	carbontetrachloride	88-72-2	56-23-5	-7.49	-1.79	-5.70
35	x		triethyl_phosphate	triethyl_phosphate	78-40-0	78-40-0	-7.48	-	-
36	x	x	aniline	aniline	62-53-3	62-53-3	-7.47	-3.23	-4.24
37	x	x	naphthalene	4-methyl-2-pentanone	91-20-3	108-10-1	-7.45	-1.99	-5.46
38	x	x	o-cresol	toluene	95-48-7	108-88-3	-7.43	-1.61	-5.82
39	x		benzoyl_chloride	benzoyl_chloride	98-88-4	98-88-4	-7.38	-	-
40	x	x	2-methylaniline	n-octanol	95-53-4	111-87-5	-7.36	-3.35	-4.01
41	x	x	aniline	chloroform	62-53-3	67-66-3	-7.34	-2.84	-4.50
42	x	x	o-cresol	chlorobenzene	95-48-7	108-90-7	-7.33	-2.70	-4.63
43	x	x	n-octanol	diethylether	111-87-5	60-29-7	-7.25	-1.76	-5.49
44	x	x	4-methylaniline	carbontetrachloride	106-49-0	56-23-5	-7.24	-1.65	-5.59
45	x		trimethyl_phosphate	trimethyl_phosphate	512-56-1	512-56-1	-7.18	-	-
46	x	x	naphthalene	cyclohexane	91-20-3	110-82-7	-7.17	-0.78	-6.39
47	x	x	4-methylaniline	xylene-mixture	106-49-0	1330-20-7	-7.17	-1.77	-5.40
48	x	x	aniline	nitrobenzene	62-53-3	98-95-3	-7.15	-4.03	-3.12
49	x	x	1-heptanol	bromoform	111-70-6	75-25-2	-7.10	-1.93	-5.17
50	x	x	benzaldehyde	chloroform	100-52-7	67-66-3	-7.09	-2.67	-4.42
51	x	x	1-hexanol	1-hexanol	111-27-3	111-27-3	-7.05	-2.70	-4.35
52	x	x	hexanoicacid	carbontetrachloride	142-62-1	56-23-5	-6.99	-1.69	-5.30
53	x	x	propanoicacid	2-methyl-1-propanol	79-09-4	78-83-1	-6.98	-3.75	-3.23
54	x	x	o-cresol	chlorobenzene	95-48-7	108-90-7	-6.96	-2.70	-4.26
55	x		2-4-6-trimethylpyridine	2-4-6-trimethylpyridine	108-75-8	108-75-8	-6.92	-	-
56	x	x	1-hexanol	ethylethanoate	111-27-3	141-78-6	-6.92	-2.31	-4.61
57	x	x	m-cresol	dichloroethane	108-39-4	107-06-2	-6.91	-3.29	-3.62
58	x		3-bromopyridine	3-bromopyridine	626-55-1	626-55-1	-6.88	-	-
59	x	x	aniline	benzene	62-53-3	71-43-2	-6.88	-1.74	-5.14

TABLE B.3: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
60	x	x	aniline	chlorobenzene	62-53-3	108-90-7	-6.72	-3.05	-3.67
61	x	x	aniline	n-octanol	62-53-3	111-87-5	-6.71	-3.51	-3.20
62	x	x	o-nitrotoluene	cyclohexane	88-72-2	110-82-7	-6.71	-1.60	-5.11
63	x	x	aniline	toluene	62-53-3	108-88-3	-6.69	-1.82	-4.87
64	x	x	1-heptanol	bromobenzene	111-70-6	108-86-1	-6.68	-2.13	-4.55
65	x	x	m-cresol	benzene	108-39-4	71-43-2	-6.66	-1.63	-5.03
66	x	x	aniline	bromobenzene	62-53-3	108-86-1	-6.66	-2.99	-3.67
67	x	x	N-methylaniline	benzene	100-61-8	71-43-2	-6.64	-1.40	-5.24
68	x	x	pentylethanoate	dichloroethane	628-63-7	107-06-2	-6.64	-2.82	-3.82
69	x	x	hexanoicacid	heptane	142-62-1	142-82-5	-6.54	-1.41	-5.13
70	x	x	1-heptanol	iodobenzene	111-70-6	591-50-4	-6.53	-1.99	-4.54
71	x	x	o-nitrotoluene	n-hexadecane	88-72-2	544-76-3	-6.52	-1.63	-4.89
72	x	x	n,n-dimethylformamide	n,n-dimethylformamide	68-12-2	68-12-2	-6.47	-4.88	-1.59
73	x	x	thioanisole	n-octanol	100-68-5	111-87-5	-6.47	-3.15	-3.32
74	x	x	methyl_hexanoate	benzene	106-70-7	71-43-2	-6.47	-1.47	-5.00
75	x	x	methyl_hexanoate	carbontetrachloride	106-70-7	56-23-5	-6.39	-1.44	-4.95
76	x	x	benzonitrile	diethylether	100-47-0	60-29-7	-6.36	-3.24	-3.12
77	x	x	m-cresol	xylene-mixture	108-39-4	1330-20-7	-6.32	-1.72	-4.60
78	x	x	bromobenzene	bromobenzene	108-86-1	108-86-1	-6.29	-1.44	-4.85
79	x	x	acetophenone	cyclohexane	98-86-2	110-82-7	-6.29	-1.52	-4.77
80	x	x	2-methylaniline	heptane	95-53-4	142-82-5	-6.28	-1.32	-4.96
81	x	x	o-cresol	carbondisulfide	95-48-7	75-15-0	-6.27	-1.76	-4.51
82	x	x	anisole	chloroform	100-66-3	67-66-3	-6.24	-1.84	-4.40
83	x	x	nitrobenzene	n-hexadecane	98-95-3	544-76-3	-6.22	-1.74	-4.48
84	x	x	pentylethanoate	ethylbenzene	628-63-7	100-41-4	-6.20	-1.57	-4.63
85	x	x	1-hexanol	bromoform	111-27-3	75-25-2	-6.20	-2.03	-4.17
86	x	x	aceticacid	acetophenone	64-19-7	98-86-2	-6.20	-3.88	-2.32
87	x	x	2-heptanone	2-heptanone	110-43-0	110-43-0	-6.19	-3.09	-3.09
88	x	x	1-4-dichlorobenzene	diethylether	106-46-7	60-29-7	-6.18	-1.42	-4.76
89	x		pyrrole	pyrrole	109-97-7	109-97-7	-6.15	-	-

TABLE B.4: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
90	x	x	4-methylaniline	heptane	106-49-0	142-82-5	-6.15	-1.37	-4.78
91	x	x	2-methoxyethanol	tributylphosphate	109-86-4	126-73-8	-6.14	-2.76	-3.38
92	x		1-1-2-2-tetrachloroethane	1-1-2-2-tetrachloroethane	79-34-5	79-34-5	-6.12	-	-
93	x	x	2-heptanone	carbontetrachloride	110-43-0	56-23-5	-6.12	-1.48	-4.64
94	x	x	acetophenone	n-dodecane	98-86-2	112-40-3	-6.11	-1.51	-4.60
95	x	x	aceticacid	butylethanoate	64-19-7	123-86-4	-6.11	-3.08	-3.03
96	x	x	methyl_hexanoate	n-butylbenzene	106-70-7	104-51-8	-6.09	-1.52	-4.57
97	x	x	4-methylaniline	n-decane	106-49-0	124-18-5	-6.05	-1.44	-4.61
98	x	x	1-butanol	1-butanol	71-36-3	71-36-3	-6.03	-2.77	-3.26
99	x	x	o-cresol	cyclohexane	95-48-7	110-82-7	-6.02	-1.34	-4.68
100	x		2-chloroethanol	2-chloroethanol	107-07-3	107-07-3	-6.01	-	-
101	x	x	morpholine	n-octanol	110-91-8	111-87-5	-5.99	-2.78	-3.21
102	x	x	pentylethanoate	sec-butylbenzene	628-63-7	135-98-8	-5.98	-1.51	-4.47
103	x	x	octanal	n-hexadecane	124-13-0	544-76-3	-5.98	-1.24	-4.74
104	x	x	methyl_pentanoate	dichloroethane	624-24-8	107-06-2	-5.97	-2.86	-3.11
105	x	x	1-4-dichlorobenzene	cyclohexane	106-46-7	110-82-7	-5.89	-0.82	-5.07
106	x	x	1-nitropropane	1-nitropropane	108-03-2	108-03-2	-5.88	-4.34	-1.53
107	x	x	2-methylpyrazine	n-octanol	109-08-0	111-87-5	-5.87	-2.82	-3.05
108	x	x	4-methylphenol	n-hexane	106-44-5	110-54-3	-5.86	-1.29	-4.57
109	x	x	ethylbenzene	chloroform	100-41-4	67-66-3	-5.84	-1.04	-4.80
110	x	x	thiophene	chloroform	110-02-1	67-66-3	-5.83	-1.29	-4.54
111	x	x	m-xylene	m-xylene	108-38-3	108-38-3	-5.81	-0.66	-5.15
112	x	x	2-Hexanone	benzene	591-78-6	71-43-2	-5.76	-1.51	-4.25
113	x	x	butanoicacid	o-nitrotoluene	107-92-6	88-72-2	-5.76	-3.84	-1.92
114	x	x	diethyl_disulfide	n-hexadecane	110-81-6	544-76-3	-5.74	-0.98	-4.76
115	x	x	ethylbenzene	ethylbenzene	100-41-4	100-41-4	-5.73	-0.67	-5.06
116	x	x	dibutylether	dibutylether	142-96-1	142-96-1	-5.71	-0.63	-5.08
117	x	x	pentylethanoate	cyclohexane	628-63-7	110-82-7	-5.71	-1.29	-4.42
118	x	x	anisole	diethylether	100-66-3	60-29-7	-5.71	-1.75	-3.96
119	x	x	chlorobenzene	chlorobenzene	108-90-7	108-90-7	-5.70	-1.47	-4.23

TABLE B.5: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
120	x	x	o-cresol	2,2,4-trimethylpentane	95-48-7	540-84-1	-5.68	-1.27	-4.41
121	x	x	methyl_hexanoate	n-pentane	106-70-7	109-66-0	-5.67	-1.14	-4.53
122	x	x	trimethyl_phosphate	cyclohexane	512-56-1	110-82-7	-5.67	-2.71	-2.96
123	x	x	thioanisole	cyclohexane	100-68-5	110-82-7	-5.66	-1.66	-4.00
124	x	x	nitromethane	n,n-dimethylacetamide	75-52-5	127-19-5	-5.62	-4.80	-0.82
125	x	x	pentylethanoate	n-pentane	628-63-7	109-66-0	-5.62	-1.14	-4.48
126	x	x	n-octane	triethylamine	111-65-9	121-44-8	-5.62	-0.08	-5.54
127	x	x	butylethanoate	butylethanoate	123-86-4	123-86-4	-5.58	-2.36	-3.22
128	x	x	ethylethanoate	chloroform	141-78-6	67-66-3	-5.58	-2.31	-3.27
129	x	x	m-xylene	diethylether	108-38-3	60-29-7	-5.56	-1.00	-4.56
130	x	x	1-heptanol	n-octane	111-70-6	111-65-9	-5.56	-1.04	-4.52
131	x	x	toluene	1,2-dibromoethane	108-88-3	106-93-4	-5.55	-1.09	-4.46
132	x	x	pyridine	dichloroethane	110-86-1	107-06-2	-5.53	-2.59	-2.94
133	x	x	aniline	cyclohexane	62-53-3	110-82-7	-5.52	-1.51	-4.01
134	x	x	2-5-dimethylpyridine	n-hexadecane	589-93-5	544-76-3	-5.52	-1.08	-4.44
135	x	x	phenol	n-decane	108-95-2	124-18-5	-5.50	-1.41	-4.09
136	x	x	butylamine	1-hexanol	109-73-9	111-27-3	-5.50	-2.31	-3.19
137	x	x	pyridine	pyridine	110-86-1	110-86-1	-5.47	-2.69	-2.78
138	x	x	2-Hexanone	carbontetrachloride	591-78-6	56-23-5	-5.47	-1.48	-3.99
139	x	x	anisole	n-octanol	100-66-3	111-87-5	-5.47	-2.27	-3.20
140	x	x	1-propanol	tributylphosphate	71-23-8	126-73-8	-5.42	-2.47	-2.95
141	x	x	pentylethanoate	heptane	628-63-7	142-82-5	-5.42	-1.20	-4.22
142	x	x	propylethanoate	dichloroethane	109-60-4	107-06-2	-5.40	-2.80	-2.60
143	x	x	ethanol	n,n-dimethylacetamide	64-17-5	127-19-5	-5.40	-2.92	-2.48
144	x	x	n-octane	toluene	111-65-9	108-88-3	-5.38	-0.08	-5.30
145	x	x	butylethanoate	1-chlorohexane	123-86-4	544-10-5	-5.37	-2.50	-2.87
146	x	x	propanoicacid	chloroform	79-09-4	67-66-3	-5.37	-2.92	-2.45
147	x	x	o-xylene	n-hexadecane	95-47-6	544-76-3	-5.37	-0.58	-4.79
148	x	x	butylamine	carbontetrachloride	109-73-9	56-23-5	-5.35	-1.08	-4.27
149	x	x	hexanoicacid	n-hexadecane	142-62-1	544-76-3	-5.35	-1.53	-3.82

TABLE B.6: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
150	x	x	butylamine	n-octanol	109-73-9	111-87-5	-5.33	-2.22	-3.11
151	x	x	benzonitrile	heptane	100-47-0	142-82-5	-5.33	-1.78	-3.55
152	x	x	pentylethanoate	n-nonane	628-63-7	111-84-2	-5.33	-1.24	-4.09
153	x	x	toluene	benzene	108-88-3	71-43-2	-5.32	-0.64	-4.68
154	x	x	2-Hexanone	n-butylbenzene	591-78-6	104-51-8	-5.31	-1.57	-3.74
155	x	x	1-butanol	methanol	71-36-3	67-56-1	-5.31	-2.91	-2.40
156	x	x	phenol	2,2,4-trimethylpentane	108-95-2	540-84-1	-5.30	-1.36	-3.94
157	x	x	2-heptanone	n-nonane	110-43-0	111-84-2	-5.24	-1.28	-3.96
158	x		nitromethane	methyl_ethyl_ketone	75-52-5	78-93-3	-5.24	-	-
159	x	x	4-methyl-2-pentanone	4-methyl-2-pentanone	108-10-1	108-10-1	-5.23	-2.94	-2.29
160	x	x	2-heptanone	heptane	110-43-0	142-82-5	-5.22	-1.23	-3.99
161	x	x	aceticacid	dibutylether	64-19-7	142-96-1	-5.21	-2.44	-2.77
162	x	x	aniline	2,2,4-trimethylpentane	62-53-3	540-84-1	-5.20	-1.43	-3.77
163	x	x	2-methylpyridine	dibutylether	109-06-8	142-96-1	-5.20	-1.52	-3.68
164	x	x	propylethanoate	chlorobenzene	109-60-4	108-90-7	-5.15	-2.45	-2.70
165	x	x	aniline	n-pentane	62-53-3	109-66-0	-5.15	-1.33	-3.82
166	x	x	2-pentanone	benzene	107-87-9	71-43-2	-5.14	-1.50	-3.64
167	x	x	2-propanol	dimethylsulfoxide	67-63-0	67-68-5	-5.14	-2.85	-2.29
168	x	x	toluene	bromobenzene	108-88-3	108-86-1	-5.13	-1.13	-4.00
169	x	x	1-1-1-trifluoro-2-propanol	n-octanol	374-01-6	111-87-5	-5.12	-2.73	-2.39
170	x	x	2-pentanone	perfluorobenzene	107-87-9	392-56-3	-5.10	-1.32	-3.78
171	x	x	1-hexanol	2,2,4-trimethylpentane	111-27-3	540-84-1	-5.10	-1.10	-4.00
172	x	x	nitromethane	cyclohexanone	75-52-5	108-94-1	-5.09	-4.53	-0.56
173	x	x	dipropylamine	benzene	142-84-7	71-43-2	-5.09	-0.69	-4.40
174	x	x	ethanol	pyridine	64-17-5	110-86-1	-5.08	-2.67	-2.41
175	x		toluene	methyl_ethyl_ketone	108-88-3	78-93-3	-5.06	-	-
176	x	x	2-methylpyridine	cyclohexane	109-06-8	110-82-7	-5.05	-1.03	-4.02
177	x	x	propylethanoate	carbontetrachloride	109-60-4	56-23-5	-5.03	-1.43	-3.60
178	x	x	ethanol	1-butanol	64-17-5	71-36-3	-5.02	-2.77	-2.25
179	x	x	1-pentanol	iodobenzene	71-41-0	591-50-4	-5.02	-2.01	-3.01

TABLE B.7: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
180	x	x	p-xylene	n-hexane	106-42-3	110-54-3	-5.01	-0.50	-4.51
181	x	x	pyridine	carbontetrachloride	110-86-1	56-23-5	-5.01	-1.28	-3.73
182	x	x	chlorobenzene	n-octanol	108-90-7	111-87-5	-5.00	-1.68	-3.32
183	x	x	3-3-dimethylbutanone	ethylbenzene	75-97-8	100-41-4	-4.92	-1.52	-3.40
184	x	x	1-butanol	dichloroethane	71-36-3	107-06-2	-4.92	-2.58	-2.34
185	x	x	1-propanol	ethylethanoate	71-23-8	141-78-6	-4.90	-2.28	-2.62
186	x	x	methylethanoate	chloroform	79-20-9	67-66-3	-4.90	-2.46	-2.44
187	x	x	n-octane	2,6-dimethylpyridine	111-65-9	108-48-5	-4.88	-0.14	-4.74
188	x	x	pyridine	diisopropylether	110-86-1	108-20-3	-4.88	-1.78	-3.10
189	x	x	ethanol	2,6-dimethylpyridine	64-17-5	108-48-5	-4.87	-2.40	-2.47
190	x	x	propylethanoate	xylene-mixture	109-60-4	1330-20-7	-4.87	-1.53	-3.34
191	x	x	butylethanoate	n-hexane	123-86-4	110-54-3	-4.86	-1.18	-3.68
192	x	x	toluene	n-hexane	108-88-3	110-54-3	-4.84	-0.51	-4.33
193	x	x	propylamine	1-hexanol	107-10-8	111-27-3	-4.83	-2.30	-2.53
194	x	x	butanal	2-methyl-1-propanol	123-72-8	78-83-1	-4.82	-2.92	-1.90
195	x	x	piperidine	diethylether	110-89-4	60-29-7	-4.82	-1.09	-3.73
196	x	x	2-propanol	2-propanol	67-63-0	67-63-0	-4.82	-2.70	-2.12
197	x	x	2,2,2-trifluoroethanol	2,2,2-trifluoroethanol	75-89-8	75-89-8	-4.82	-4.28	-0.54
198	x	x	butanoicacid	carbontetrachloride	107-92-6	56-23-5	-4.81	-1.78	-3.03
199	x	x	2-pentanone	mesitylene	107-87-9	108-67-8	-4.80	-1.50	-3.30
200	x	x	propylethanoate	isopropylbenzene	109-60-4	98-82-8	-4.78	-1.52	-3.26
201	x	x	methyl_pentanoate	n-decane	624-24-8	124-18-5	-4.77	-1.26	-3.51
202	x		n-octane	ethoxybenzene	111-65-9	103-73-1	-4.75	-	-
203	x	x	1-propanol	methanol	71-23-8	67-56-1	-4.75	-2.89	-1.86
204	x	x	2-pentanone	n-butylbenzene	107-87-9	104-51-8	-4.74	-1.56	-3.18
205	x	x	propylamine	chloroform	107-10-8	67-66-3	-4.73	-1.79	-2.94
206	x	x	n-octane	2-methylpyridine	111-65-9	109-06-8	-4.73	-0.15	-4.58
207	x	x	nitromethane	m-cresol	75-52-5	108-39-4	-4.73	-4.41	-0.32
208	x	x	1-pentanol	ethylbenzene	71-41-0	100-41-4	-4.72	-1.35	-3.37
209	x	x	butylethanoate	decalin-mixture	123-86-4	91-17-8	-4.71	-1.42	-3.29

TABLE B.8: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
210	x	x	ethanol	2-methoxyethanol	64-17-5	109-86-4	-4.71	-2.76	-1.95
211	x	x	pentylamine	xylene-mixture	110-58-7	1330-20-7	-4.70	-1.16	-3.54
212	x	x	dichloroethane	dichloroethane	107-06-2	107-06-2	-4.70	-2.34	-2.36
213	x	x	butylethanoate	n-nonane	123-86-4	111-84-2	-4.69	-1.24	-3.45
214	x	x	nitromethane	anisole	75-52-5	100-66-3	-4.69	-3.38	-1.31
215	x	x	aceticacid	o-nitrotoluene	64-19-7	88-72-2	-4.68	-4.00	-0.68
216	x	x	2-methylpyridine	n-hexadecane	109-06-8	544-76-3	-4.68	-1.05	-3.63
217	x	x	2-Hexanone	n-hexane	591-78-6	110-54-3	-4.68	-1.21	-3.47
218	x	x	nitromethane	chloroform	75-52-5	67-66-3	-4.68	-3.53	-1.15
219	x	x	1,4-dioxane	ethanol	123-91-1	64-17-5	-4.68	-2.95	-1.73
220	x	x	butylethanoate	n-decane	123-86-4	124-18-5	-4.66	-1.26	-3.40
221	x		acrylonitrile	acrylonitrile	107-13-1	107-13-1	-4.66	-	-
222	x	x	propylethanoate	1-fluorooctane	109-60-4	463-11-6	-4.65	-2.12	-2.53
223	x	x	1-butanol	1,2-dibromoethane	71-36-3	106-93-4	-4.65	-2.14	-2.51
224	x	x	2-methoxypropane	n-octanol	598-53-8	111-87-5	-4.64	-1.45	-3.19
225	x	x	ethylethanoate	chlorobenzene	141-78-6	108-90-7	-4.63	-2.46	-2.17
226	x	x	butanal	n-octanol	123-72-8	111-87-5	-4.62	-2.71	-1.91
227	x	x	butylethanoate	n-hexadecane	123-86-4	544-76-3	-4.61	-1.31	-3.30
228	x	x	chlorobenzene	decalin-mixture	108-90-7	91-17-8	-4.61	-0.83	-3.78
229	x	x	ethylbenzene	1-nonanol	100-41-4	143-08-8	-4.61	-1.27	-3.34
230	x	x	methylpropanoate	benzene	554-12-1	71-43-2	-4.58	-1.45	-3.13
231	x	x	toluene	m-cresol	108-88-3	108-39-4	-4.58	-1.39	-3.19
232	x	x	ethanol	tributylphosphate	64-17-5	126-73-8	-4.57	-2.47	-2.10
233	x		methyl_ethyl_ketone	methyl_ethyl_ketone	78-93-3	78-93-3	-4.55	-	-
234	x	x	toluene	n-octanol	108-88-3	111-87-5	-4.55	-1.34	-3.21
235	x	x	nitromethane	1,2-dibromoethane	75-52-5	106-93-4	-4.54	-3.59	-0.95
236	x	x	ethylethanoate	benzene	141-78-6	71-43-2	-4.53	-1.46	-3.07
237	x	x	n-octane	2-propanol	111-65-9	67-63-0	-4.50	-0.17	-4.33
238	x	x	butylethanoate	n-pentadecane	123-86-4	629-62-9	-4.49	-1.30	-3.19
239	x	x	butylamine	tetrachloroethene	109-73-9	127-18-4	-4.49	-1.10	-3.39

TABLE B.9: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
240	x	x	1-nitropropane	carbontetrachloride	108-03-2	56-23-5	-4.49	-2.03	-2.46
241	x	x	ethylethanoate	ethylethanoate	141-78-6	141-78-6	-4.46	-2.50	-1.96
242	x	x	n-octane	1-butanol	111-65-9	71-36-3	-4.45	-0.16	-4.29
243	x	x	2-propanol	diethylether	67-63-0	60-29-7	-4.44	-1.93	-2.51
244	x	x	butylamine	ethylbenzene	109-73-9	100-41-4	-4.43	-1.18	-3.25
245	x	x	butanal	butanal	123-72-8	123-72-8	-4.42	-2.84	-1.58
246	x	x	perfluorobenzene	perfluorobenzene	392-56-3	392-56-3	-4.42	-0.53	-3.89
247	x	x	3-3-dimethylbutanone	cyclohexane	75-97-8	110-82-7	-4.42	-1.25	-3.17
248	x	x	ethanol	methanol	64-17-5	67-56-1	-4.41	-2.90	-1.51
249	x	x	ethylethanoate	carbontetrachloride	141-78-6	56-23-5	-4.40	-1.44	-2.96
250	x	x	ethanol	n-octanol	64-17-5	111-87-5	-4.36	-2.56	-1.80
251	x	x	propylethanoate	cyclohexane	109-60-4	110-82-7	-4.36	-1.28	-3.08
252	x	x	methyl_ethyl_ketone	2,6-dimethylpyridine	78-93-3	108-48-5	-4.34	-2.77	-1.57
253	x	x	butylamine	toluene	109-73-9	108-88-3	-4.33	-1.15	-3.18
254	x	x	nitromethane	chlorobenzene	75-52-5	108-90-7	-4.32	-3.77	-0.55
255	x	x	3-3-dimethylbutanone	heptane	75-97-8	142-82-5	-4.30	-1.16	-3.14
256	x	x	1-hexanol	n-dodecane	111-27-3	112-40-3	-4.28	-1.15	-3.13
257	x	x	ethylethanoate	xylene-mixture	141-78-6	1330-20-7	-4.26	-1.54	-2.72
258	x	x	tetrachloroethene	n-octanol	127-18-4	111-87-5	-4.24	-0.71	-3.53
259	x	x	methyl_ethyl_ketone	xylene-mixture	78-93-3	1330-20-7	-4.23	-1.58	-2.65
260	x	x	diethylsulfide	n-hexadecane	352-93-2	544-76-3	-4.23	-0.90	-3.33
261	x	x	toluene	tetrahydrothiophene-s,s-dioxide	108-88-3	126-33-0	-4.23	-1.56	-2.67
262	x	x	tetrahydrofuran	tetrahydrofuran	109-99-9	109-99-9	-4.23	-1.75	-2.48
263	x	x	3-3-dimethylbutanone	n-octane	75-97-8	111-65-9	-4.21	-1.19	-3.02
264	x	x	ethylamine	1-hexanol	75-04-7	111-27-3	-4.20	-2.30	-1.90
265	x	x	1-butanol	carbontetrachloride	71-36-3	56-23-5	-4.20	-1.29	-2.91
266	x	x	nitromethane	diethylether	75-52-5	60-29-7	-4.19	-3.39	-0.80
267	x	x	nitromethane	diphenylether	75-52-5	101-84-8	-4.19	-3.20	-0.99
268	x	x	benzene	cyclohexane	71-43-2	110-82-7	-4.19	-0.56	-3.63
269	x	x	3-3-dimethylbutanone	n-nonane	75-97-8	111-84-2	-4.19	-1.20	-2.99

TABLE B.10: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
270	x	x	methylpropanoate	tert-butylbenzene	554-12-1	98-06-6	-4.17	-1.50	-2.67
271	x	x	1-pentanol	2,2,4-trimethylpentane	71-41-0	540-84-1	-4.17	-1.05	-3.12
272	x	x	2-pentanone	n-pentane	107-87-9	109-66-0	-4.16	-1.16	-3.00
273	x	x	methanol	tributylphosphate	67-56-1	126-73-8	-4.16	-2.53	-1.63
274	x	x	ethylamine	1-heptanol	75-04-7	111-70-6	-4.15	-2.26	-1.89
275	x	x	2-pentanone	2,2,4-trimethylpentane	107-87-9	540-84-1	-4.14	-1.24	-2.90
276	x	x	methylethanoate	methylethanoate	79-20-9	79-20-9	-4.14	-2.76	-1.37
277	x	x	pyridine	n-hexadecane	110-86-1	544-76-3	-4.10	-1.15	-2.95
278	x	x	methyl_ethyl_ketone	carbontetrachloride	78-93-3	56-23-5	-4.09	-1.47	-2.62
279	x	x	propylethanoate	n-octane	109-60-4	111-65-9	-4.09	-1.22	-2.87
280	x	x	propanoicacid	carbontetrachloride	79-09-4	56-23-5	-4.09	-1.84	-2.25
281	x	x	ethylethanoate	carbondisulfide	141-78-6	75-15-0	-4.08	-1.66	-2.42
282	x	x	1-butanol	bromobenzene	71-36-3	108-86-1	-4.08	-2.21	-1.87
283	x	x	propylethanoate	decalin-mixture	109-60-4	91-17-8	-4.05	-1.41	-2.64
284	x	x	ethylmethanoate	ethylmethanoate	109-94-4	109-94-4	-4.03	-2.61	-1.42
285	x	x	ethylamine	chloroform	75-04-7	67-66-3	-4.02	-1.79	-2.23
286	x	x	diethylamine	dichloroethane	109-89-7	107-06-2	-4.00	-1.47	-2.53
287	x		methyl_ethyl_ketone	1-2-3-trimethylbenzene	78-93-3	526-73-8	-3.97	-	-
288	x	x	ethanol	butylethanoate	64-17-5	123-86-4	-3.97	-2.14	-1.83
289	x	x	methyl_ethyl_ketone	diisopropylether	78-93-3	108-20-3	-3.96	-2.05	-1.91
290	x	x	1-pentanol	n-pentane	71-41-0	109-66-0	-3.92	-0.98	-2.94
291	x	x	1-pentanol	n-decane	71-41-0	124-18-5	-3.92	-1.09	-2.83
292	x	x	propylethanoate	n-pentadecane	109-60-4	629-62-9	-3.91	-1.29	-2.62
293	x	x	fluorobenzene	n-octanol	462-06-6	111-87-5	-3.87	-1.45	-2.42
294	x	x	propanal	diethylether	123-38-6	60-29-7	-3.85	-2.19	-1.66
295	x	x	1-propanol	1,2-dibromoethane	71-23-8	106-93-4	-3.82	-2.13	-1.69
296	x	x	1-propanol	chlorobenzene	71-23-8	108-90-7	-3.82	-2.24	-1.58
297	x	x	1-propanethiol	2,2,4-trimethylpentane	107-03-9	540-84-1	-3.78	-0.85	-2.93
298	x	x	methyl_ethyl_ketone	dibutylether	78-93-3	142-96-1	-3.78	-1.92	-1.86
299	x	x	1-butanol	ethylbenzene	71-36-3	100-41-4	-3.77	-1.40	-2.37

TABLE B.11: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
300	x	x	2-pentanone	n-hexadecane	107-87-9	544-76-3	-3.76	-1.33	-2.43
301	x	x	methylethanoate	ethylbenzene	79-20-9	100-41-4	-3.74	-1.68	-2.06
302	x	x	1-propanol	bromobenzene	71-23-8	108-86-1	-3.74	-2.20	-1.54
303	x	x	benzene	1-heptanol	71-43-2	111-70-6	-3.73	-1.38	-2.35
304	x	x	butylamine	decalin-mixture	109-73-9	91-17-8	-3.72	-1.06	-2.66
305	x	x	methylpropanoate	cyclohexane	554-12-1	110-82-7	-3.71	-1.27	-2.44
306	x	x	methylethanoate	xylene-mixture	79-20-9	1330-20-7	-3.70	-1.65	-2.05
307	x	x	t-butanol	benzene	75-65-0	71-43-2	-3.70	-1.18	-2.52
308	x	x	methylmethanoate	methylmethanoate	107-31-3	107-31-3	-3.66	-2.82	-0.85
309	x	x	fluorobenzene	carbontetrachloride	462-06-6	56-23-5	-3.64	-0.72	-2.92
310	x	x	1-methoxypropane	n-octanol	557-17-5	111-87-5	-3.63	-1.42	-2.21
311	x	x	methylpropanoate	heptane	554-12-1	142-82-5	-3.63	-1.19	-2.44
312	x	x	methylethanoate	tetrachloroethene	79-20-9	127-18-4	-3.63	-1.57	-2.06
313	x	x	propylamine	carbontetrachloride	107-10-8	56-23-5	-3.59	-1.08	-2.51
314	x	x	acetone	toluene	67-64-1	108-88-3	-3.59	-1.71	-1.88
315	x		perfluorohexane	perfluorohexane	355-42-0	355-42-0	-3.58	-	-
316	x	x	propylamine	bromobenzene	107-10-8	108-86-1	-3.57	-1.89	-1.68
317	x	x	1-butanol	n-hexadecane	71-36-3	544-76-3	-3.55	-1.16	-2.39
318	x	x	propylamine	iodobenzene	107-10-8	591-50-4	-3.54	-1.76	-1.78
319	x	x	benzene	1-pentanol	71-43-2	71-41-0	-3.53	-1.44	-2.09
320	x	x	1-propanol	iodobenzene	71-23-8	591-50-4	-3.52	-2.06	-1.46
321	x		nitromethane	carbontetrachloride	75-52-5	56-23-5	-3.52	-	-
322	x	x	methylpropanoate	n-decane	554-12-1	124-18-5	-3.49	-1.25	-2.24
323	x	x	ethylethanoate	n-octane	141-78-6	111-65-9	-3.48	-1.22	-2.26
324	x	x	ethylethanoate	decalin-mixture	141-78-6	91-17-8	-3.47	-1.41	-2.06
325	x	x	diethylamine	n-octane	109-89-7	111-65-9	-3.42	-0.57	-2.85
326	x	x	ethanol	chlorobenzene	64-17-5	108-90-7	-3.30	-2.24	-1.06
327	x	x	nitromethane	carbondisulfide	75-52-5	75-15-0	-3.30	-2.56	-0.74
328	x	x	ethylamine	tributylphosphate	75-04-7	126-73-8	-3.29	-2.12	-1.17
329	x	x	acetone	xylene-mixture	67-64-1	1330-20-7	-3.26	-1.72	-1.54

TABLE B.12: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
330	x	x	ethylethanoate	n-hexadecane	141-78-6	544-76-3	-3.25	-1.30	-1.95
331	x	x	ethanol	diphenylether	64-17-5	101-84-8	-3.22	-1.88	-1.34
332	x	x	z-1,2-dichloroethene	n-hexadecane	156-59-2	544-76-3	-3.11	-0.85	-2.26
333	x	x	n-pentane	benzene	109-66-0	71-43-2	-2.99	-0.06	-2.93
334	x	x	methylethanoate	heptane	79-20-9	142-82-5	-2.97	-1.29	-1.68
335	x	x	1,2-ethanediol	n-hexadecane	107-21-1	544-76-3	-2.81	-2.09	-0.72
336	x	x	methanol	bromoform	67-56-1	75-25-2	-2.79	-2.07	-0.72
337	x	x	trimethylamine	diethylether	75-50-3	60-29-7	-2.78	-0.82	-1.96
338	x	x	1-propanol	n-octane	71-23-8	111-65-9	-2.76	-1.09	-1.67
339	x	x	1-propanol	n-nonane	71-23-8	111-84-2	-2.76	-1.11	-1.65
340	x	x	ethylamine	chlorobenzene	75-04-7	108-90-7	-2.73	-1.92	-0.81
341	x	x	ethanol	carbondisulfide	64-17-5	75-15-0	-2.72	-1.49	-1.23
342	x	x	methylamine	benzene	74-89-5	71-43-2	-2.66	-1.18	-1.48
343	x	x	trimethylamine	ethylbenzene	75-50-3	100-41-4	-2.64	-0.56	-2.08
344	x	x	water	dichloromethane	7732-18-5	75-09-2	-2.63	-3.74	1.11
345	x	x	trimethylamine	xylene-mixture	75-50-3	1330-20-7	-2.63	-0.55	-2.08
346	x	x	fluorotrichloromethane	cyclohexane	75-69-4	110-82-7	-2.63	-0.23	-2.40
347	x	x	acetone	heptane	67-64-1	142-82-5	-2.61	-1.34	-1.27
348	x	x	acetone	n-hexane	67-64-1	110-54-3	-2.60	-1.31	-1.29
349	x	x	methylamine	carbontetrachloride	74-89-5	56-23-5	-2.53	-1.15	-1.38
350	x	x	acetone	2,2,4-trimethylpentane	67-64-1	540-84-1	-2.44	-1.36	-1.08
351	x	x	ethanol	2,2,4-trimethylpentane	64-17-5	540-84-1	-2.44	-1.09	-1.35
352	x	x	methanol	chlorobenzene	67-56-1	108-90-7	-2.44	-2.30	-0.14
353	x	x	methylamine	diethylether	74-89-5	60-29-7	-2.32	-1.81	-0.51
354	x	x	ethylamine	n-pentane	75-04-7	109-66-0	-2.18	-0.84	-1.34
355	x	x	water	chloroform	7732-18-5	67-66-3	-2.05	-3.14	1.09
356	x	x	ethylamine	cyclohexane	75-04-7	110-82-7	-2.04	-0.96	-1.08
357	x	x	difluorodichloromethane	cyclohexane	75-71-8	110-82-7	-1.81	-0.18	-1.63
358	x	x	water	benzene	7732-18-5	71-43-2	-1.71	-2.00	0.29
359	x	x	methanol	n-nonane	67-56-1	111-84-2	-1.29	-1.15	-0.14

TABLE B.13: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
360	x	x	propene	n-hexadecane	115-07-1	544-76-3	-1.29	-0.27	-1.02
361	x	x	naproxen	n-octanol	22204-53-1	111-87-5	-14.55	-5.63	-8.92
362	x		ethyl_octadecanoate	n-hexadecane	111-61-5	544-76-3	-13.69	-	-
363	x	x	ethyl-4-hydroxybenzoate	n-octanol	120-47-8	111-87-5	-12.57	-5.14	-7.43
364	x	x	fenthion	n-octanol	55-38-9	111-87-5	-12.55	-5.85	-6.70
365	x	x	acetylsalicylic_acid	n-octanol	50-78-2	111-87-5	-11.56	-6.02	-5.54
366	x	x	anthracene	n-octanol	120-12-7	111-87-5	-10.47	-2.43	-8.04
367	x		dimethyl_phthalate	dimethyl_phthalate	131-11-3	131-11-3	-10.39	-	-
368	x	x	3-hydroxybenzaldehyde	dichloroethane	100-83-4	107-06-2	-10.11	-5.42	-4.69
369	x	x	anthracene	heptane	120-12-7	142-82-5	-10.00	-0.92	-9.08
370	x		4-chloroacetophenone	4-chloroacetophenone	99-91-2	99-91-2	-9.93	-	-
371	x		n-heptanoic_acid	n-heptanoic_acid	111-14-8	111-14-8	-9.88	-	-
372	x	x	1-decanol	n-octanol	112-30-1	111-87-5	-9.88	-2.63	-7.25
373	x	x	1-decanol	1-decanol	112-30-1	112-30-1	-9.58	-2.49	-7.09
374	x	x	1-decanol	1-decanol	112-30-1	112-30-1	-9.58	-2.49	-7.09
375	x	x	2,2-dichlorobiphenyl	n-octanol	13029-08-8	111-87-5	-9.41	-3.03	-6.38
376	x	x	tripropyl_phosphate	benzene	513-08-6	71-43-2	-9.34	-2.90	-6.44
377	x	x	3-hydroxybenzaldehyde	benzene	100-83-4	71-43-2	-9.29	-2.74	-6.55
378	x	x	4-methylphenol	1-pentanol	106-44-5	71-41-0	-9.25	-3.48	-5.77
379	x	x	2-3-dichlorobiphenyl	n-octanol	16605-91-7	111-87-5	-9.23	-2.65	-6.58
380	x	x	2,2-dichlorobiphenyl	heptane	13029-08-8	142-82-5	-9.22	-1.21	-8.01
381	x	x	2-6-dichlorobenzonitrile	n-octanol	1194-65-6	111-87-5	-9.18	-4.40	-4.78
382	x	x	4-methylphenol	1-heptanol	106-44-5	111-70-6	-9.16	-3.34	-5.82
383	x		cinnamaldehyde	cinnamaldehyde	104-55-2	104-55-2	-9.12	-	-
384	x	x	dichlorphos	benzene	62-73-7	71-43-2	-9.09	-2.73	-6.36
385	x		dibenzyl_ether	dibenzyl_ether	103-50-4	103-50-4	-9.05	-	-
386	x	x	1-nonanol	1-nonanol	143-08-8	143-08-8	-9.05	-2.44	-6.61
387	x		4-methoxybenzaldehyde	4-methoxybenzaldehyde	123-11-5	123-11-5	-9.02	-	-
388	x	x	4-methylphenol	n-octanol	106-44-5	111-87-5	-8.84	-3.26	-5.58
389	x	x	hexanoicacid	2-methyl-1-propanol	142-62-1	78-83-1	-8.77	-3.54	-5.23

TABLE B.14: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
390	x	x	o-cresol	1-hexanol	95-48-7	111-27-3	-8.76	-3.24	-5.52
391	x	x	hexanoicacid	1-butanol	142-62-1	71-36-3	-8.75	-3.55	-5.20
392	x		hexamethylphosphoramide	hexamethylphosphoramide	680-31-9	680-31-9	-8.68	-	-
393	x	x	tripropyl_phosphate	n-octanol	513-08-6	111-87-5	-8.65	-5.76	-2.89
394	x	x	tripropyl_phosphate	carbontetrachloride	513-08-6	56-23-5	-8.60	-2.85	-5.75
395	x	x	dichlorphos	n-octanol	62-73-7	111-87-5	-8.59	-5.22	-3.37
396	x	x	o-cresol	1-pentanol	95-48-7	71-41-0	-8.57	-3.33	-5.24
397	x	x	quinoline	quinoline	91-22-5	91-22-5	-8.56	-2.85	-5.71
398	x	x	n-octanol	n-octanol	111-87-5	111-87-5	-8.51	-2.29	-6.22
399	x	x	o-cresol	n-octanol	95-48-7	111-87-5	-8.49	-3.11	-5.38
400	x	x	benzylalcohol	benzylalcohol	100-51-6	100-51-6	-8.46	-3.62	-4.84
401	x	x	quinoline	n-octanol	91-22-5	111-87-5	-8.43	-2.89	-5.54
402	x	x	m-cresol	1-hexanol	108-39-4	111-27-3	-8.42	-3.41	-5.01
403	x	x	m-cresol	m-cresol	108-39-4	108-39-4	-8.40	-3.41	-4.99
404	x		4-methylphenol	4-methylphenol	106-44-5	106-44-5	-8.37	-	-
405	x	x	m-cresol	m-cresol	108-39-4	108-39-4	-8.32	-3.41	-4.91
406	x	x	o-nitrotoluene	chloroform	88-72-2	67-66-3	-8.30	-2.91	-5.39
407	x	x	hexanoicacid	diisopropylether	142-62-1	108-20-3	-8.23	-2.32	-5.91
408	x	x	2-methylaniline	chloroform	95-53-4	67-66-3	-8.23	-2.69	-5.54
409	x	x	m-cresol	n-octanol	108-39-4	111-87-5	-8.20	-3.28	-4.92
410	x		propiophenone	propiophenone	93-55-0	93-55-0	-8.19	-	-
411	x	x	n-octanol	butylethanoate	111-87-5	123-86-4	-8.17	-1.89	-6.28
412	x	x	pentanoicacid	1-pentanol	109-52-4	71-41-0	-8.17	-3.70	-4.47
413	x	x	o-cresol	nitrobenzene	95-48-7	98-95-3	-8.16	-3.57	-4.59
414	x	x	p_hydroxybenzaldehyde	carbontetrachloride	123-08-0	56-23-5	-8.16	-2.79	-5.37
415	x	x	pentanoicacid	pentanoicacid	109-52-4	109-52-4	-8.15	-2.12	-6.03
416	x		2-4-dimethylphenol	2-4-dimethylphenol	105-67-9	105-67-9	-8.15	-	-
417	x		2-methoxyphenol	2-methoxyphenol	90-05-1	90-05-1	-8.14	-	-
418	x	x	4-methylphenol	nitrobenzene	106-44-5	98-95-3	-8.13	-3.73	-4.40
419	x	x	4-methylphenol	acetonitrile	106-44-5	75-05-8	-8.08	-3.73	-4.35

TABLE B.15: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
420	x		hexanoicacid	methyl_ethyl_ketone	142-62-1	78-93-3	-8.07	-	-
421	x		dichloroacetic_acid	dichloroacetic_acid	79-43-6	79-43-6	-8.07	-	-
422	x	x	pentanoicacid	2-methyl-1-propanol	109-52-4	78-83-1	-8.06	-3.74	-4.32
423	x	x	o-nitrotoluene	o-nitrotoluene	88-72-2	88-72-2	-8.04	-3.95	-4.09
424	x	x	3-methylaniline	acetonitrile	108-44-1	75-05-8	-8.04	-3.98	-4.06
425	x	x	o-nitrotoluene	o-nitrotoluene	88-72-2	88-72-2	-8.04	-3.95	-4.09
426	x	x	m-cresol	1-decanol	108-39-4	112-30-1	-8.01	-3.09	-4.92
427	x	x	1-heptanol	tributylphosphate	111-70-6	126-73-8	-7.98	-2.39	-5.59
428	x	x	nitrobenzene	nitrobenzene	98-95-3	98-95-3	-7.97	-4.20	-3.77
429	x	x	m-cresol	diethylether	108-39-4	60-29-7	-7.95	-2.53	-5.42
430	x	x	nitrobenzene	nitrobenzene	98-95-3	98-95-3	-7.94	-4.20	-3.74
431	x	x	o-cresol	acetonitrile	95-48-7	75-05-8	-7.92	-3.57	-4.35
432	x	x	m-cresol	acetonitrile	108-39-4	75-05-8	-7.90	-3.75	-4.15
433	x	x	naphthalene	chloroform	91-20-3	67-66-3	-7.89	-1.51	-6.38
434	x	x	o-cresol	o-cresol	95-48-7	95-48-7	-7.86	-2.85	-5.01
435	x	x	phenol	nitrobenzene	108-95-2	98-95-3	-7.86	-3.78	-4.08
436	x	x	2-methylaniline	acetonitrile	95-53-4	75-05-8	-7.85	-3.87	-3.98
437	x	x	4-methylphenol	methanol	106-44-5	67-56-1	-7.84	-3.71	-4.13
438	x	x	1-heptanol	1-heptanol	111-70-6	111-70-6	-7.84	-2.55	-5.29
439	x		meso-2-3-butanediol	meso-2-3-butanediol	5341-95-7	5341-95-7	-7.81	-	-
440	x	x	acetophenone	chloroform	98-86-2	67-66-3	-7.81	-2.79	-5.02
441	x	x	4-methylaniline	butylethanoate	106-49-0	123-86-4	-7.81	-2.83	-4.98
442	x	x	phenol	o-nitrotoluene	108-95-2	88-72-2	-7.79	-3.71	-4.08
443	x	x	1-heptanol	n-octanol	111-70-6	111-87-5	-7.75	-2.49	-5.26
444	x	x	4-methylphenol	dichloroethane	106-44-5	107-06-2	-7.75	-3.28	-4.47
445	x	x	3-methylaniline	methanol	108-44-1	67-56-1	-7.75	-3.96	-3.79
446	x	x	o-cresol	dichloroethane	95-48-7	107-06-2	-7.73	-3.13	-4.60
447	x	x	o-cresol	methanol	95-48-7	67-56-1	-7.72	-3.56	-4.16
448	x		ethyl_benzoate	ethyl_benzoate	93-89-0	93-89-0	-7.72	-	-
449	x		diethyl_malonate	diethyl_malonate	105-53-3	105-53-3	-7.72	-	-

TABLE B.16: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
450	x	x	tripropyl_phosphate	cyclohexane	513-08-6	110-82-7	-7.71	-2.54	-5.17
451	x	x	acetophenone	acetophenone	98-86-2	98-86-2	-7.69	-3.70	-3.98
452	x	x	m-cresol	methanol	108-39-4	67-56-1	-7.68	-3.73	-3.95
453	x	x	1-decanol	n-hexadecane	112-30-1	544-76-3	-7.68	-1.20	-6.48
454	x	x	butanoicacid	1-butanol	107-92-6	71-36-3	-7.66	-3.72	-3.94
455	x		hydrogen_peroxide	hydrogen_peroxide	7722-84-1	7722-84-1	-7.64	-	-
456	x	x	butanoicacid	2-methyl-1-propanol	107-92-6	78-83-1	-7.62	-3.71	-3.91
457	x	x	3-methylaniline	benzene	108-44-1	71-43-2	-7.61	-1.70	-5.91
458	x	x	pentanoicacid	2-butanol	109-52-4	78-92-2	-7.61	-3.72	-3.89
459	x	x	aniline	aniline	62-53-3	62-53-3	-7.61	-3.23	-4.38
460	x		2-chloroaniline	2-chloroaniline	95-51-2	95-51-2	-7.61	-	-
461	x	x	N-methylaniline	N-methylaniline	100-61-8	100-61-8	-7.61	-2.52	-5.09
462	x	x	methylbenzoate	methylbenzoate	93-58-3	93-58-3	-7.60	-2.91	-4.69
463	x	x	aniline	tributylphosphate	62-53-3	126-73-8	-7.60	-3.38	-4.22
464	x	x	nitrobenzene	benzene	98-95-3	71-43-2	-7.60	-1.95	-5.65
465	x	x	pentanoicacid	diisopropylether	109-52-4	108-20-3	-7.59	-2.46	-5.13
466	x	x	naphthalene	butylethanoate	91-20-3	123-86-4	-7.59	-1.54	-6.05
467	x	x	4-methylaniline	benzene	106-49-0	71-43-2	-7.59	-1.68	-5.91
468	x	x	4-methylphenol	chloroform	106-44-5	67-66-3	-7.59	-2.64	-4.95
469	x		1-2-4-trichlorobenzene	1-2-4-trichlorobenzene	120-82-1	120-82-1	-7.57	-	-
470	x	x	3-methylaniline	n-octanol	108-44-1	111-87-5	-7.57	-3.46	-4.11
471	x	x	1-heptanol	ethylethanoate	111-70-6	141-78-6	-7.56	-2.20	-5.36
472	x	x	2-methylaniline	methanol	95-53-4	67-56-1	-7.54	-3.85	-3.69
473	x	x	aniline	4-methyl-2-pentanone	62-53-3	108-10-1	-7.54	-3.68	-3.86
474	x		pentanoicacid	methyl_ethyl_ketone	109-52-4	78-93-3	-7.54	-	-
475	x	x	4-methylaniline	chlorobenzene	106-49-0	108-90-7	-7.54	-2.97	-4.57
476	x	x	1-heptanol	chloroform	111-70-6	67-66-3	-7.53	-2.02	-5.51
477	x	x	4-methylphenol	1,2-dibromoethane	106-44-5	106-93-4	-7.52	-2.69	-4.83
478	x	x	triethyl_phosphate	carbontetrachloride	78-40-0	56-23-5	-7.51	-2.86	-4.65
479	x	x	1-heptanol	diethylether	111-70-6	60-29-7	-7.51	-1.93	-5.58

TABLE B.17: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
480	x	x	tripropyl_phosphate	heptane	513-08-6	142-82-5	-7.50	-2.37	-5.13
481	x		1-2-dimethoxybenzene	1-2-dimethoxybenzene	91-16-7	91-16-7	-7.49	-	-
482	x	x	phenol	dichloroethane	108-95-2	107-06-2	-7.48	-3.32	-4.16
483	x	x	o-cresol	bromoform	95-48-7	75-25-2	-7.45	-2.40	-5.05
484	x	x	4-methylaniline	n-octanol	106-49-0	111-87-5	-7.44	-3.43	-4.01
485	x	x	o-cresol	benzene	95-48-7	71-43-2	-7.44	-1.54	-5.90
486	x	x	butanoicacid	4-methyl-2-pentanone	107-92-6	108-10-1	-7.44	-3.60	-3.84
487	x	x	benzonitrile	benzonitrile	100-47-0	100-47-0	-7.43	-4.50	-2.93
488	x		c-hexanol	c-hexanol	108-93-0	108-93-0	-7.40	-	-
489	x	x	4-methylaniline	toluene	106-49-0	108-88-3	-7.39	-1.76	-5.63
490	x	x	aniline	dichloroethane	62-53-3	107-06-2	-7.39	-3.53	-3.86
491	x	x	quinoline	cyclohexane	91-22-5	110-82-7	-7.38	-1.24	-6.14
492	x	x	2-methylaniline	benzene	95-53-4	71-43-2	-7.37	-1.63	-5.74
493	x	x	butanoicacid	butanoicacid	107-92-6	107-92-6	-7.37	-2.27	-5.10
494	x	x	pentylethanoate	chloroform	628-63-7	67-66-3	-7.36	-2.31	-5.05
495	x	x	4-methylphenol	benzene	106-44-5	71-43-2	-7.35	-1.62	-5.73
496	x	x	butanoicacid	ethylethanoate	107-92-6	141-78-6	-7.34	-3.09	-4.25
497	x	x	butanoicacid	2-butanol	107-92-6	78-92-2	-7.34	-3.69	-3.65
498	x		butanoicacid	methyl_ethyl_ketone	107-92-6	78-93-3	-7.34	-	-
499	x	x	hexanoicacid	dichloroethane	142-62-1	107-06-2	-7.33	-3.32	-4.01
500	x		benzylamine	benzylamine	100-46-9	100-46-9	-7.32	-	-
501	x	x	aniline	butylethanoate	62-53-3	123-86-4	-7.30	-2.91	-4.39
502	x	x	benzonitrile	benzonitrile	100-47-0	100-47-0	-7.28	-4.50	-2.78
503	x	x	m-cresol	bromobenzene	108-39-4	108-86-1	-7.26	-2.80	-4.46
504	x	x	benzamide	heptane	55-21-0	142-82-5	-7.26	-2.33	-4.93
505	x	x	methylbenzoate	n-octanol	93-58-3	111-87-5	-7.26	-3.17	-4.09
506	x		N-N-dimethylaniline	N-N-dimethylaniline	121-69-7	121-69-7	-7.25	-	-
507	x	x	o-cresol	xylene-mixture	95-48-7	1330-20-7	-7.25	-1.62	-5.63
508	x	x	o-cresol	ethylbenzene	95-48-7	100-41-4	-7.25	-1.65	-5.60
509	x	x	naphthalene	diethylether	91-20-3	60-29-7	-7.25	-1.43	-5.82

TABLE B.18: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
510	x	x	methyl_hexanoate	chloroform	106-70-7	67-66-3	-7.24	-2.34	-4.90
511	x	x	tripropyl_phosphate	n-hexane	513-08-6	110-54-3	-7.24	-2.32	-4.92
512	x	x	naphthalene	diisopropylether	91-20-3	108-20-3	-7.24	-1.26	-5.98
513	x	x	4-methylphenol	chlorobenzene	106-44-5	108-90-7	-7.23	-2.83	-4.40
514	x	x	3-methylaniline	carbontetrachloride	108-44-1	56-23-5	-7.23	-1.67	-5.56
515	x	x	benzaldehyde	dichloroethane	100-52-7	107-06-2	-7.23	-3.27	-3.96
516	x	x	benzonitrile	chloroform	100-47-0	67-66-3	-7.22	-3.38	-3.84
517	x	x	phenol	1,2-dibromoethane	108-95-2	106-93-4	-7.22	-2.73	-4.49
518	x	x	o-nitrotoluene	diethylether	88-72-2	60-29-7	-7.21	-2.78	-4.43
519	x		diiodomethane	diiodomethane	75-11-6	75-11-6	-7.21	-	-
520	x	x	methylbenzoate	carbontetrachloride	93-58-3	56-23-5	-7.19	-1.60	-5.59
521	x	x	propanoicacid	cyclohexanone	79-09-4	108-94-1	-7.18	-3.73	-3.45
522	x	x	4-methylphenol	xylene-mixture	106-44-5	1330-20-7	-7.18	-1.71	-5.47
523	x	x	1-hexanol	1-hexanol	111-27-3	111-27-3	-7.17	-2.70	-4.48
524	x	x	ethoxybenzene	chloroform	103-73-1	67-66-3	-7.16	-1.75	-5.41
525	x	x	2-methylaniline	carbontetrachloride	95-53-4	56-23-5	-7.16	-1.60	-5.56
526	x	x	o-cresol	iodobenzene	95-48-7	591-50-4	-7.14	-2.48	-4.66
527	x	x	phenol	chloroform	108-95-2	67-66-3	-7.14	-2.68	-4.46
528	x	x	1-heptanol	butylethanoate	111-70-6	123-86-4	-7.14	-2.07	-5.07
529	x		iodobenzene	iodobenzene	591-50-4	591-50-4	-7.14	-	-
530	x		tetramethylurea	tetramethylurea	632-22-4	632-22-4	-7.13	-	-
531	x	x	phenol	benzene	108-95-2	71-43-2	-7.12	-1.65	-5.47
532	x	x	4-methylphenol	bromobenzene	106-44-5	108-86-1	-7.12	-2.78	-4.34
533	x		dibutyl_sulfide	dibutyl_sulfide	544-40-1	544-40-1	-7.11	-	-
534	x	x	thiophenol	dichloromethane	108-98-5	75-09-2	-7.11	-2.52	-4.59
535	x	x	acetophenone	carbontetrachloride	98-86-2	56-23-5	-7.10	-1.70	-5.40
536	x	x	1-hexanol	n-octanol	111-27-3	111-87-5	-7.06	-2.59	-4.47
537	x		propanoicacid	methyl_ethyl_ketone	79-09-4	78-93-3	-7.05	-	-
538	x	x	benzaldehyde	benzaldehyde	100-52-7	100-52-7	-7.04	-3.53	-3.50
539	x	x	naphthalene	heptane	91-20-3	142-82-5	-7.02	-0.72	-6.30

TABLE B.19: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
540	x	x	methylbenzoate	cyclohexane	93-58-3	110-82-7	-7.01	-1.43	-5.58
541	x	x	4-methylphenol	iodobenzene	106-44-5	591-50-4	-7.01	-2.60	-4.41
542	x	x	propanoicacid	2-butanol	79-09-4	78-92-2	-7.00	-3.74	-3.26
543	x	x	naphthalene	n-octanol	91-20-3	111-87-5	-6.97	-1.89	-5.08
544	x	x	hexanoicacid	toluene	142-62-1	108-88-3	-6.97	-1.80	-5.17
545	x	x	N-methylaniline	n-octanol	100-61-8	111-87-5	-6.94	-2.87	-4.07
546	x	x	hexanoicacid	benzene	142-62-1	71-43-2	-6.94	-1.72	-5.22
547	x	x	phenol	toluene	108-95-2	108-88-3	-6.93	-1.72	-5.21
548	x	x	nitrobenzene	carbontetrachloride	98-95-3	56-23-5	-6.92	-1.91	-5.01
549	x	x	o-dichlorobenzene	o-dichlorobenzene	95-50-1	95-50-1	-6.91	-1.97	-4.95
550	x		N-N-diethylformamide	N-N-diethylformamide	617-84-5	617-84-5	-6.91	-	-
551	x	x	o-dichlorobenzene	o-dichlorobenzene	95-50-1	95-50-1	-6.89	-1.97	-4.92
552	x	x	n,n-dimethylacetamide	n,n-dimethylacetamide	127-19-5	127-19-5	-6.88	-4.85	-2.03
553	x	x	phenol	bromoform	108-95-2	75-25-2	-6.88	-2.56	-4.32
554	x	x	o-cresol	bromobenzene	95-48-7	108-86-1	-6.87	-2.65	-4.22
555	x	x	n-butylbenzene	n-butylbenzene	104-51-8	104-51-8	-6.86	-0.63	-6.23
556	x	x	n-butylbenzene	n-butylbenzene	104-51-8	104-51-8	-6.86	-0.63	-6.23
557	x	x	2-ethyl-3-methoxypyrazine	n-octanol	25680-58-4	111-87-5	-6.85	-2.53	-4.32
558	x	x	1-heptanol	benzene	111-70-6	71-43-2	-6.85	-1.25	-5.60
559	x	x	nitrobenzene	diethylether	98-95-3	60-29-7	-6.85	-2.94	-3.91
560	x	x	propanoicacid	4-methyl-2-pentanone	79-09-4	108-10-1	-6.85	-3.65	-3.20
561	x	x	butanoicacid	diisopropylether	107-92-6	108-20-3	-6.85	-2.45	-4.40
562	x		2-chloroethyl_ether	2-chloroethyl_ether	111-44-4	111-44-4	-6.83	-	-
563	x	x	phenol	xylene-mixture	108-95-2	1330-20-7	-6.83	-1.73	-5.10
564	x	x	1-hexanol	diethylether	111-27-3	60-29-7	-6.82	-2.03	-4.79
565	x	x	phenol	ethylbenzene	108-95-2	100-41-4	-6.82	-1.76	-5.06
566	x		ethyl_trichloroacetate	ethyl_trichloroacetate	515-84-4	515-84-4	-6.81	-	-
567	x		ethoxybenzene	ethoxybenzene	103-73-1	103-73-1	-6.79	-	-
568	x	x	1-heptanol	dichloroethane	111-70-6	107-06-2	-6.79	-2.50	-4.29
569	x	x	acetophenone	diethylether	98-86-2	60-29-7	-6.79	-2.67	-4.12

TABLE B.20: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
570	x	x	1-heptanol	chlorobenzene	111-70-6	108-90-7	-6.78	-2.17	-4.61
571	x	x	triethyl_phosphate	n-hexane	78-40-0	110-54-3	-6.78	-2.33	-4.45
572	x	x	n,n-dimethylacetamide	n,n-dimethylacetamide	127-19-5	127-19-5	-6.77	-4.85	-1.92
573	x	x	methylbenzoate	decalin-mixture	93-58-3	91-17-8	-6.76	-1.57	-5.19
574	x	x	phenol	n-butylbenzene	108-95-2	104-51-8	-6.76	-1.71	-5.05
575	x	x	o-cresol	iodobenzene	95-48-7	591-50-4	-6.76	-2.48	-4.28
576	x		ethoxybenzene	ethoxybenzene	103-73-1	103-73-1	-6.75	-	-
577	x	x	1-heptanol	toluene	111-70-6	108-88-3	-6.75	-1.31	-5.44
578	x	x	acetophenone	n-octanol	98-86-2	111-87-5	-6.74	-3.42	-3.32
579	x	x	1-heptanol	xylene-mixture	111-70-6	1330-20-7	-6.74	-1.31	-5.43
580	x	x	methylbenzoate	2,2,4-trimethylpentane	93-58-3	540-84-1	-6.71	-1.35	-5.36
581	x	x	butylethanoate	chloroform	123-86-4	67-66-3	-6.71	-2.31	-4.40
582	x	x	m-cresol	chloroform	108-39-4	67-66-3	-6.70	-2.65	-4.05
583	x	x	1-heptanol	ethylbenzene	111-70-6	100-41-4	-6.70	-1.34	-5.36
584	x	x	1-pentanol	tributylphosphate	71-41-0	126-73-8	-6.69	-2.41	-4.28
585	x		m-dichlorobenzene	m-dichlorobenzene	541-73-1	541-73-1	-6.68	-	-
586	x	x	methyl_pentanoate	chloroform	624-24-8	67-66-3	-6.68	-2.34	-4.34
587	x	x	triethyl_phosphate	heptane	78-40-0	142-82-5	-6.67	-2.38	-4.29
588	x	x	1-hexanol	chloroform	111-27-3	67-66-3	-6.67	-2.12	-4.55
589	x	x	aniline	diisopropylether	62-53-3	108-20-3	-6.67	-2.39	-4.28
590	x	x	hexanoicacid	xylene-mixture	142-62-1	1330-20-7	-6.67	-1.81	-4.86
591	x	x	cyclopentanone	cyclopentanone	120-92-3	120-92-3	-6.66	-3.32	-3.34
592	x		4-butyrolactone	4-butyrolactone	96-48-0	96-48-0	-6.65	-	-
593	x		dimethylcyanamide	dimethylcyanamide	1467-79-4	1467-79-4	-6.65	-	-
594	x	x	1-heptanol	1,2-dibromoethane	111-70-6	106-93-4	-6.64	-2.06	-4.58
595	x	x	1-pentanol	1-pentanol	71-41-0	71-41-0	-6.63	-2.67	-3.97
596	x	x	nitrobenzene	n-octanol	98-95-3	111-87-5	-6.63	-3.72	-2.91
597	x		3-methyl-1-butanol	3-methyl-1-butanol	123-51-3	123-51-3	-6.63	-	-
598	x	x	propanoicacid	propanoicacid	79-09-4	79-09-4	-6.63	-2.53	-4.09
599	x	x	1-hexanol	butylethanoate	111-27-3	123-86-4	-6.62	-2.17	-4.45

TABLE B.21: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
600	x	x	nitrobenzene	cyclohexane	98-95-3	110-82-7	-6.62	-1.71	-4.91
601	x	x	pentanoicacid	chloroform	109-52-4	67-66-3	-6.61	-2.88	-3.73
602	x	x	4-methylpyridine	n-octanol	108-89-4	111-87-5	-6.60	-2.73	-3.87
603	x	x	N-methylaniline	carbontetrachloride	100-61-8	56-23-5	-6.58	-1.37	-5.21
604	x	x	methyl_hexanoate	dichloroethane	106-70-7	107-06-2	-6.57	-2.87	-3.70
605	x	x	pentylethanoate	benzene	628-63-7	71-43-2	-6.53	-1.47	-5.06
606	x	x	n,n-dimethylformamide	n,n-dimethylformamide	68-12-2	68-12-2	-6.52	-4.88	-1.64
607	x		1-8-cineole	1-8-cineole	470-82-6	470-82-6	-6.51	-	-
608	x	x	o-cresol	carbontetrachloride	95-48-7	56-23-5	-6.51	-1.51	-5.00
609	x	x	aniline	diethylether	62-53-3	60-29-7	-6.51	-2.71	-3.80
610	x	x	1-heptanol	o-dichlorobenzene	111-70-6	95-50-1	-6.50	-2.49	-4.01
611	x	x	1-heptanol	carbontetrachloride	111-70-6	56-23-5	-6.49	-1.23	-5.26
612	x	x	pentylethanoate	chlorobenzene	628-63-7	108-90-7	-6.49	-2.47	-4.02
613	x	x	pentanoicacid	dichloroethane	109-52-4	107-06-2	-6.47	-3.50	-2.97
614	x		1-2-3-trimethylbenzene	1-2-3-trimethylbenzene	526-73-8	526-73-8	-6.47	-	-
615	x		1,2,4-trimethylbenzene	1-2-3-trimethylbenzene	95-63-6	526-73-8	-6.47	-	-
616	x	x	3-methylaniline	cyclohexane	108-44-1	110-82-7	-6.47	-1.48	-4.99
617	x	x	5-nonanone	n-hexadecane	502-56-7	544-76-3	-6.46	-1.13	-5.33
618	x	x	2-heptanone	chlorobenzene	110-43-0	108-90-7	-6.46	-2.62	-3.84
619	x	x	2,4-dimethylpyridine	2,4-dimethylpyridine	108-47-4	108-47-4	-6.46	-2.51	-3.95
620	x	x	aniline	1-pentanol	62-53-3	71-41-0	-6.44	-3.76	-2.68
621	x	x	2-methylaniline	cyclohexane	95-53-4	110-82-7	-6.44	-1.42	-5.02
622	x	x	tert-butylbenzene	tert-butylbenzene	98-06-6	98-06-6	-6.43	-0.65	-5.78
623	x	x	aceticacid	cyclohexanone	64-19-7	108-94-1	-6.43	-3.84	-2.59
624	x	x	tert-butylbenzene	tert-butylbenzene	98-06-6	98-06-6	-6.43	-0.65	-5.78
625	x	x	N-methylaniline	decalin-mixture	100-61-8	91-17-8	-6.41	-1.35	-5.06
626	x	x	pentylethanoate	toluene	628-63-7	108-88-3	-6.41	-1.53	-4.88
627	x	x	2-ethylpyrazine	n-octanol	13925-00-3	111-87-5	-6.40	-2.67	-3.73
628	x	x	3-methylpyridine	n-octanol	108-99-6	111-87-5	-6.40	-2.66	-3.74
629	x	x	1-pentanol	n-octanol	71-41-0	111-87-5	-6.40	-2.50	-3.90

TABLE B.22: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
630	x	x	2,6-dimethylpyridine	benzene	108-48-5	71-43-2	-6.39	-1.08	-5.31
631	x	x	anisole	anisole	100-66-3	100-66-3	-6.39	-1.75	-4.64
632	x		dibutylamine	dibutylamine	111-92-2	111-92-2	-6.38	-	-
633	x	x	2-octanone	n-octanol	111-13-7	111-87-5	-6.38	-3.04	-3.34
634	x	x	methyl_hexanoate	toluene	106-70-7	108-88-3	-6.38	-1.53	-4.85
635	x	x	nitrobenzene	decalin-mixture	98-95-3	91-17-8	-6.36	-1.88	-4.48
636	x	x	2-heptanone	benzene	110-43-0	71-43-2	-6.36	-1.51	-4.85
637	x		diethyl_sulfite	diethyl_sulfite	623-81-4	623-81-4	-6.35	-	-
638	x	x	pentylethanoate	carbontetrachloride	628-63-7	56-23-5	-6.35	-1.44	-4.91
639	x	x	3-methylaniline	heptane	108-44-1	142-82-5	-6.35	-1.38	-4.97
640	x	x	pentylethanoate	bromobenzene	628-63-7	108-86-1	-6.35	-2.43	-3.92
641	x	x	N-methylaniline	cyclohexane	100-61-8	110-82-7	-6.33	-1.22	-5.11
642	x	x	anisole	anisole	100-66-3	100-66-3	-6.33	-1.75	-4.58
643	x		water	water	7732-18-5	7732-18-5	-6.32	-4.39	-1.93
644	x	x	1-4-dichlorobenzene	chloroform	106-46-7	67-66-3	-6.32	-1.48	-4.84
645	x	x	4-methylphenol	carbontetrachloride	106-44-5	56-23-5	-6.32	-1.59	-4.73
646	x	x	methylbenzoate	n-hexadecane	93-58-3	544-76-3	-6.31	-1.45	-4.86
647	x	x	4-methylaniline	cyclohexane	106-49-0	110-82-7	-6.30	-1.47	-4.83
648	x	x	n-octanol	n-hexadecane	111-87-5	544-76-3	-6.30	-1.00	-5.30
649	x	x	2-heptanone	toluene	110-43-0	108-88-3	-6.30	-1.58	-4.72
650	x	x	aceticacid	aniline	64-19-7	62-53-3	-6.30	-3.37	-2.93
651	x		hydrazine	hydrazine	302-01-2	302-01-2	-6.29	-	-
652	x	x	1-butanol	tributylphosphate	71-36-3	126-73-8	-6.28	-2.47	-3.81
653	x	x	benzonitrile	carbontetrachloride	100-47-0	56-23-5	-6.28	-2.12	-4.16
654	x	x	1-4-dichlorobenzene	carbontetrachloride	106-46-7	56-23-5	-6.28	-0.92	-5.36
655	x	x	piperidine	n-octanol	110-89-4	111-87-5	-6.27	-1.43	-4.84
656	x	x	methylbenzoate	benzene	93-58-3	71-43-2	-6.27	-1.63	-4.64
657	x	x	methyl_hexanoate	xylene-mixture	106-70-7	1330-20-7	-6.26	-1.54	-4.72
658	x	x	cyclohexanone	cyclohexanone	108-94-1	108-94-1	-6.26	-3.35	-2.90
659	x		acetic_anhydride	acetic_anhydride	108-24-7	108-24-7	-6.25	-	-

TABLE B.23: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
660	x		2-methoxyethyl_ether	2-methoxyethyl_ether	111-96-6	111-96-6	-6.25	-	-
661	x	x	bromobenzene	bromobenzene	108-86-1	108-86-1	-6.25	-1.44	-4.81
662	x	x	o-cresol	n-hexane	95-48-7	110-54-3	-6.25	-1.22	-5.03
663	x	x	cyclohexanone	cyclohexanone	108-94-1	108-94-1	-6.25	-3.35	-2.90
664	x	x	o-xylene	chloroform	95-47-6	67-66-3	-6.23	-1.10	-5.13
665	x	x	acetophenone	decalin-mixture	98-86-2	91-17-8	-6.23	-1.68	-4.55
666	x		2-ethoxyethanol	2-ethoxyethanol	110-80-5	110-80-5	-6.21	-	-
667	x	x	methyl_hexanoate	perfluorobenzene	106-70-7	392-56-3	-6.21	-1.30	-4.91
668	x	x	o-cresol	n-nonane	95-48-7	111-84-2	-6.20	-1.29	-4.91
669	x	x	propanoicacid	aniline	79-09-4	62-53-3	-6.20	-3.27	-2.93
670	x	x	4-methylphenol	n-octane	106-44-5	111-65-9	-6.19	-1.35	-4.84
671	x	x	methyl_hexanoate	isopropylbenzene	106-70-7	98-82-8	-6.19	-1.53	-4.66
672	x	x	pentylethanoate	xylene-mixture	628-63-7	1330-20-7	-6.19	-1.54	-4.65
673	x	x	N-methylaniline	n-hexadecane	100-61-8	544-76-3	-6.19	-1.24	-4.95
674	x	x	4-methylaniline	n-hexane	106-49-0	110-54-3	-6.18	-1.34	-4.84
675	x	x	4-methylpyridine	4-methylpyridine	108-89-4	108-89-4	-6.17	-2.82	-3.35
676	x	x	piperidine	2-methyl-1-propanol	110-89-4	78-83-1	-6.17	-1.56	-4.61
677	x	x	4-methylpyridine	benzene	108-89-4	71-43-2	-6.17	-1.39	-4.78
678	x	x	o-cresol	n-octane	95-48-7	111-65-9	-6.16	-1.28	-4.88
679	x		methyl_hexanoate	1-2-3-trimethylbenzene	106-70-7	526-73-8	-6.16	-	-
680	x	x	pentylethanoate	perfluorobenzene	628-63-7	392-56-3	-6.16	-1.30	-4.86
681	x	x	o-dichlorobenzene	n-hexadecane	95-50-1	544-76-3	-6.16	-0.88	-5.28
682	x	x	1-butanol	1-butanol	71-36-3	71-36-3	-6.16	-2.77	-3.39
683	x	x	3-methylaniline	n-octane	108-44-1	111-65-9	-6.15	-1.41	-4.74
684	x	x	2-heptanone	perfluorobenzene	110-43-0	392-56-3	-6.15	-1.33	-4.82
685	x	x	2-heptanone	xylene-mixture	110-43-0	1330-20-7	-6.15	-1.59	-4.56
686	x	x	2-methylpyridine	n-octanol	109-06-8	111-87-5	-6.14	-2.39	-3.75
687	x	x	phenol	carbontetrachloride	108-95-2	56-23-5	-6.14	-1.61	-4.53
688	x	x	acetophenone	n-hexadecane	98-86-2	544-76-3	-6.14	-1.54	-4.60
689	x	x	nitrobenzene	heptane	98-95-3	142-82-5	-6.14	-1.60	-4.54

TABLE B.24: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
690	x	x	acetophenone	heptane	98-86-2	142-82-5	-6.14	-1.42	-4.72
691	x	x	3-methylpyridine	3-methylpyridine	108-99-6	108-99-6	-6.13	-2.74	-3.40
692	x	x	pentylethanoate	isopropylbenzene	628-63-7	98-82-8	-6.13	-1.53	-4.60
693	x	x	benzaldehyde	n-octanol	100-52-7	111-87-5	-6.13	-3.26	-2.87
694	x	x	1-hexanol	benzene	111-27-3	71-43-2	-6.13	-1.32	-4.81
695	x	x	1-pentanol	ethylethanoate	71-41-0	141-78-6	-6.13	-2.22	-3.91
696	x	x	methyl_hexanoate	tert-butylbenzene	106-70-7	98-06-6	-6.13	-1.51	-4.62
697	x	x	1-hexanol	toluene	111-27-3	108-88-3	-6.12	-1.38	-4.74
698	x	x	propanoicacid	dibutylether	79-09-4	142-96-1	-6.11	-2.36	-3.75
699	x	x	o-dichlorobenzene	n-undecane	95-50-1	1120-21-4	-6.11	-0.85	-5.26
700	x	x	1-pentanol	diethylether	71-41-0	60-29-7	-6.11	-1.95	-4.16
701	x	x	benzaldehyde	carbontetrachloride	100-52-7	56-23-5	-6.11	-1.64	-4.47
702	x	x	2-heptanone	ethylbenzene	110-43-0	100-41-4	-6.10	-1.62	-4.48
703	x	x	phenol	tetrachloroethene	108-95-2	127-18-4	-6.10	-1.65	-4.45
704	x	x	aniline	carbontetrachloride	62-53-3	56-23-5	-6.10	-1.70	-4.40
705	x	x	aniline	xylene-mixture	62-53-3	1330-20-7	-6.10	-1.83	-4.27
706	x	x	benzonitrile	n-octanol	100-47-0	111-87-5	-6.09	-4.07	-2.02
707	x		pentylethanoate	1-2-3-trimethylbenzene	628-63-7	526-73-8	-6.09	-	-
708	x	x	nitrobenzene	n-hexane	98-95-3	110-54-3	-6.09	-1.57	-4.52
709	x	x	isopropylbenzene	isopropylbenzene	98-82-8	98-82-8	-6.09	-0.65	-5.44
710	x	x	benzaldehyde	diethylether	100-52-7	60-29-7	-6.08	-2.55	-3.53
711	x	x	1-hexanol	1,2-dibromoethane	111-27-3	106-93-4	-6.08	-2.16	-3.92
712	x	x	2-methylaniline	n-hexadecane	95-53-4	544-76-3	-6.08	-1.44	-4.64
713	x	x	2,6-dimethylpyridine	2,6-dimethylpyridine	108-48-5	108-48-5	-6.08	-2.06	-4.02
714	x	x	o-xylene	carbontetrachloride	95-47-6	56-23-5	-6.07	-0.65	-5.42
715	x	x	2-methylaniline	n-octane	95-53-4	111-65-9	-6.06	-1.35	-4.71
716	x	x	methyl_hexanoate	p-isopropyltoluene	106-70-7	99-87-6	-6.06	-1.44	-4.62
717	x	x	acetophenone	n-hexane	98-86-2	110-54-3	-6.05	-1.39	-4.66
718	x	x	isopropylbenzene	isopropylbenzene	98-82-8	98-82-8	-6.04	-0.65	-5.39
719	x	x	2,6-dimethylpyridine	2,6-dimethylpyridine	108-48-5	108-48-5	-6.04	-2.06	-3.98

TABLE B.25: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
720	x	x	4-methylaniline	n-hexadecane	106-49-0	544-76-3	-6.04	-1.49	-4.55
721	x	x	1-4-dichlorobenzene	n-hexadecane	106-46-7	544-76-3	-6.02	-0.84	-5.18
722	x	x	1-hexanol	dichloroethane	111-27-3	107-06-2	-6.02	-2.61	-3.41
723	x	x	pentylethanoate	p-isopropyltoluene	628-63-7	99-87-6	-6.02	-1.44	-4.58
724	x	x	1-heptanol	cyclohexane	111-70-6	110-82-7	-6.02	-1.09	-4.93
725	x	x	dipropylamine	n-octanol	142-84-7	111-87-5	-6.02	-1.45	-4.57
726	x		styrene	styrene	100-42-5	100-42-5	-6.01	-	-
727	x	x	o-dichlorobenzene	heptane	95-50-1	142-82-5	-6.01	-0.81	-5.20
728	x	x	pentanoicacid	benzene	109-52-4	71-43-2	-6.01	-1.83	-4.18
729	x		2-heptanone	1-2-3-trimethylbenzene	110-43-0	526-73-8	-6.01	-	-
730	x	x	o-dichlorobenzene	n-octanol	95-50-1	111-87-5	-6.01	-1.96	-4.05
731	x	x	o-cresol	heptane	95-48-7	142-82-5	-6.01	-1.25	-4.76
732	x	x	4-methylaniline	n-octane	106-49-0	111-65-9	-6.00	-1.40	-4.60
733	x	x	ethoxybenzene	cyclohexane	103-73-1	110-82-7	-6.00	-0.93	-5.07
734	x	x	4-methylphenol	n-decane	106-44-5	124-18-5	-6.00	-1.39	-4.61
735	x	x	2-heptanone	mesitylene	110-43-0	108-67-8	-5.99	-1.51	-4.48
736	x	x	butanoicacid	chloroform	107-92-6	67-66-3	-5.99	-2.85	-3.14
737	x	x	thiophenol	n-octanol	108-98-5	111-87-5	-5.99	-2.57	-3.42
738	x	x	2-heptanone	isopropylbenzene	110-43-0	98-82-8	-5.99	-1.58	-4.41
739	x		pentanonitrile	pentanonitrile	110-59-8	110-59-8	-5.98	-	-
740	x	x	thioanisole	chloroform	100-68-5	67-66-3	-5.98	-2.65	-3.33
741	x	x	methyl_ethyl_ketone	m-cresol	78-93-3	108-39-4	-5.98	-3.09	-2.89
742	x	x	1-hexanol	chlorobenzene	111-27-3	108-90-7	-5.98	-2.27	-3.71
743	x	x	2-methoxyethanol	2-methoxyethanol	109-86-4	109-86-4	-5.98	-3.07	-2.90
744	x	x	o-xylene	o-xylene	95-47-6	95-47-6	-5.96	-0.74	-5.22
745	x		pentamethylene_sulfide	pentamethylene_sulfide	1613-51-0	1613-51-0	-5.94	-	-
746	x	x	butylethanoate	dichloroethane	123-86-4	107-06-2	-5.93	-2.81	-3.12
747	x	x	2-heptanone	n-butylbenzene	110-43-0	104-51-8	-5.93	-1.57	-4.36
748	x		1-2-diaminoethane	1-2-diaminoethane	107-15-3	107-15-3	-5.92	-	-
749	x	x	1-hexanol	bromobenzene	111-27-3	108-86-1	-5.92	-2.23	-3.69

TABLE B.26: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
750	x	x	pentylethanoate	tert-butylbenzene	628-63-7	98-06-6	-5.92	-1.51	-4.41
751	x	x	m-cresol	n-hexadecane	108-39-4	544-76-3	-5.91	-1.45	-4.46
752	x		morpholine	morpholine	110-91-8	110-91-8	-5.90	-	-
753	x	x	1-pentanol	chloroform	71-41-0	67-66-3	-5.90	-2.04	-3.86
754	x	x	aceticacid	aceticacid	64-19-7	64-19-7	-5.89	-3.29	-2.60
755	x	x	pentanoicacid	toluene	109-52-4	108-88-3	-5.89	-1.91	-3.98
756	x	x	4-methylphenol	n-hexadecane	106-44-5	544-76-3	-5.88	-1.44	-4.44
757	x	x	2-heptanone	tert-butylbenzene	110-43-0	98-06-6	-5.88	-1.56	-4.32
758	x	x	pyridine	2-methyl-1-propanol	110-86-1	78-83-1	-5.87	-2.78	-3.09
759	x	x	dipropylamine	2-methyl-1-propanol	142-84-7	78-83-1	-5.87	-1.57	-4.30
760	x	x	2-ethylpyrazine	dibutylether	13925-00-3	142-96-1	-5.87	-1.73	-4.14
761	x	x	2-methylpyridine	benzene	109-06-8	71-43-2	-5.86	-1.19	-4.67
762	x	x	benzonitrile	decalin-mixture	100-47-0	91-17-8	-5.86	-2.09	-3.77
763	x	x	m-xylene	chloroform	108-38-3	67-66-3	-5.86	-1.05	-4.81
764	x	x	2-methyl-1-propanol	2-methyl-1-propanol	78-83-1	78-83-1	-5.86	-2.63	-3.22
765	x	x	1-hexanol	xylene-mixture	111-27-3	1330-20-7	-5.85	-1.39	-4.46
766	x	x	2-Hexanone	chlorobenzene	591-78-6	108-90-7	-5.84	-2.62	-3.22
767	x	x	methyl_pentanoate	benzene	624-24-8	71-43-2	-5.83	-1.47	-4.36
768	x	x	2-methoxyethanol	n-octanol	109-86-4	111-87-5	-5.83	-2.86	-2.97
769	x	x	butanoicacid	dichloroethane	107-92-6	107-06-2	-5.83	-3.47	-2.36
770	x	x	methyl_pentanoate	chlorobenzene	624-24-8	108-90-7	-5.83	-2.50	-3.33
771	x	x	2-octanone	n-hexadecane	111-13-7	544-76-3	-5.81	-1.35	-4.46
772	x	x	pentylethanoate	1-bromooctane	628-63-7	111-83-1	-5.81	-2.37	-3.44
773	x	x	1-4-dichlorobenzene	heptane	106-46-7	142-82-5	-5.81	-0.77	-5.04
774	x	x	piperazine	n-octanol	110-85-0	111-87-5	-5.80	-3.02	-2.78
775	x	x	2-methyl-1-propanol	2-methyl-1-propanol	78-83-1	78-83-1	-5.79	-2.63	-3.16
776	x	x	butylethanoate	benzene	123-86-4	71-43-2	-5.78	-1.47	-4.31
777	x	x	aniline	decalin-mixture	62-53-3	91-17-8	-5.78	-1.68	-4.10
778	x	x	1-pentanol	butylethanoate	71-41-0	123-86-4	-5.78	-2.08	-3.70
779	x	x	o-cresol	n-hexadecane	95-48-7	544-76-3	-5.78	-1.36	-4.42

TABLE B.27: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
780	x	x	p-xylene	p-xylene	106-42-3	106-42-3	-5.77	-0.63	-5.14
781	x	x	4-methylphenol	heptane	106-44-5	142-82-5	-5.77	-1.32	-4.45
782	x	x	1-butanol	ethylethanoate	71-36-3	141-78-6	-5.77	-2.28	-3.49
783	x	x	dibutylether	dibutylether	142-96-1	142-96-1	-5.76	-0.63	-5.13
784	x	x	1-1-1-3-3-3-hexafluoro-2-propanol	n-octanol	920-66-1	111-87-5	-5.76	-3.28	-2.48
785	x	x	2-methylpyridine	2-methylpyridine	109-06-8	109-06-8	-5.76	-2.39	-3.37
786	x	x	1-heptanol	n-hexane	111-70-6	110-54-3	-5.75	-1.00	-4.75
787	x	x	methyl_hexanoate	cyclohexane	106-70-7	110-82-7	-5.75	-1.29	-4.46
788	x	x	4-methylpyridine	acetonitrile	108-89-4	75-05-8	-5.75	-3.11	-2.64
789	x	x	butylethanoate	chlorobenzene	123-86-4	108-90-7	-5.74	-2.46	-3.28
790	x	x	aceticacid	diisopropylether	64-19-7	108-20-3	-5.73	-2.59	-3.14
791	x	x	1-hexanol	iodobenzene	111-27-3	591-50-4	-5.71	-2.09	-3.62
792	x	x	methyl_pentanoate	carbontetrachloride	624-24-8	56-23-5	-5.71	-1.44	-4.27
793	x	x	pentanoicacid	xylene-mixture	109-52-4	1330-20-7	-5.71	-1.92	-3.79
794	x	x	1-butanol	n-octanol	71-36-3	111-87-5	-5.71	-2.57	-3.14
795	x	x	benzaldehyde	cyclohexane	100-52-7	110-82-7	-5.71	-1.46	-4.25
796	x	x	m-xylene	carbontetrachloride	108-38-3	56-23-5	-5.71	-0.62	-5.09
797	x	x	2-methylpyridine	2-methylpyridine	109-06-8	109-06-8	-5.71	-2.39	-3.32
798	x	x	1-hexanol	o-dichlorobenzene	111-27-3	95-50-1	-5.70	-2.60	-3.10
799	x	x	1-butanol	diethylether	71-36-3	60-29-7	-5.69	-2.01	-3.68
800	x	x	1-4-dichlorobenzene	n-hexane	106-46-7	110-54-3	-5.69	-0.75	-4.94
801	x	x	2-octanone	heptane	111-13-7	142-82-5	-5.68	-1.24	-4.44
802	x	x	1-propanol	dimethylsulfoxide	71-23-8	67-68-5	-5.68	-2.94	-2.74
803	x	x	1-hexanol	ethylbenzene	111-27-3	100-41-4	-5.68	-1.42	-4.26
804	x	x	1-4-dichlorobenzene	n-octanol	106-46-7	111-87-5	-5.67	-1.79	-3.88
805	x	x	m-xylene	heptane	108-38-3	142-82-5	-5.67	-0.51	-5.16
806	x	x	ethylbenzene	ethylbenzene	100-41-4	100-41-4	-5.67	-0.67	-5.00
807	x	x	phenol	n-pentane	108-95-2	109-66-0	-5.67	-1.27	-4.40
808	x	x	ethylbenzene	carbontetrachloride	100-41-4	56-23-5	-5.67	-0.61	-5.06
809	x		3-methyl-2-butanone	3-methyl-2-butanone	563-80-4	563-80-4	-5.67	-	-

TABLE B.28: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
810	x		diethyl_carbonate	diethyl_carbonate	105-58-8	105-58-8	-5.66	-	-
811	x	x	chlorobenzene	chlorobenzene	108-90-7	108-90-7	-5.66	-1.47	-4.19
812	x	x	nitromethane	dimethylsulfoxide	75-52-5	67-68-5	-5.66	-4.84	-0.82
813	x	x	nitromethane	n,n-dimethylformamide	75-52-5	68-12-2	-5.66	-4.80	-0.86
814	x	x	thioanisole	carbontetrachloride	100-68-5	56-23-5	-5.66	-1.80	-3.86
815	x		1,1,1-trichloroethane	1,1,1-trichloroethane	71-55-6	71-55-6	-5.66	-	-
816	x	x	ethoxybenzene	n-octanol	103-73-1	111-87-5	-5.65	-2.18	-3.47
817	x	x	methyl_pentanoate	toluene	624-24-8	108-88-3	-5.65	-1.53	-4.12
818	x	x	ethoxybenzene	n-hexadecane	103-73-1	544-76-3	-5.64	-0.94	-4.70
819	x	x	methyl_hexanoate	n-hexane	106-70-7	110-54-3	-5.64	-1.18	-4.46
820	x	x	methyl_hexanoate	heptane	106-70-7	142-82-5	-5.63	-1.20	-4.43
821	x	x	n-octane	diethylether	111-65-9	60-29-7	-5.62	-0.12	-5.50
822	x	x	1-heptanol	n-pentane	111-70-6	109-66-0	-5.62	-0.96	-4.66
823	x	x	1-heptanol	n-decane	111-70-6	124-18-5	-5.62	-1.07	-4.55
824	x	x	nitromethane	acetonitrile	75-52-5	75-05-8	-5.62	-4.79	-0.83
825	x	x	1-heptanol	n-nonane	111-70-6	111-84-2	-5.62	-1.05	-4.57
826	x	x	1-heptanol	n-hexadecane	111-70-6	544-76-3	-5.62	-1.11	-4.51
827	x		acetylacetone	acetylacetone	123-54-6	123-54-6	-5.62	-	-
828	x	x	methyl_pentanoate	xylene-mixture	624-24-8	1330-20-7	-5.61	-1.54	-4.07
829	x	x	2-methoxyethanol	1-nonanol	109-86-4	143-08-8	-5.61	-2.79	-2.82
830	x	x	dipropyl_sulfide	n-hexadecane	111-47-7	544-76-3	-5.61	-0.74	-4.87
831	x	x	thiophenol	n-hexadecane	108-98-5	544-76-3	-5.61	-1.12	-4.49
832	x	x	4-methylpyridine	methanol	108-89-4	67-56-1	-5.60	-3.10	-2.50
833	x	x	1-heptanol	heptane	111-70-6	142-82-5	-5.60	-1.02	-4.58
834	x	x	phenol	n-nonane	108-95-2	111-84-2	-5.60	-1.39	-4.21
835	x	x	2-Hexanone	toluene	591-78-6	108-88-3	-5.60	-1.58	-4.02
836	x	x	4-methylphenol	2,2,4-trimethylpentane	106-44-5	540-84-1	-5.59	-1.34	-4.25
837	x	x	methyl_pentanoate	perfluorobenzene	624-24-8	392-56-3	-5.59	-1.30	-4.29
838	x	x	butylethanoate	carbontetrachloride	123-86-4	56-23-5	-5.59	-1.44	-4.15
839	x		2-methyl-2-butanol	2-methyl-2-butanol	75-85-4	75-85-4	-5.59	-	-

TABLE B.29: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
840	x	x	nitroethane	nitroethane	79-24-3	79-24-3	-5.58	-4.46	-1.12
841	x	x	butylethanoate	bromobenzene	123-86-4	108-86-1	-5.58	-2.42	-3.16
842	x	x	o-xylene	diethylether	95-47-6	60-29-7	-5.58	-1.04	-4.54
843	x	x	butylethanoate	toluene	123-86-4	108-88-3	-5.57	-1.53	-4.04
844	x	x	phenol	cyclohexane	108-95-2	110-82-7	-5.57	-1.44	-4.13
845	x	x	methyl_pentanoate	ethylbenzene	624-24-8	100-41-4	-5.56	-1.57	-3.99
846	x	x	1-propanol	1-propanol	71-23-8	71-23-8	-5.56	-2.80	-2.75
847	x	x	butylamine	1-pentanol	109-73-9	71-41-0	-5.55	-2.37	-3.18
848	x	x	2-Hexanone	perfluorobenzene	591-78-6	392-56-3	-5.55	-1.33	-4.22
849	x	x	thioanisole	decalin-mixture	100-68-5	91-17-8	-5.54	-1.78	-3.76
850	x	x	n-octane	1,2-dibromoethane	111-65-9	106-93-4	-5.54	-0.12	-5.42
851	x	x	o-xylene	cyclohexane	95-47-6	110-82-7	-5.54	-0.57	-4.97
852	x	x	benzonitrile	cyclohexane	100-47-0	110-82-7	-5.54	-1.90	-3.64
853	x	x	formicacid	formicacid	64-18-6	64-18-6	-5.54	-6.96	1.42
854	x	x	2-nitropropane	2-nitropropane	79-46-9	79-46-9	-5.54	-4.09	-1.45
855	x		tetrahydrothiophene	tetrahydrothiophene	110-01-0	110-01-0	-5.54	-	-
856	x	x	nitroethane	nitroethane	79-24-3	79-24-3	-5.53	-4.46	-1.07
857	x	x	methyl_hexanoate	n-octane	106-70-7	111-65-9	-5.53	-1.23	-4.30
858	x	x	benzaldehyde	n-hexane	100-52-7	110-54-3	-5.53	-1.34	-4.19
859	x	x	toluene	dichloromethane	108-88-3	75-09-2	-5.53	-1.31	-4.22
860	x	x	p-xylene	heptane	106-42-3	142-82-5	-5.52	-0.51	-5.01
861	x	x	butylethanoate	butylethanoate	123-86-4	123-86-4	-5.52	-2.36	-3.16
862	x	x	2,4-dimethylpyridine	n-hexadecane	108-47-4	544-76-3	-5.52	-1.12	-4.40
863	x	x	butylamine	2-butanol	109-73-9	78-92-2	-5.52	-2.39	-3.13
864	x	x	m-xylene	cyclohexane	108-38-3	110-82-7	-5.52	-0.55	-4.97
865	x	x	propanoicacid	1,2-dibromoethane	79-09-4	106-93-4	-5.52	-2.96	-2.56
866	x	x	pentylethanoate	n-hexane	628-63-7	110-54-3	-5.52	-1.18	-4.34
867	x	x	butylethanoate	perfluorobenzene	123-86-4	392-56-3	-5.52	-1.30	-4.22
868	x	x	o-xylene	heptane	95-47-6	142-82-5	-5.52	-0.53	-4.99
869	x	x	pyridine	pyridine	110-86-1	110-86-1	-5.52	-2.69	-2.83

TABLE B.30: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
870	x	x	2-butanol	2-butanol	78-92-2	78-92-2	-5.51	-2.52	-2.99
871	x	x	benzonitrile	n-hexadecane	100-47-0	544-76-3	-5.51	-1.93	-3.58
872	x	x	2,6-dimethylpyridine	cyclohexane	108-48-5	110-82-7	-5.51	-0.94	-4.57
873	x	x	methyl_hexanoate	n-nonane	106-70-7	111-84-2	-5.51	-1.24	-4.27
874	x	x	methyl_hexanoate	decalin-mixture	106-70-7	91-17-8	-5.51	-1.42	-4.09
875	x	x	2-ethylpyrazine	n-octane	13925-00-3	111-65-9	-5.51	-1.13	-4.38
876	x	x	pyrrole	chloroform	109-97-7	67-66-3	-5.50	-2.62	-2.88
877	x	x	benzaldehyde	heptane	100-52-7	142-82-5	-5.50	-1.37	-4.13
878	x	x	toluene	tetrahydrofuran	108-88-3	109-99-9	-5.50	-1.25	-4.25
879	x	x	phenol	n-hexane	108-95-2	110-54-3	-5.49	-1.31	-4.18
880	x	x	2-Hexanone	xylene-mixture	591-78-6	1330-20-7	-5.49	-1.59	-3.90
881	x	x	anisole	carbontetrachloride	100-66-3	56-23-5	-5.49	-1.10	-4.39
882	x	x	2-Hexanone	ethylbenzene	591-78-6	100-41-4	-5.49	-1.62	-3.87
883	x	x	butylethanoate	ethylbenzene	123-86-4	100-41-4	-5.48	-1.56	-3.92
884	x	x	2-butanol	2-butanol	78-92-2	78-92-2	-5.48	-2.52	-2.96
885	x	x	methyl_hexanoate	n-decane	106-70-7	124-18-5	-5.48	-1.26	-4.22
886	x	x	methylpropanoate	chloroform	554-12-1	67-66-3	-5.48	-2.31	-3.17
887	x	x	toluene	chloroform	108-88-3	67-66-3	-5.48	-1.07	-4.41
888	x	x	2-propen-1-ol	2-propen-1-ol	107-18-6	107-18-6	-5.48	-3.28	-2.20
889	x	x	phenol	n-octane	108-95-2	111-65-9	-5.47	-1.37	-4.10
890	x	x	2-heptanone	cyclohexane	110-43-0	110-82-7	-5.47	-1.32	-4.15
891	x	x	methyl_pentanoate	isopropylbenzene	624-24-8	98-82-8	-5.45	-1.53	-3.92
892	x	x	hydrogen_peroxide	nitrobenzene	7722-84-1	98-95-3	-5.45	-4.55	-0.90
893	x	x	4-butyrolactone	nitromethane	96-48-0	75-52-5	-5.45	-5.94	0.49
894	x	x	ethylbenzene	diethylether	100-41-4	60-29-7	-5.45	-0.99	-4.46
895	x	x	chlorobenzene	chloroform	108-90-7	67-66-3	-5.45	-1.37	-4.08
896	x	x	1-pentanol	dichloroethane	71-41-0	107-06-2	-5.45	-2.51	-2.94
897	x	x	tetrachloroethene	tetrachloroethene	127-18-4	127-18-4	-5.44	-0.35	-5.09
898	x	x	ethylbenzene	n-undecane	100-41-4	1120-21-4	-5.44	-0.53	-4.91
899	x	x	1-pentanol	1,2-dibromoethane	71-41-0	106-93-4	-5.44	-2.07	-3.37

TABLE B.31: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
900	x	x	benzaldehyde	n-hexadecane	100-52-7	544-76-3	-5.44	-1.49	-3.95
901	x	x	aniline	n-hexadecane	62-53-3	544-76-3	-5.44	-1.54	-3.90
902	x	x	pentylethanoate	decalin-mixture	628-63-7	91-17-8	-5.44	-1.42	-4.02
903	x	x	methyl_ethyl_ketone	chloroform	78-93-3	67-66-3	-5.43	-2.42	-3.01
904	x	x	o-xylene	acetonitrile	95-47-6	75-05-8	-5.43	-1.59	-3.84
905	x	x	aniline	n-hexane	62-53-3	110-54-3	-5.43	-1.38	-4.05
906	x	x	methyl_hexanoate	n-hexadecane	106-70-7	544-76-3	-5.43	-1.30	-4.13
907	x	x	2-Hexanone	1-chlorohexane	591-78-6	544-10-5	-5.42	-2.65	-2.77
908	x	x	chlorobenzene	diethylether	108-90-7	60-29-7	-5.42	-1.31	-4.11
909	x	x	nitromethane	nitromethane	75-52-5	75-52-5	-5.42	-4.80	-0.62
910	x	x	methyl_pentanoate	1-chlorohexane	624-24-8	544-10-5	-5.41	-2.53	-2.88
911	x	x	2-methoxyethanol	1-decanol	109-86-4	112-30-1	-5.41	-2.71	-2.70
912	x		methyl_pentanoate	1-2-3-trimethylbenzene	624-24-8	526-73-8	-5.41	-	-
913	x	x	methyl_pentanoate	tetrachloroethene	624-24-8	127-18-4	-5.41	-1.47	-3.94
914	x	x	1-heptanol	n-dodecane	111-70-6	112-40-3	-5.41	-1.09	-4.32
915	x	x	butanonitrile	butanonitrile	109-74-0	109-74-0	-5.40	-4.49	-0.91
916	x	x	2-heptanone	n-pentane	110-43-0	109-66-0	-5.40	-1.17	-4.23
917	x	x	butylamine	1-heptanol	109-73-9	111-70-6	-5.40	-2.28	-3.12
918	x	x	butylethanoate	xylene-mixture	123-86-4	1330-20-7	-5.40	-1.54	-3.86
919	x	x	methyl_pentanoate	tert-butylbenzene	624-24-8	98-06-6	-5.39	-1.52	-3.87
920	x	x	n-octane	tetrahydrofuran	111-65-9	109-99-9	-5.39	-0.14	-5.25
921	x		2-Hexanone	1-2-3-trimethylbenzene	591-78-6	526-73-8	-5.39	-	-
922	x	x	2-Hexanone	isopropylbenzene	591-78-6	98-82-8	-5.39	-1.58	-3.81
923	x	x	tetrachloroethene	tetrachloroethene	127-18-4	127-18-4	-5.39	-0.35	-5.04
924	x	x	n-octane	carbontetrachloride	111-65-9	56-23-5	-5.39	-0.07	-5.32
925	x	x	toluene	carbondisulfide	108-88-3	75-15-0	-5.39	-0.74	-4.65
926	x	x	phenol	decalin-mixture	108-95-2	91-17-8	-5.38	-1.59	-3.79
927	x	x	n-octane	diisopropylether	111-65-9	108-20-3	-5.38	-0.10	-5.28
928	x	x	aniline	heptane	62-53-3	142-82-5	-5.38	-1.41	-3.97
929	x	x	anisole	cyclohexane	100-66-3	110-82-7	-5.38	-0.98	-4.40

TABLE B.32: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
930	x	x	nitromethane	nitromethane	75-52-5	75-52-5	-5.38	-4.80	-0.58
931	x	x	methyl_pentanoate	n-butylbenzene	624-24-8	104-51-8	-5.37	-1.53	-3.84
932	x	x	ethylbenzene	acetonitrile	100-41-4	75-05-8	-5.36	-1.52	-3.84
933	x	x	2-heptanone	n-hexane	110-43-0	110-54-3	-5.36	-1.21	-4.15
934	x	x	butylethanoate	isopropylbenzene	123-86-4	98-82-8	-5.36	-1.53	-3.83
935	x	x	pentylethanoate	n-octane	628-63-7	111-65-9	-5.36	-1.23	-4.13
936	x	x	n-octane	benzene	111-65-9	71-43-2	-5.35	-0.08	-5.27
937	x	x	anisole	n-hexadecane	100-66-3	544-76-3	-5.35	-1.00	-4.35
938	x	x	anisole	heptane	100-66-3	142-82-5	-5.35	-0.92	-4.43
939	x	x	butylamine	1-nonanol	109-73-9	143-08-8	-5.35	-2.16	-3.19
940	x	x	methyl_hexanoate	n-pentadecane	106-70-7	629-62-9	-5.35	-1.30	-4.05
941	x	x	dipropylamine	xylene-mixture	142-84-7	1330-20-7	-5.35	-0.72	-4.63
942	x	x	butylethanoate	tetrachloroethene	123-86-4	127-18-4	-5.35	-1.46	-3.89
943	x	x	nitromethane	nitroethane	75-52-5	79-24-3	-5.35	-4.74	-0.61
944	x	x	2-Hexanone	mesitylene	591-78-6	108-67-8	-5.34	-1.51	-3.83
945	x	x	pyridine	n-octanol	110-86-1	111-87-5	-5.34	-2.58	-2.76
946	x	x	1-pentanol	bromoform	71-41-0	75-25-2	-5.34	-1.95	-3.39
947	x	x	2-heptanone	tetralin	110-43-0	119-64-2	-5.33	-1.81	-3.52
948	x	x	methyl_pentanoate	p-isopropyltoluene	624-24-8	99-87-6	-5.33	-1.44	-3.89
949	x	x	pyridine	4-methyl-2-pentanone	110-86-1	108-10-1	-5.33	-2.69	-2.64
950	x	x	methyl_pentanoate	1-fluorooctane	624-24-8	463-11-6	-5.33	-2.15	-3.18
951	x	x	phenol	heptane	108-95-2	142-82-5	-5.32	-1.34	-3.98
952	x	x	pyridine	butylethanoate	110-86-1	123-86-4	-5.31	-2.15	-3.16
953	x	x	1-hexanol	cyclohexane	111-27-3	110-82-7	-5.31	-1.16	-4.15
954	x	x	pentylethanoate	n-decane	628-63-7	124-18-5	-5.31	-1.26	-4.05
955	x	x	butanoicacid	benzene	107-92-6	71-43-2	-5.30	-1.82	-3.48
956	x	x	aceticacid	aceticacid	64-19-7	64-19-7	-5.30	-3.29	-2.01
957	x	x	butanoicacid	xylene-mixture	107-92-6	1330-20-7	-5.30	-1.90	-3.40
958	x	x	diethylamine	1-pentanol	109-89-7	71-41-0	-5.30	-1.57	-3.73
959	x	x	propanoicacid	o-nitrotoluene	79-09-4	88-72-2	-5.30	-3.88	-1.42

TABLE B.33: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
960	x	x	1-propanol	1-propanol	71-23-8	71-23-8	-5.29	-2.80	-2.49
961	x	x	2-pentanone	chlorobenzene	107-87-9	108-90-7	-5.29	-2.59	-2.70
962	x	x	1-butanol	chloroform	71-36-3	67-66-3	-5.28	-2.10	-3.18
963	x	x	pyrrole	n-octanol	109-97-7	111-87-5	-5.28	-3.23	-2.05
964	x	x	butylethanoate	n-butylbenzene	123-86-4	104-51-8	-5.28	-1.52	-3.76
965	x	x	m-xylene	acetonitrile	108-38-3	75-05-8	-5.28	-1.53	-3.75
966	x	x	nitromethane	tetrahydrothiophene-s,s-dioxide	75-52-5	126-33-0	-5.28	-4.83	-0.45
967	x	x	2-propen-1-ol	n-octanol	107-18-6	111-87-5	-5.27	-3.00	-2.27
968	x	x	toluene	fluorobenzene	108-88-3	462-06-6	-5.27	-1.13	-4.14
969	x	x	2-Hexanone	tert-butylbenzene	591-78-6	98-06-6	-5.27	-1.56	-3.71
970	x	x	p-xylene	acetonitrile	106-42-3	75-05-8	-5.27	-1.51	-3.76
971	x	x	2,6-dimethylpyridine	n-hexadecane	108-48-5	544-76-3	-5.27	-0.96	-4.31
972	x	x	3-3-dimethylbutanone	perfluorobenzene	75-97-8	392-56-3	-5.26	-1.26	-4.00
973	x	x	cyclopentanone	carbontetrachloride	120-92-3	56-23-5	-5.26	-1.63	-3.63
974	x	x	1-nitropropane	toluene	108-03-2	108-88-3	-5.25	-2.15	-3.10
975	x	x	3-3-dimethylbutanone	chlorobenzene	75-97-8	108-90-7	-5.25	-2.45	-2.80
976	x	x	ethylbenzene	n-decane	100-41-4	124-18-5	-5.25	-0.53	-4.72
977	x	x	1-pentanol	chlorobenzene	71-41-0	108-90-7	-5.25	-2.18	-3.07
978	x	x	2-heptanone	n-octane	110-43-0	111-65-9	-5.25	-1.26	-3.99
979	x	x	n-octane	chloroform	111-65-9	67-66-3	-5.25	-0.12	-5.13
980	x	x	butylethanoate	tert-butylbenzene	123-86-4	98-06-6	-5.25	-1.51	-3.74
981	x	x	ethanol	dimethylsulfoxide	64-17-5	67-68-5	-5.25	-2.95	-2.30
982	x	x	1-fluorooctane	n-hexadecane	463-11-6	544-76-3	-5.25	-0.57	-4.68
983	x	x	m-xylene	n-octanol	108-38-3	111-87-5	-5.25	-1.32	-3.93
984	x	x	ethanol	aceticacid	64-17-5	64-19-7	-5.25	-2.31	-2.94
985	x	x	n-octane	dibutylether	111-65-9	142-96-1	-5.24	-0.10	-5.14
986	x	x	butanoicacid	toluene	107-92-6	108-88-3	-5.24	-1.89	-3.35
987	x	x	m-xylene	n-hexadecane	108-38-3	544-76-3	-5.24	-0.56	-4.68
988	x	x	p-xylene	n-hexadecane	106-42-3	544-76-3	-5.24	-0.56	-4.68
989	x	x	dipropylamine	toluene	142-84-7	108-88-3	-5.24	-0.72	-4.52

TABLE B.34: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
990	x	x	pentanoicacid	heptane	109-52-4	142-82-5	-5.23	-1.50	-3.73
991	x	x	4-methylpyridine	cyclohexane	108-89-4	110-82-7	-5.23	-1.21	-4.02
992	x	x	diethylamine	chloroform	109-89-7	67-66-3	-5.23	-1.16	-4.07
993	x	x	ethanol	n,n-dimethylformamide	64-17-5	68-12-2	-5.23	-2.92	-2.31
994	x	x	toluene	diethylether	108-88-3	60-29-7	-5.23	-1.02	-4.21
995	x	x	1-butanol	butylethanoate	71-36-3	123-86-4	-5.23	-2.15	-3.08
996	x	x	butylamine	1-decanol	109-73-9	112-30-1	-5.22	-2.09	-3.13
997	x	x	o-xylene	n-hexane	95-47-6	110-54-3	-5.22	-0.52	-4.70
998	x	x	butylethanoate	1-fluorooctane	123-86-4	463-11-6	-5.22	-2.13	-3.09
999	x	x	butylethanoate	sec-butylbenzene	123-86-4	135-98-8	-5.22	-1.51	-3.71
1000	x		diisopropyl_sulfide	diisopropyl_sulfide	625-80-9	625-80-9	-5.22	-	-
1001	x	x	propylethanoate	benzene	109-60-4	71-43-2	-5.21	-1.46	-3.75
1002	x	x	chlorobenzene	carbontetrachloride	108-90-7	56-23-5	-5.21	-0.84	-4.37
1003	x	x	1-butanol	acetonitrile	71-36-3	75-05-8	-5.20	-2.92	-2.28
1004	x	x	pentylethanoate	n-hexadecane	628-63-7	544-76-3	-5.20	-1.31	-3.89
1005	x	x	4-heptanone	n-hexadecane	123-19-3	544-76-3	-5.20	-1.21	-3.99
1006	x	x	p-xylene	n-octanol	106-42-3	111-87-5	-5.19	-1.31	-3.88
1007	x	x	toluene	chlorobenzene	108-88-3	108-90-7	-5.18	-1.15	-4.03
1008	x	x	n-octane	dichloromethane	111-65-9	75-09-2	-5.18	-0.15	-5.03
1009	x	x	2-heptanone	n-decane	110-43-0	124-18-5	-5.18	-1.30	-3.88
1010	x	x	pentylethanoate	n-pentadecane	628-63-7	629-62-9	-5.18	-1.30	-3.88
1011	x	x	1-pentanol	toluene	71-41-0	108-88-3	-5.17	-1.32	-3.85
1012	x	x	toluene	toluene	108-88-3	108-88-3	-5.16	-0.68	-4.49
1013	x	x	n-octane	chlorobenzene	111-65-9	108-90-7	-5.16	-0.13	-5.03
1014	x	x	piperidine	xylene-mixture	110-89-4	1330-20-7	-5.15	-0.73	-4.42
1015	x	x	chlorobenzene	heptane	108-90-7	142-82-5	-5.15	-0.70	-4.45
1016	x	x	ethylbenzene	n-hexadecane	100-41-4	544-76-3	-5.15	-0.55	-4.60
1017	x	x	butylamine	benzylalcohol	109-73-9	100-51-6	-5.15	-2.31	-2.84
1018	x	x	diethylamine	1-butanol	109-89-7	71-36-3	-5.15	-1.60	-3.55
1019	x	x	phenol	n-hexadecane	108-95-2	544-76-3	-5.14	-1.46	-3.68

TABLE B.35: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1020	x	x	chlorobenzene	n-hexane	108-90-7	110-54-3	-5.14	-0.68	-4.46
1021	x	x	3-methylpyridine	cyclohexane	108-99-6	110-82-7	-5.14	-1.18	-3.96
1022	x	x	1-hexanol	n-hexane	111-27-3	110-54-3	-5.14	-1.06	-4.08
1023	x		piperidine	piperidine	110-89-4	110-89-4	-5.14	-	-
1024	x	x	pyridine	toluene	110-86-1	108-88-3	-5.13	-1.36	-3.77
1025	x	x	methyl_pentanoate	n-octanol	624-24-8	111-87-5	-5.13	-2.85	-2.28
1026	x	x	2-heptanone	n-hexadecane	110-43-0	544-76-3	-5.13	-1.34	-3.79
1027	x	x	methyl_ethyl_ketone	1,2-dibromoethane	78-93-3	106-93-4	-5.13	-2.46	-2.67
1028	x	x	chlorobenzene	n-undecane	108-90-7	1120-21-4	-5.12	-0.74	-4.38
1029	x	x	propanoicacid	dichloroethane	79-09-4	107-06-2	-5.12	-3.53	-1.59
1030	x	x	ethanol	n-methylformamide-mixture	64-17-5	123-39-7	-5.12	-3.03	-2.09
1031	x	x	m-xylene	2,2,4-trimethylpentane	108-38-3	540-84-1	-5.12	-0.52	-4.60
1032	x	x	toluene	toluene	108-88-3	108-88-3	-5.12	-0.68	-4.44
1033	x	x	toluene	carbontetrachloride	108-88-3	56-23-5	-5.12	-0.63	-4.49
1034	x	x	2-methylpyrazine	dibutylether	109-08-0	142-96-1	-5.12	-1.84	-3.28
1035	x	x	2-methoxyethanol	diethylether	109-86-4	60-29-7	-5.12	-2.25	-2.87
1036	x		phosphorus_oxychloride	phosphorus_oxychloride	10025-87-3	10025-87-3	-5.12	-	-
1037	x	x	propanonitrile	propanonitrile	107-12-0	107-12-0	-5.11	-4.56	-0.55
1038	x	x	nitromethane	aniline	75-52-5	62-53-3	-5.11	-3.96	-1.15
1039	x	x	nitromethane	pyridine	75-52-5	110-86-1	-5.11	-4.43	-0.68
1040	x	x	nitromethane	n-methylformamide-mixture	75-52-5	123-39-7	-5.11	-4.96	-0.15
1041	x	x	butylethanoate	1-bromooctane	123-86-4	111-83-1	-5.11	-2.36	-2.75
1042	x	x	toluene	pyridine	108-88-3	110-86-1	-5.10	-1.40	-3.70
1043	x	x	chlorobenzene	cyclohexane	108-90-7	110-82-7	-5.10	-0.75	-4.35
1044	x	x	1-pentanol	benzene	71-41-0	71-43-2	-5.10	-1.27	-3.83
1045	x	x	nitromethane	tetrahydrofuran	75-52-5	109-99-9	-5.09	-4.03	-1.06
1046	x	x	2-propanol	2-propanol	67-63-0	67-63-0	-5.09	-2.70	-2.38
1047	x	x	ethylbenzene	n-octanol	100-41-4	111-87-5	-5.08	-1.31	-3.77
1048	x	x	ethanol	ethanol	64-17-5	64-17-5	-5.08	-2.85	-2.23
1049	x	x	o-xylene	n-octanol	95-47-6	111-87-5	-5.07	-1.37	-3.70

TABLE B.36: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1050	x	x	1-pentanol	bromobenzene	71-41-0	108-86-1	-5.06	-2.14	-2.92
1051	x	x	toluene	xylene-mixture	108-88-3	1330-20-7	-5.06	-0.68	-4.38
1052	x	x	toluene	2-methylpyridine	108-88-3	109-06-8	-5.06	-1.34	-3.72
1053	x	x	nitromethane	2-methoxyethanol	75-52-5	109-86-4	-5.06	-4.57	-0.49
1054	x	x	nitromethane	ethylethanoate	75-52-5	141-78-6	-5.06	-3.82	-1.24
1055	x	x	propylethanoate	perfluorobenzene	109-60-4	392-56-3	-5.06	-1.29	-3.77
1056	x	x	toluene	ethylethanoate	108-88-3	141-78-6	-5.05	-1.17	-3.88
1057	x	x	nitromethane	dichloromethane	75-52-5	75-09-2	-5.05	-4.19	-0.86
1058	x	x	nitromethane	benzonitrile	75-52-5	100-47-0	-5.05	-4.71	-0.34
1059	x	x	butanoicacid	heptane	107-92-6	142-82-5	-5.05	-1.49	-3.56
1060	x	x	toluene	cyclohexanone	108-88-3	108-94-1	-5.05	-1.44	-3.61
1061	x	x	2-pentanone	2-pentanone	107-87-9	107-87-9	-5.04	-3.18	-1.86
1062	x	x	ethanol	ethanol	64-17-5	64-17-5	-5.04	-2.85	-2.19
1063	x	x	propylamine	1-butanol	107-10-8	71-36-3	-5.04	-2.40	-2.64
1064	x	x	methyl_pentanoate	cyclohexane	624-24-8	110-82-7	-5.04	-1.29	-3.75
1065	x	x	1-hexanol	carbontetrachloride	111-27-3	56-23-5	-5.04	-1.30	-3.74
1066	x	x	1,4-dioxane	n,n-dimethylformamide	123-91-1	68-12-2	-5.03	-3.02	-2.01
1067	x	x	piperidine	benzene	110-89-4	71-43-2	-5.03	-0.69	-4.34
1068	x	x	toluene	2,6-dimethylpyridine	108-88-3	108-48-5	-5.03	-1.24	-3.79
1069	x	x	3-pentanone	3-pentanone	96-22-0	96-22-0	-5.02	-3.01	-2.01
1070	x	x	1-propanol	n-octanol	71-23-8	111-87-5	-5.02	-2.56	-2.46
1071	x	x	n-octane	bromobenzene	111-65-9	108-86-1	-5.02	-0.13	-4.89
1072	x	x	2-pentanone	toluene	107-87-9	108-88-3	-5.02	-1.57	-3.45
1073	x	x	propylethanoate	propylethanoate	109-60-4	109-60-4	-5.02	-2.43	-2.59
1074	x	x	ethanol	2-methylpyridine	64-17-5	109-06-8	-5.01	-2.57	-2.44
1075	x	x	1,4-dioxane	n,n-dimethylacetamide	123-91-1	127-19-5	-5.01	-3.02	-1.99
1076	x	x	ethanol	1-propanol	64-17-5	71-23-8	-5.01	-2.81	-2.20
1077	x	x	cyclopentanone	n-octanol	120-92-3	111-87-5	-5.01	-3.17	-1.84
1078	x		t-butanol	t-butanol	75-65-0	75-65-0	-5.00	-	-
1079	x	x	3-3-dimethylbutanone	toluene	75-97-8	108-88-3	-5.00	-1.49	-3.51

TABLE B.37: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1080	x	x	propylethanoate	toluene	109-60-4	108-88-3	-5.00	-1.52	-3.48
1081	x	x	anisole	decalin-mixture	100-66-3	91-17-8	-5.00	-1.09	-3.91
1082	x	x	n-octane	fluorobenzene	111-65-9	462-06-6	-4.99	-0.13	-4.86
1083	x	x	toluene	iodobenzene	108-88-3	591-50-4	-4.99	-1.05	-3.94
1084	x		toluene	ethoxybenzene	108-88-3	103-73-1	-4.99	-	-
1085	x	x	m-xylene	n-hexane	108-38-3	110-54-3	-4.99	-0.50	-4.49
1086	x	x	chlorobenzene	n-hexadecane	108-90-7	544-76-3	-4.99	-0.76	-4.23
1087	x	x	ethylbenzene	n-hexane	100-41-4	110-54-3	-4.99	-0.49	-4.50
1088	x	x	3-3-dimethylbutanone	1-chlorohexane	75-97-8	544-10-5	-4.98	-2.48	-2.50
1089	x	x	toluene	triethylamine	108-88-3	121-44-8	-4.98	-0.68	-4.30
1090	x	x	1-hexanol	n-nonane	111-27-3	111-84-2	-4.97	-1.12	-3.85
1091	x	x	ethylbenzene	cyclohexane	100-41-4	110-82-7	-4.97	-0.54	-4.43
1092	x	x	1-hexanol	n-pentane	111-27-3	109-66-0	-4.97	-1.02	-3.95
1093	x	x	1-hexanol	n-decane	111-27-3	124-18-5	-4.97	-1.14	-3.83
1094	x	x	trichloroethene	trichloroethene	79-01-6	79-01-6	-4.97	-0.81	-4.16
1095	x		ethyl_propanoate	ethyl_propanoate	105-37-3	105-37-3	-4.96	-	-
1096	x	x	dipropylamine	diethylether	142-84-7	60-29-7	-4.96	-1.09	-3.87
1097	x	x	methyl_pentanoate	n-pentane	624-24-8	109-66-0	-4.96	-1.14	-3.82
1098	x	x	butylethanoate	n-octanol	123-86-4	111-87-5	-4.96	-2.80	-2.16
1099	x	x	toluene	anisole	108-88-3	100-66-3	-4.95	-1.02	-3.93
1100	x	x	toluene	benzonitrile	108-88-3	100-47-0	-4.95	-1.51	-3.44
1101	x	x	propylethanoate	ethylbenzene	109-60-4	100-41-4	-4.95	-1.56	-3.39
1102	x	x	butylethanoate	cyclohexane	123-86-4	110-82-7	-4.94	-1.29	-3.65
1103	x	x	methyl_pentanoate	n-hexane	624-24-8	110-54-3	-4.94	-1.18	-3.76
1104	x	x	propylethanoate	bromobenzene	109-60-4	108-86-1	-4.93	-2.41	-2.52
1105	x	x	ethylethanoate	dichloroethane	141-78-6	107-06-2	-4.93	-2.81	-2.12
1106	x	x	1-pentanol	o-dichlorobenzene	71-41-0	95-50-1	-4.93	-2.51	-2.42
1107	x	x	butanoicacid	isopropylbenzene	107-92-6	98-82-8	-4.93	-1.89	-3.04
1108	x	x	chlorobenzene	n-decane	108-90-7	124-18-5	-4.93	-0.73	-4.20
1109	x	x	1-hexanol	n-hexadecane	111-27-3	544-76-3	-4.92	-1.18	-3.74

TABLE B.38: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1110	x	x	methyl_pentanoate	heptane	624-24-8	142-82-5	-4.92	-1.20	-3.72
1111	x	x	nitromethane	acetophenone	75-52-5	98-86-2	-4.92	-4.57	-0.35
1112	x	x	1,4-dioxane	2-methoxyethanol	123-91-1	109-86-4	-4.91	-2.86	-2.05
1113	x	x	3-methylpyridine	n-hexadecane	108-99-6	544-76-3	-4.91	-1.20	-3.71
1114	x	x	3-3-dimethylbutanone	xylene-mixture	75-97-8	1330-20-7	-4.91	-1.50	-3.41
1115	x	x	toluene	diisopropylether	108-88-3	108-20-3	-4.91	-0.90	-4.01
1116	x	x	nitromethane	nitrobenzene	75-52-5	98-95-3	-4.90	-4.79	-0.11
1117	x	x	toluene	acetophenone	108-88-3	98-86-2	-4.90	-1.46	-3.44
1118	x	x	1-propanol	diethylether	71-23-8	60-29-7	-4.90	-2.00	-2.90
1119	x	x	1,4-dioxane	dimethylsulfoxide	123-91-1	67-68-5	-4.90	-3.05	-1.85
1120	x	x	toluene	cyclohexane	108-88-3	110-82-7	-4.90	-0.56	-4.34
1121	x	x	1-hexanol	heptane	111-27-3	142-82-5	-4.89	-1.08	-3.81
1122	x	x	4-methylpyridine	n-hexadecane	108-89-4	544-76-3	-4.89	-1.23	-3.66
1123	x	x	aceticacid	dichloroethane	64-19-7	107-06-2	-4.89	-3.64	-1.25
1124	x	x	acetonitrile	acetonitrile	75-05-8	75-05-8	-4.89	-4.83	-0.06
1125	x	x	propylamine	1-pentanol	107-10-8	71-41-0	-4.88	-2.36	-2.52
1126	x	x	nitromethane	aceticacid	75-52-5	64-19-7	-4.88	-3.87	-1.01
1127	x	x	butylethanoate	n-pentane	123-86-4	109-66-0	-4.88	-1.14	-3.74
1128	x	x	methyl_ethyl_ketone	aniline	78-93-3	62-53-3	-4.87	-2.74	-2.13
1129	x	x	methylpropanoate	dichloroethane	554-12-1	107-06-2	-4.87	-2.83	-2.04
1130	x	x	2-propen-1-ol	diethylether	107-18-6	60-29-7	-4.87	-2.34	-2.53
1131	x	x	2-pentanone	xylene-mixture	107-87-9	1330-20-7	-4.87	-1.58	-3.29
1132	x	x	toluene	dibutylether	108-88-3	142-96-1	-4.87	-0.84	-4.03
1133	x	x	1,4-dioxane	n-methylformamide-mixture	123-91-1	123-39-7	-4.86	-3.13	-1.73
1134	x	x	1-hexanol	n-octane	111-27-3	111-65-9	-4.86	-1.10	-3.76
1135	x	x	diethylamine	2-methyl-1-propanol	109-89-7	78-83-1	-4.86	-1.59	-3.27
1136	x	x	methyl_pentanoate	n-octane	624-24-8	111-65-9	-4.86	-1.23	-3.63
1137	x	x	toluene	diphenylether	108-88-3	101-84-8	-4.86	-0.95	-3.91
1138	x	x	methanol	methanol	67-56-1	67-56-1	-4.86	-2.96	-1.90
1139	x	x	methyl_pentanoate	n-nonane	624-24-8	111-84-2	-4.85	-1.24	-3.61

TABLE B.39: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1140	x	x	2-pentanone	ethylbenzene	107-87-9	100-41-4	-4.85	-1.60	-3.25
1141	x		pyrrolidine	pyrrolidine	123-75-1	123-75-1	-4.84	-	-
1142	x	x	ethanol	benzylalcohol	64-17-5	100-51-6	-4.84	-2.66	-2.18
1143	x	x	ethanol	2-propanol	64-17-5	67-63-0	-4.84	-2.79	-2.05
1144	x	x	propylethanoate	1-chlorohexane	109-60-4	544-10-5	-4.84	-2.48	-2.36
1145	x	x	2-pentanone	1-chlorohexane	107-87-9	544-10-5	-4.84	-2.63	-2.21
1146	x	x	2-pentanone	isopropylbenzene	107-87-9	98-82-8	-4.84	-1.57	-3.27
1147	x	x	dimethyl_disulfide	n-hexadecane	624-92-0	544-76-3	-4.84	-1.11	-3.73
1148	x	x	aniline	n-octane	62-53-3	111-65-9	-4.84	-1.44	-3.40
1149	x	x	butylethanoate	heptane	123-86-4	142-82-5	-4.83	-1.20	-3.63
1150	x		2-pentanone	1-2-3-trimethylbenzene	107-87-9	526-73-8	-4.83	-	-
1151	x	x	methyl_pentanoate	decalin-mixture	624-24-8	91-17-8	-4.83	-1.42	-3.41
1152	x	x	chlorobenzene	1-decanol	108-90-7	112-30-1	-4.83	-1.59	-3.24
1153	x	x	toluene	n-octane	108-88-3	111-65-9	-4.82	-0.53	-4.29
1154	x	x	toluene	n-undecane	108-88-3	1120-21-4	-4.81	-0.55	-4.26
1155	x	x	3-3-dimethylbutanone	isopropylbenzene	75-97-8	98-82-8	-4.81	-1.49	-3.32
1156	x	x	pyridine	diethylether	110-86-1	60-29-7	-4.81	-2.01	-2.80
1157	x	x	2,2,2-trifluoroethanol	n-octanol	75-89-8	111-87-5	-4.81	-3.82	-0.99
1158	x	x	2-pentanone	carbontetrachloride	107-87-9	56-23-5	-4.81	-1.47	-3.34
1159	x	x	n,n-dimethylacetamide	heptane	127-19-5	142-82-5	-4.80	-1.76	-3.04
1160	x		3-3-dimethylbutanone	1-2-3-trimethylbenzene	75-97-8	526-73-8	-4.80	-	-
1161	x	x	methyl_ethyl_ketone	aceticacid	78-93-3	64-19-7	-4.80	-2.67	-2.13
1162	x	x	propylamine	1-heptanol	107-10-8	111-70-6	-4.80	-2.27	-2.53
1163	x	x	t-butanol	diethylether	75-65-0	60-29-7	-4.80	-1.83	-2.97
1164	x	x	propylethanoate	tetrachloroethene	109-60-4	127-18-4	-4.80	-1.46	-3.34
1165	x	x	butylethanoate	n-octane	123-86-4	111-65-9	-4.80	-1.23	-3.57
1166	x	x	3-3-dimethylbutanone	tert-butylbenzene	75-97-8	98-06-6	-4.79	-1.47	-3.32
1167	x	x	2-Hexanone	n-pentane	591-78-6	109-66-0	-4.79	-1.17	-3.62
1168	x	x	t-butanol	n-octanol	75-65-0	111-87-5	-4.78	-2.37	-2.41
1169	x	x	toluene	heptane	108-88-3	142-82-5	-4.78	-0.52	-4.26

TABLE B.40: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1170	x	x	3-3-dimethylbutanone	n-butylbenzene	75-97-8	104-51-8	-4.77	-1.48	-3.29
1171	x	x	2-Hexanone	cyclohexane	591-78-6	110-82-7	-4.77	-1.32	-3.45
1172	x	x	propylamine	n-octanol	107-10-8	111-87-5	-4.77	-2.21	-2.56
1173	x	x	3-3-dimethylbutanone	mesitylene	75-97-8	108-67-8	-4.77	-1.42	-3.35
1174	x	x	propanoicacid	benzene	79-09-4	71-43-2	-4.75	-1.87	-2.88
1175	x	x	diethylamine	n-octanol	109-89-7	111-87-5	-4.75	-1.46	-3.29
1176	x	x	aceticacid	chloroform	64-19-7	67-66-3	-4.74	-3.01	-1.73
1177	x	x	methanol	1-butanol	67-56-1	71-36-3	-4.73	-2.83	-1.90
1178	x	x	methyl_ethyl_ketone	nitroethane	78-93-3	79-24-3	-4.73	-3.35	-1.38
1179	x	x	methyl_ethyl_ketone	acetonitrile	78-93-3	75-05-8	-4.73	-3.40	-1.33
1180	x	x	1-pentanol	carbontetrachloride	71-41-0	56-23-5	-4.73	-1.24	-3.49
1181	x	x	diethylsulfide	diethylsulfide	352-93-2	352-93-2	-4.73	-1.78	-2.95
1182	x	x	1-butanol	bromoform	71-36-3	75-25-2	-4.72	-2.01	-2.71
1183	x	x	n-octane	ethylethanoate	111-65-9	141-78-6	-4.72	-0.13	-4.59
1184	x	x	1-pentanol	xylene-mixture	71-41-0	1330-20-7	-4.72	-1.33	-3.39
1185	x	x	propylethanoate	tert-butylbenzene	109-60-4	98-06-6	-4.72	-1.50	-3.22
1186	x	x	propanoicacid	xylene-mixture	79-09-4	1330-20-7	-4.72	-1.96	-2.76
1187	x	x	2-pentanone	tert-butylbenzene	107-87-9	98-06-6	-4.72	-1.55	-3.17
1188	x	x	2-Hexanone	2,2,4-trimethylpentane	591-78-6	540-84-1	-4.72	-1.26	-3.46
1189	x	x	n-octane	iodobenzene	111-65-9	591-50-4	-4.72	-0.12	-4.60
1190	x	x	methyl_ethyl_ketone	nitromethane	78-93-3	75-52-5	-4.72	-3.40	-1.32
1191	x	x	hydrogen_peroxide	chloroform	7722-84-1	67-66-3	-4.70	-3.30	-1.40
1192	x	x	4-butyrolactone	toluene	96-48-0	108-88-3	-4.70	-2.84	-1.86
1193	x	x	2-methylpyrazine	n-octane	109-08-0	111-65-9	-4.70	-1.21	-3.49
1194	x	x	methyl_pentanoate	n-hexadecane	624-24-8	544-76-3	-4.69	-1.31	-3.38
1195	x	x	dichloroethane	dichloroethane	107-06-2	107-06-2	-4.69	-2.34	-2.35
1196	x	x	toluene	2,2,4-trimethylpentane	108-88-3	540-84-1	-4.68	-0.53	-4.15
1197	x	x	1-nitrobutane	n-hexadecane	627-05-4	544-76-3	-4.66	-1.84	-2.82
1198	x	x	propylamine	1-nonanol	107-10-8	143-08-8	-4.66	-2.15	-2.51
1199	x	x	pyridine	dibutylether	110-86-1	142-96-1	-4.65	-1.67	-2.98

TABLE B.41: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1200	x	x	toluene	n-decane	108-88-3	124-18-5	-4.65	-0.55	-4.10
1201	x	x	1-propanol	acetonitrile	71-23-8	75-05-8	-4.65	-2.91	-1.74
1202	x	x	2-Hexanone	tetralin	591-78-6	119-64-2	-4.64	-1.81	-2.83
1203	x		n-octane	methyl_ethyl_ketone	111-65-9	78-93-3	-4.64	-	-
1204	x	x	1-nitrobutane	n-hexane	627-05-4	110-54-3	-4.64	-1.66	-2.98
1205	x	x	benzene	chloroform	71-43-2	67-66-3	-4.64	-1.07	-3.57
1206	x	x	fluorobenzene	fluorobenzene	462-06-6	462-06-6	-4.64	-1.24	-3.39
1207	x	x	tetrachloroethene	n-undecane	127-18-4	1120-21-4	-4.63	-0.30	-4.33
1208	x	x	propylethanoate	carbondisulfide	109-60-4	75-15-0	-4.63	-1.65	-2.98
1209	x	x	nitromethane	fluorobenzene	75-52-5	462-06-6	-4.62	-3.71	-0.91
1210	x	x	n-octane	anisole	111-65-9	100-66-3	-4.62	-0.12	-4.50
1211	x	x	propylethanoate	sec-butylbenzene	109-60-4	135-98-8	-4.62	-1.50	-3.12
1212	x	x	2-propanol	n-octanol	67-63-0	111-87-5	-4.62	-2.47	-2.15
1213	x	x	methylpropanoate	toluene	554-12-1	108-88-3	-4.62	-1.52	-3.10
1214	x	x	pentanoicacid	n-hexadecane	109-52-4	544-76-3	-4.61	-1.63	-2.98
1215	x	x	1,4-dioxane	1-propanol	123-91-1	71-23-8	-4.61	-2.91	-1.70
1216	x	x	methyl_ethyl_ketone	pyridine	78-93-3	110-86-1	-4.61	-3.11	-1.50
1217	x	x	2-Hexanone	n-decane	591-78-6	124-18-5	-4.61	-1.30	-3.31
1218	x	x	fluorobenzene	fluorobenzene	462-06-6	462-06-6	-4.60	-1.24	-3.36
1219	x	x	methyl_ethyl_ketone	fluorobenzene	78-93-3	462-06-6	-4.60	-2.55	-2.05
1220	x	x	2-Hexanone	n-octane	591-78-6	111-65-9	-4.60	-1.26	-3.34
1221	x	x	propylamine	1-decanol	107-10-8	112-30-1	-4.59	-2.09	-2.50
1222	x	x	2-Hexanone	n-nonane	591-78-6	111-84-2	-4.59	-1.28	-3.31
1223	x	x	methyl_pentanoate	n-pentadecane	624-24-8	629-62-9	-4.59	-1.30	-3.29
1224	x	x	methylbutanoate	n-octanol	623-42-7	111-87-5	-4.59	-2.83	-1.76
1225	x	x	methylpropanoate	methylpropanoate	554-12-1	554-12-1	-4.58	-2.52	-2.07
1226	x	x	methyl_ethyl_ketone	benzonitrile	78-93-3	100-47-0	-4.58	-3.33	-1.25
1227	x	x	4-butyrolactone	ethanol	96-48-0	64-17-5	-4.58	-5.81	1.23
1228	x	x	2,2,4-trimethylpentane	2,2,4-trimethylpentane	540-84-1	540-84-1	-4.58	-0.09	-4.49
1229	x	x	methyl_ethyl_ketone	benzylalcohol	78-93-3	100-51-6	-4.57	-3.09	-1.48

TABLE B.42: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1230	x	x	propanoicacid	toluene	79-09-4	108-88-3	-4.57	-1.95	-2.62
1231	x	x	dipropylamine	n-hexadecane	142-84-7	544-76-3	-4.57	-0.61	-3.96
1232	x	x	ethylethanoate	bromobenzene	141-78-6	108-86-1	-4.57	-2.42	-2.15
1233	x	x	n-octane	cyclohexanone	111-65-9	108-94-1	-4.57	-0.16	-4.41
1234	x	x	ethanol	tetrahydrofuran	64-17-5	109-99-9	-4.56	-2.42	-2.14
1235	x	x	ethylethanoate	perfluorobenzene	141-78-6	392-56-3	-4.56	-1.29	-3.27
1236	x	x	benzene	benzene	71-43-2	71-43-2	-4.56	-0.65	-3.91
1237	x	x	propylethanoate	n-octanol	109-60-4	111-87-5	-4.55	-2.78	-1.77
1238	x	x	1-2-dimethoxyethane	n-octanol	110-71-4	111-87-5	-4.55	-2.22	-2.33
1239	x	x	2-Hexanone	heptane	591-78-6	142-82-5	-4.55	-1.23	-3.32
1240	x	x	methylethanoate	dichloroethane	79-20-9	107-06-2	-4.55	-2.99	-1.56
1241	x	x	methylpropanoate	chlorobenzene	554-12-1	108-90-7	-4.55	-2.47	-2.08
1242	x	x	benzene	benzene	71-43-2	71-43-2	-4.55	-0.65	-3.90
1243	x	x	methyl_ethyl_ketone	tetrahydrofuran	78-93-3	109-99-9	-4.54	-2.80	-1.74
1244	x	x	toluene	n-hexadecane	108-88-3	544-76-3	-4.54	-0.57	-3.97
1245	x	x	aceticacid	bromoform	64-19-7	75-25-2	-4.54	-2.89	-1.65
1246	x	x	2-propanol	methanol	67-63-0	67-56-1	-4.53	-2.81	-1.72
1247	x	x	nitromethane	benzylalcohol	75-52-5	100-51-6	-4.53	-4.41	-0.12
1248	x		dipropyl_ether	dipropyl_ether	111-43-3	111-43-3	-4.53	-	-
1249	x	x	methyl_ethyl_ketone	2-methylpyridine	78-93-3	109-06-8	-4.52	-2.98	-1.54
1250	x	x	1-propanol	butylethanoate	71-23-8	123-86-4	-4.52	-2.14	-2.38
1251	x	x	ethylethanoate	ethylethanoate	141-78-6	141-78-6	-4.50	-2.50	-2.00
1252	x	x	triethylamine	triethylamine	121-44-8	121-44-8	-4.50	-0.45	-4.05
1253	x	x	n-octane	pyridine	111-65-9	110-86-1	-4.50	-0.16	-4.34
1254	x	x	benzene	carbontetrachloride	71-43-2	56-23-5	-4.50	-0.63	-3.87
1255	x		methyl_ethyl_ketone	methyl_ethyl_ketone	78-93-3	78-93-3	-4.50	-	-
1256	x	x	nitromethane	benzene	75-52-5	71-43-2	-4.50	-2.26	-2.24
1257	x	x	ethylamine	1-butanol	75-04-7	71-36-3	-4.50	-2.39	-2.11
1258	x	x	1-nitropropane	carbondisulfide	108-03-2	75-15-0	-4.50	-2.34	-2.16
1259	x	x	butylamine	butylamine	109-73-9	109-73-9	-4.49	-1.78	-2.71

TABLE B.43: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1260	x	x	1,4-dioxane	2-propanol	123-91-1	67-63-0	-4.49	-2.89	-1.60
1261	x	x	1,1,2-trichloroethane	n-hexadecane	79-00-5	544-76-3	-4.49	-1.20	-3.29
1262	x	x	aceticacid	decalin-mixture	64-19-7	91-17-8	-4.49	-1.87	-2.62
1263	x		2-methyltetrahydrofuran	2-methyltetrahydrofuran	96-47-9	96-47-9	-4.49	-	-
1264	x	x	propylethanoate	1-bromooctane	109-60-4	111-83-1	-4.48	-2.35	-2.13
1265	x	x	t-butanol	chloroform	75-65-0	67-66-3	-4.48	-1.92	-2.56
1266	x		4-butyrolactone	methyl_ethyl_ketone	96-48-0	78-93-3	-4.47	-	-
1267	x	x	methyl_ethyl_ketone	chlorobenzene	78-93-3	108-90-7	-4.47	-2.59	-1.88
1268	x	x	methyl_ethyl_ketone	benzene	78-93-3	71-43-2	-4.46	-1.50	-2.96
1269	x		ethanol	methyl_ethyl_ketone	64-17-5	78-93-3	-4.46	-	-
1270	x	x	2-Hexanone	n-hexadecane	591-78-6	544-76-3	-4.45	-1.34	-3.11
1271	x		nitromethane	ethoxybenzene	75-52-5	103-73-1	-4.45	-	-
1272	x	x	1-butanol	benzene	71-36-3	71-43-2	-4.45	-1.31	-3.14
1273	x	x	ethanol	aniline	64-17-5	62-53-3	-4.45	-2.37	-2.08
1274	x	x	1-nitropropane	n-octanol	108-03-2	111-87-5	-4.44	-3.93	-0.51
1275	x	x	butylamine	anisole	109-73-9	100-66-3	-4.44	-1.71	-2.73
1276	x	x	triethylamine	triethylamine	121-44-8	121-44-8	-4.44	-0.45	-3.99
1277	x	x	3-3-dimethylbutanone	n-pentane	75-97-8	109-66-0	-4.43	-1.10	-3.33
1278	x	x	methylpropanoate	carbontetrachloride	554-12-1	56-23-5	-4.43	-1.42	-3.01
1279	x	x	dimethylamine	2-methyl-1-propanol	124-40-3	78-83-1	-4.43	-1.77	-2.66
1280	x	x	ethanol	acetonitrile	64-17-5	75-05-8	-4.43	-2.91	-1.52
1281	x	x	methyl_ethyl_ketone	anisole	78-93-3	100-66-3	-4.43	-2.31	-2.12
1282	x	x	methyl_ethyl_ketone	cyclohexanone	78-93-3	108-94-1	-4.42	-3.18	-1.24
1283	x	x	perfluorobenzene	perfluorobenzene	392-56-3	392-56-3	-4.42	-0.53	-3.89
1284	x	x	acetone	chloroform	67-64-1	67-66-3	-4.42	-2.60	-1.82
1285	x	x	propanoicacid	decalin-mixture	79-09-4	91-17-8	-4.42	-1.81	-2.61
1286	x	x	cyclopentanol	n-hexadecane	96-41-3	544-76-3	-4.42	-1.08	-3.34
1287	x	x	ethylethanoate	toluene	141-78-6	108-88-3	-4.41	-1.53	-2.88
1288	x	x	1-propanol	chloroform	71-23-8	67-66-3	-4.41	-2.09	-2.32
1289	x	x	ethanol	cyclohexanone	64-17-5	108-94-1	-4.41	-2.74	-1.67

TABLE B.44: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1290	x	x	toluene	1-iodohexadecane	108-88-3	544-77-4	-4.41	-0.92	-3.49
1291	x	x	ethanol	diethylether	64-17-5	60-29-7	-4.41	-2.01	-2.40
1292	x	x	carbontetrachloride	carbontetrachloride	56-23-5	56-23-5	-4.40	-0.34	-4.06
1293	x	x	n-octane	1-propanol	111-65-9	71-23-8	-4.39	-0.17	-4.22
1294	x	x	methyl_ethyl_ketone	acetophenone	78-93-3	98-86-2	-4.39	-3.22	-1.17
1295	x	x	2-propanol	acetonitrile	67-63-0	75-05-8	-4.39	-2.82	-1.57
1296	x	x	cyclopentanone	n-hexadecane	120-92-3	544-76-3	-4.39	-1.48	-2.91
1297	x	x	methylpropanoate	tetrachloroethene	554-12-1	127-18-4	-4.39	-1.45	-2.94
1298	x	x	propanoicacid	chlorobenzene	79-09-4	108-90-7	-4.38	-3.11	-1.27
1299	x	x	1-pentanol	n-hexane	71-41-0	110-54-3	-4.38	-1.01	-3.37
1300	x	x	n-octane	diphenylether	111-65-9	101-84-8	-4.38	-0.11	-4.27
1301	x	x	methyl_ethyl_ketone	bromobenzene	78-93-3	108-86-1	-4.37	-2.55	-1.82
1302	x	x	toluene	decalin-mixture	108-88-3	91-17-8	-4.37	-0.62	-3.75
1303	x	x	3-pentanone	n-octanol	96-22-0	111-87-5	-4.36	-2.78	-1.58
1304	x	x	carbontetrachloride	carbontetrachloride	56-23-5	56-23-5	-4.35	-0.34	-4.01
1305	x		1-chlorobutane	1-chlorobutane	109-69-3	109-69-3	-4.35	-	-
1306	x	x	toluene	1-nonanol	108-88-3	143-08-8	-4.34	-1.30	-3.04
1307	x	x	2-propen-1-ol	chloroform	107-18-6	67-66-3	-4.34	-2.44	-1.90
1308	x	x	nitromethane	ethanol	75-52-5	64-17-5	-4.34	-4.70	0.36
1309	x	x	3-3-dimethylbutanone	n-hexane	75-97-8	110-54-3	-4.34	-1.14	-3.20
1310	x	x	butylamine	dichloroethane	109-73-9	107-06-2	-4.34	-2.23	-2.11
1311	x	x	diethylether	chloroform	60-29-7	67-66-3	-4.32	-1.05	-3.27
1312	x	x	methyl_ethyl_ketone	ethanol	78-93-3	64-17-5	-4.32	-3.32	-1.00
1313	x	x	nitromethane	toluene	75-52-5	108-88-3	-4.31	-2.36	-1.95
1314	x	x	1-butanol	toluene	71-36-3	108-88-3	-4.31	-1.37	-2.94
1315	x	x	ethylethanoate	ethylbenzene	141-78-6	100-41-4	-4.31	-1.56	-2.75
1316	x	x	1-butanol	chlorobenzene	71-36-3	108-90-7	-4.31	-2.25	-2.06
1317	x	x	ethanol	tetrahydrothiophene-s,s-dioxide	64-17-5	126-33-0	-4.30	-2.94	-1.36
1318	x	x	pyridine	cyclohexane	110-86-1	110-82-7	-4.30	-1.14	-3.16
1319	x	x	3-pentanone	cyclohexane	96-22-0	110-82-7	-4.30	-1.19	-3.11

TABLE B.45: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1320	x	x	nitromethane	perfluorobenzene	75-52-5	392-56-3	-4.30	-2.01	-2.29
1321	x	x	methylpropanoate	ethylbenzene	554-12-1	100-41-4	-4.29	-1.55	-2.74
1322	x	x	trichloroethene	cyclohexane	79-01-6	110-82-7	-4.29	-0.51	-3.78
1323	x	x	pyridine	heptane	110-86-1	142-82-5	-4.28	-1.06	-3.22
1324	x	x	2-propanol	chloroform	67-63-0	67-66-3	-4.28	-2.02	-2.26
1325	x	x	pentylamine	n-hexadecane	110-58-7	544-76-3	-4.28	-0.98	-3.30
1326	x		methyl_ethyl_ketone	ethoxybenzene	78-93-3	103-73-1	-4.28	-	-
1327	x	x	ethylamine	1-pentanol	75-04-7	71-41-0	-4.27	-2.36	-1.91
1328	x	x	methyl_ethyl_ketone	toluene	78-93-3	108-88-3	-4.27	-1.57	-2.70
1329	x	x	ethylethanoate	2-methyl-1-propanol	141-78-6	78-83-1	-4.27	-3.00	-1.27
1330	x	x	butylamine	tetrahydrothiophene-s,s-dioxide	109-73-9	126-33-0	-4.25	-2.57	-1.68
1331	x	x	fluorobenzene	chloroform	462-06-6	67-66-3	-4.25	-1.18	-3.07
1332	x	x	ethylethanoate	1-chlorohexane	141-78-6	544-10-5	-4.25	-2.50	-1.75
1333	x	x	benzene	acetonitrile	71-43-2	75-05-8	-4.25	-1.55	-2.70
1334	x	x	nitromethane	bromobenzene	75-52-5	108-86-1	-4.25	-3.70	-0.55
1335	x	x	butanonitrile	n-octanol	109-74-0	111-87-5	-4.25	-4.08	-0.17
1336	x	x	tetrahydrofuran	tetrahydrofuran	109-99-9	109-99-9	-4.25	-1.75	-2.50
1337	x	x	butylamine	diethylether	109-73-9	60-29-7	-4.24	-1.71	-2.53
1338	x	x	1-pentanol	n-hexadecane	71-41-0	544-76-3	-4.24	-1.12	-3.12
1339	x	x	2,2,4-trimethylpentane	n-hexadecane	540-84-1	544-76-3	-4.24	-0.09	-4.15
1340	x	x	pentanal	2,2,4-trimethylpentane	110-62-3	540-84-1	-4.24	-1.05	-3.19
1341	x	x	ethanol	ethylethanoate	64-17-5	141-78-6	-4.24	-2.28	-1.96
1342	x	x	methylethanoate	perfluorobenzene	79-20-9	392-56-3	-4.23	-1.39	-2.84
1343	x	x	propanoicacid	isopropylbenzene	79-09-4	98-82-8	-4.23	-1.95	-2.28
1344	x	x	butylamine	bromobenzene	109-73-9	108-86-1	-4.22	-1.89	-2.33
1345	x	x	butylamine	p-isopropyltoluene	109-73-9	99-87-6	-4.22	-1.08	-3.14
1346	x	x	ethylethanoate	tert-butylbenzene	141-78-6	98-06-6	-4.22	-1.51	-2.71
1347	x	x	methyl_ethyl_ketone	iodobenzene	78-93-3	591-50-4	-4.22	-2.38	-1.84
1348	x	x	ethylethanoate	isopropylbenzene	141-78-6	98-82-8	-4.22	-1.53	-2.69
1349	x	x	ethylethanoate	tetrachloroethene	141-78-6	127-18-4	-4.22	-1.46	-2.76

TABLE B.46: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1350	x	x	tetrahydropyran	n-octanol	142-68-7	111-87-5	-4.21	-1.56	-2.65
1351	x	x	benzene	diethylether	71-43-2	60-29-7	-4.21	-1.02	-3.19
1352	x	x	propylethanoate	n-pentane	109-60-4	109-66-0	-4.21	-1.13	-3.08
1353	x	x	methylpropanoate	1-chlorohexane	554-12-1	544-10-5	-4.20	-2.50	-1.70
1354	x	x	methylpropanoate	xylene-mixture	554-12-1	1330-20-7	-4.20	-1.52	-2.68
1355	x	x	nitromethane	xylene-mixture	75-52-5	1330-20-7	-4.20	-2.37	-1.83
1356	x	x	acetone	acetone	67-64-1	67-64-1	-4.20	-3.47	-0.73
1357	x	x	butylamine	iodobenzene	109-73-9	591-50-4	-4.19	-1.77	-2.42
1358	x	x	methylpropanoate	n-butylbenzene	554-12-1	104-51-8	-4.19	-1.51	-2.68
1359	x	x	2-pentanone	cyclohexane	107-87-9	110-82-7	-4.19	-1.31	-2.88
1360	x	x	methylpropanoate	isopropylbenzene	554-12-1	98-82-8	-4.19	-1.51	-2.68
1361	x	x	n-octane	n-octanol	111-65-9	111-87-5	-4.18	-0.15	-4.03
1362	x	x	1-butanol	xylene-mixture	71-36-3	1330-20-7	-4.17	-1.38	-2.79
1363	x	x	ethanol	nitromethane	64-17-5	75-52-5	-4.16	-2.92	-1.24
1364	x	x	ethylethanoate	1-fluorooctane	141-78-6	463-11-6	-4.16	-2.12	-2.04
1365	x	x	methyl_ethyl_ketone	1-propanol	78-93-3	71-23-8	-4.15	-3.27	-0.88
1366	x	x	3-3-dimethylbutanone	n-decane	75-97-8	124-18-5	-4.15	-1.22	-2.93
1367	x	x	fluorobenzene	n-hexane	462-06-6	110-54-3	-4.15	-0.59	-3.56
1368	x	x	methylamine	4-methyl-2-pentanone	74-89-5	108-10-1	-4.14	-2.43	-1.71
1369	x	x	methylpropanoate	p-isopropyltoluene	554-12-1	99-87-6	-4.14	-1.43	-2.71
1370	x		methylpropanoate	1-2-3-trimethylbenzene	554-12-1	526-73-8	-4.14	-	-
1371	x		perfluoroheptane	perfluoroheptane	335-57-9	335-57-9	-4.14	-	-
1372	x	x	fluorobenzene	heptane	462-06-6	142-82-5	-4.13	-0.60	-3.53
1373	x	x	propanal	n-octanol	123-38-6	111-87-5	-4.13	-2.77	-1.36
1374	x	x	butylamine	o-dichlorobenzene	109-73-9	95-50-1	-4.13	-2.23	-1.90
1375	x	x	butylamine	perfluorobenzene	109-73-9	392-56-3	-4.13	-0.97	-3.16
1376	x	x	diethylamine	carbontetrachloride	109-89-7	56-23-5	-4.12	-0.68	-3.44
1377	x	x	ethanol	acetophenone	64-17-5	98-86-2	-4.12	-2.77	-1.35
1378	x	x	methyl_ethyl_ketone	ethylbenzene	78-93-3	100-41-4	-4.12	-1.61	-2.51
1379	x	x	methyl_ethyl_ketone	1-butanol	78-93-3	71-36-3	-4.12	-3.22	-0.90

TABLE B.47: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1380	x	x	ethylethanoate	sec-butylbenzene	141-78-6	135-98-8	-4.11	-1.51	-2.60
1381	x	x	nitromethane	iodobenzene	75-52-5	591-50-4	-4.10	-3.49	-0.61
1382	x	x	propylethanoate	n-hexane	109-60-4	110-54-3	-4.10	-1.17	-2.93
1383	x	x	methyl_ethyl_ketone	1-chlorohexane	78-93-3	544-10-5	-4.10	-2.63	-1.47
1384	x	x	dichloromethane	dimethylsulfoxide	75-09-2	67-68-5	-4.10	-2.54	-1.56
1385	x	x	1-pentanol	n-octane	71-41-0	111-65-9	-4.10	-1.05	-3.05
1386	x	x	1-pentanol	heptane	71-41-0	142-82-5	-4.09	-1.03	-3.06
1387	x	x	1-pentanol	n-dodecane	71-41-0	112-40-3	-4.09	-1.10	-2.99
1388	x	x	methyl_ethyl_ketone	tetrahydrothiophene-s,s-dioxide	78-93-3	126-33-0	-4.09	-3.43	-0.66
1389	x	x	methyl_ethyl_ketone	diethylether	78-93-3	60-29-7	-4.09	-2.31	-1.78
1390	x	x	methylpropanoate	1-fluorooctane	554-12-1	463-11-6	-4.09	-2.12	-1.97
1391	x	x	ethylamine	n-octanol	75-04-7	111-87-5	-4.09	-2.21	-1.88
1392	x	x	thiophene	heptane	110-02-1	142-82-5	-4.09	-0.64	-3.45
1393	x	x	propylethanoate	heptane	109-60-4	142-82-5	-4.09	-1.20	-2.89
1394	x	x	1,1,1-trichloroethane	cyclohexane	71-55-6	110-82-7	-4.08	-0.71	-3.37
1395	x	x	trichloroethene	n-hexadecane	79-01-6	544-76-3	-4.08	-0.52	-3.56
1396	x	x	methyl_ethyl_ketone	diphenylether	78-93-3	101-84-8	-4.08	-2.17	-1.91
1397	x	x	aceticacid	xylene-mixture	64-19-7	1330-20-7	-4.08	-2.03	-2.05
1398	x	x	propylethanoate	n-nonane	109-60-4	111-84-2	-4.07	-1.24	-2.83
1399	x	x	2-pentanone	heptane	107-87-9	142-82-5	-4.07	-1.22	-2.85
1400	x	x	methylpropanoate	n-octanol	554-12-1	111-87-5	-4.06	-2.81	-1.25
1401	x	x	propanoicacid	heptane	79-09-4	142-82-5	-4.06	-1.54	-2.52
1402	x	x	1-nitropropane	cyclohexane	108-03-2	110-82-7	-4.06	-1.81	-2.25
1403	x	x	butylamine	isopropylbenzene	109-73-9	98-82-8	-4.06	-1.15	-2.91
1404	x	x	ethylethanoate	n-octanol	141-78-6	111-87-5	-4.06	-2.80	-1.26
1405	x	x	1-butanol	iodobenzene	71-36-3	591-50-4	-4.05	-2.07	-1.98
1406	x	x	ethanol	benzonitrile	64-17-5	100-47-0	-4.05	-2.86	-1.19
1407	x	x	propylamine	dichloroethane	107-10-8	107-06-2	-4.04	-2.22	-1.82
1408	x	x	nitromethane	1-propanol	75-52-5	71-23-8	-4.04	-4.64	0.60
1409	x	x	methylethanoate	benzene	79-20-9	71-43-2	-4.04	-1.57	-2.47

TABLE B.48: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1410	x	x	1-propanol	bromoform	71-23-8	75-25-2	-4.03	-2.00	-2.03
1411	x	x	fluorobenzene	n-hexadecane	462-06-6	544-76-3	-4.03	-0.65	-3.38
1412	x		perfluoromethylcyclohexane	perfluoromethylcyclohexane	355-02-2	355-02-2	-4.03	-	-
1413	x	x	diisopropylether	diisopropylether	108-20-3	108-20-3	-4.02	-0.91	-3.12
1414	x	x	aceticacid	benzene	64-19-7	71-43-2	-4.02	-1.94	-2.08
1415	x	x	ethylamine	1-nonanol	75-04-7	143-08-8	-4.02	-2.14	-1.88
1416	x	x	diethylamine	benzene	109-89-7	71-43-2	-4.02	-0.70	-3.32
1417	x	x	propylethanoate	n-decane	109-60-4	124-18-5	-4.02	-1.26	-2.76
1418	x	x	methyl_ethyl_ketone	isopropylbenzene	78-93-3	98-82-8	-4.02	-1.57	-2.45
1419	x	x	diisopropylether	n-hexadecane	108-20-3	544-76-3	-4.02	-0.57	-3.45
1420	x	x	ethanol	triethylamine	64-17-5	121-44-8	-4.02	-1.38	-2.64
1421	x	x	thiophene	n-hexadecane	110-02-1	544-76-3	-4.01	-0.70	-3.31
1422	x	x	methylbutanoate	n-hexadecane	623-42-7	544-76-3	-4.01	-1.30	-2.71
1423	x	x	benzene	heptane	71-43-2	142-82-5	-4.00	-0.52	-3.48
1424	x	x	aceticacid	toluene	64-19-7	108-88-3	-4.00	-2.02	-1.98
1425	x	x	methylethanoate	chlorobenzene	79-20-9	108-90-7	-4.00	-2.63	-1.37
1426	x	x	nitromethane	2-propanol	75-52-5	67-63-0	-4.00	-4.61	0.61
1427	x	x	carbondisulfide	carbondisulfide	75-15-0	75-15-0	-4.00	-0.12	-3.88
1428	x	x	2-pentanone	tetralin	107-87-9	119-64-2	-3.99	-1.79	-2.20
1429	x	x	propylamine	tributylphosphate	107-10-8	126-73-8	-3.98	-2.13	-1.85
1430	x	x	ethanol	nitroethane	64-17-5	79-24-3	-3.98	-2.88	-1.10
1431	x	x	diisopropylether	diisopropylether	108-20-3	108-20-3	-3.97	-0.91	-3.06
1432	x	x	2-pentanone	n-octane	107-87-9	111-65-9	-3.97	-1.25	-2.72
1433	x	x	2-pentanone	n-nonane	107-87-9	111-84-2	-3.97	-1.26	-2.71
1434	x	x	ethylethanoate	1-bromooctane	141-78-6	111-83-1	-3.97	-2.36	-1.61
1435	x	x	propanal	propanal	123-38-6	123-38-6	-3.96	-2.99	-0.97
1436	x	x	benzene	n-hexane	71-43-2	110-54-3	-3.96	-0.51	-3.45
1437	x	x	benzene	dimethylsulfoxide	71-43-2	67-68-5	-3.96	-1.58	-2.38
1438	x	x	1-nitropropane	n-hexadecane	108-03-2	544-76-3	-3.95	-1.84	-2.11
1439	x	x	1-nitropropane	n-octane	108-03-2	111-65-9	-3.95	-1.73	-2.22

TABLE B.49: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1440	x	x	carbondisulfide	carbondisulfide	75-15-0	75-15-0	-3.95	-0.12	-3.83
1441	x	x	methyl_ethyl_ketone	mesitylene	78-93-3	108-67-8	-3.95	-1.50	-2.45
1442	x	x	methylamine	1-pentanol	74-89-5	71-41-0	-3.95	-2.48	-1.47
1443	x	x	1-nitropropane	2,2,4-trimethylpentane	108-03-2	540-84-1	-3.94	-1.72	-2.22
1444	x	x	ethanol	chloroform	64-17-5	67-66-3	-3.94	-2.10	-1.84
1445	x	x	methyl_ethyl_ketone	tert-butylbenzene	78-93-3	98-06-6	-3.94	-1.55	-2.39
1446	x	x	3-3-dimethylbutanone	n-hexadecane	75-97-8	544-76-3	-3.94	-1.27	-2.67
1447	x	x	2-pentanone	n-decane	107-87-9	124-18-5	-3.93	-1.28	-2.65
1448	x	x	diethylamine	xylene-mixture	109-89-7	1330-20-7	-3.93	-0.73	-3.20
1449	x	x	propylethanoate	n-hexadecane	109-60-4	544-76-3	-3.93	-1.30	-2.63
1450	x	x	nitromethane	1-butanol	75-52-5	71-36-3	-3.93	-4.57	0.64
1451	x	x	tetrahydrofuran	n-octanol	109-99-9	111-87-5	-3.93	-1.85	-2.08
1452	x	x	diethylamine	diethylamine	109-89-7	109-89-7	-3.93	-1.01	-2.92
1453	x	x	1-pentanol	n-nonane	71-41-0	111-84-2	-3.92	-1.07	-2.85
1454	x	x	methylethanoate	sec-butylbenzene	79-20-9	135-98-8	-3.91	-1.62	-2.29
1455	x	x	ethylamine	1-decanol	75-04-7	112-30-1	-3.91	-2.08	-1.83
1456	x	x	trimethylamine	chloroform	75-50-3	67-66-3	-3.90	-0.86	-3.04
1457	x	x	trimethylamine	2-methyl-1-propanol	75-50-3	78-83-1	-3.90	-1.16	-2.74
1458	x	x	nitromethane	diisopropylether	75-52-5	108-20-3	-3.90	-3.04	-0.86
1459	x	x	ethanol	diisopropylether	64-17-5	108-20-3	-3.90	-1.78	-2.12
1460	x		tetrahydropyran	tetrahydropyran	142-68-7	142-68-7	-3.90	-	-
1461	x	x	thiophene	n-octanol	110-02-1	111-87-5	-3.89	-1.60	-2.29
1462	x	x	pentanal	n-hexadecane	110-62-3	544-76-3	-3.89	-1.13	-2.76
1463	x	x	e-1,2-dichloroethene	e-1,2-dichloroethene	156-60-5	156-60-5	-3.88	-0.68	-3.20
1464	x	x	methylhydrazine	n-octanol	60-34-4	111-87-5	-3.88	-3.47	-0.41
1465	x	x	trichloroethene	n-undecane	79-01-6	1120-21-4	-3.87	-0.51	-3.36
1466	x	x	2,4-dimethylpentane	n-hexadecane	108-08-7	544-76-3	-3.87	-0.08	-3.79
1467	x	x	methanol	n-octanol	67-56-1	111-87-5	-3.87	-2.62	-1.25
1468	x	x	1-propanol	benzene	71-23-8	71-43-2	-3.87	-1.31	-2.56
1469	x	x	methylethanoate	bromobenzene	79-20-9	108-86-1	-3.87	-2.58	-1.29

TABLE B.50: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1470	x	x	methyl_ethyl_ketone	triethylamine	78-93-3	121-44-8	-3.86	-1.57	-2.29
1471	x	x	acetone	chlorobenzene	67-64-1	108-90-7	-3.86	-2.78	-1.08
1472	x	x	1-propanol	dichloroethane	71-23-8	107-06-2	-3.85	-2.57	-1.28
1473	x	x	methyl_ethyl_ketone	carbondisulfide	78-93-3	75-15-0	-3.85	-1.71	-2.14
1474	x	x	dichloromethane	dichloromethane	75-09-2	75-09-2	-3.85	-2.18	-1.67
1475	x	x	trichloroethene	n-decane	79-01-6	124-18-5	-3.84	-0.50	-3.34
1476	x	x	diethylamine	diethylether	109-89-7	60-29-7	-3.83	-1.11	-2.72
1477	x	x	3-pentanone	n-hexadecane	96-22-0	544-76-3	-3.83	-1.21	-2.62
1478	x	x	1,1,1-trichloroethane	n-undecane	71-55-6	1120-21-4	-3.82	-0.70	-3.12
1479	x	x	ethanol	dichloromethane	64-17-5	75-09-2	-3.82	-2.52	-1.30
1480	x	x	benzene	1-nonanol	71-43-2	143-08-8	-3.82	-1.31	-2.51
1481	x	x	methylethanoate	carbontetrachloride	79-20-9	56-23-5	-3.82	-1.54	-2.28
1482	x	x	n,n-dimethylformamide	cyclohexane	68-12-2	110-82-7	-3.82	-1.95	-1.87
1483	x	x	acetone	perfluorobenzene	67-64-1	392-56-3	-3.82	-1.45	-2.37
1484	x	x	pyridine	n-hexane	110-86-1	110-54-3	-3.81	-1.04	-2.77
1485	x	x	methylethanoate	toluene	79-20-9	108-88-3	-3.81	-1.64	-2.17
1486	x	x	dichloromethane	dichloromethane	75-09-2	75-09-2	-3.80	-2.18	-1.62
1487	x	x	benzene	n-hexadecane	71-43-2	544-76-3	-3.80	-0.57	-3.23
1488	x	x	benzene	n-decane	71-43-2	124-18-5	-3.80	-0.55	-3.25
1489	x	x	diethylamine	dibutylether	109-89-7	142-96-1	-3.80	-0.91	-2.89
1490	x	x	acetone	benzene	67-64-1	71-43-2	-3.79	-1.64	-2.15
1491	x	x	methyl_ethyl_ketone	n-octanol	78-93-3	111-87-5	-3.78	-2.98	-0.80
1492	x	x	propanoicacid	cyclohexane	79-09-4	110-82-7	-3.78	-1.65	-2.13
1493	x	x	diisopropylether	chloroform	108-20-3	67-66-3	-3.78	-1.09	-2.69
1494	x	x	diethylamine	diisopropylether	109-89-7	108-20-3	-3.78	-0.97	-2.81
1495	x	x	methylamine	n-octanol	74-89-5	111-87-5	-3.78	-2.33	-1.45
1496	x	x	pyrrole	cyclohexane	109-97-7	110-82-7	-3.77	-1.40	-2.37
1497	x	x	1-butanol	n-pentane	71-36-3	109-66-0	-3.77	-1.01	-2.76
1498	x	x	1-butanol	n-decane	71-36-3	124-18-5	-3.77	-1.13	-2.64
1499	x	x	1-butanol	n-hexane	71-36-3	110-54-3	-3.77	-1.05	-2.72

TABLE B.51: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1500	x	x	1-butanol	n-nonane	71-36-3	111-84-2	-3.77	-1.11	-2.66
1501	x	x	acetone	dimethylsulfoxide	67-64-1	67-68-5	-3.76	-3.63	-0.13
1502	x	x	diethylamine	toluene	109-89-7	108-88-3	-3.75	-0.73	-3.02
1503	x	x	heptane	n-octanol	142-82-5	111-87-5	-3.74	-0.15	-3.59
1504	x	x	1,1,1-trichloroethane	n-hexadecane	71-55-6	544-76-3	-3.73	-0.72	-3.01
1505	x	x	benzene	n-octanol	71-43-2	111-87-5	-3.72	-1.34	-2.38
1506	x	x	1-propanol	toluene	71-23-8	108-88-3	-3.71	-1.37	-2.34
1507	x	x	1-propanol	ethylbenzene	71-23-8	100-41-4	-3.71	-1.40	-2.31
1508	x	x	dimethyldisulfide	dimethyldisulfide	75-18-3	75-18-3	-3.70	-2.05	-1.66
1509	x	x	propylamine	xylene-mixture	107-10-8	1330-20-7	-3.69	-1.16	-2.53
1510	x	x	methylpropanoate	n-pentane	554-12-1	109-66-0	-3.69	-1.13	-2.56
1511	x	x	dimethylamine	chloroform	124-40-3	67-66-3	-3.69	-1.33	-2.36
1512	x	x	ethylethanoate	n-pentane	141-78-6	109-66-0	-3.69	-1.14	-2.55
1513	x	x	1-butanol	n-octane	71-36-3	111-65-9	-3.69	-1.09	-2.60
1514	x	x	benzene	1-hexanol	71-43-2	111-27-3	-3.68	-1.40	-2.28
1515	x	x	propylamine	benzene	107-10-8	71-43-2	-3.68	-1.10	-2.58
1516	x	x	nitromethane	dibutylether	75-52-5	142-96-1	-3.67	-2.86	-0.81
1517	x	x	methylethanoate	carbondisulfide	79-20-9	75-15-0	-3.67	-1.78	-1.89
1518	x	x	1-butanol	heptane	71-36-3	142-82-5	-3.66	-1.07	-2.59
1519	x	x	methylethanoate	1-chlorohexane	79-20-9	544-10-5	-3.66	-2.66	-1.00
1520	x	x	1-propanethiol	n-hexadecane	107-03-9	544-76-3	-3.66	-0.91	-2.75
1521	x	x	propanonitrile	n-octanol	107-12-0	111-87-5	-3.66	-4.10	0.44
1522	x	x	methylpropanoate	n-hexane	554-12-1	110-54-3	-3.65	-1.16	-2.49
1523	x	x	acetaldehyde	chloroform	75-07-0	67-66-3	-3.65	-2.35	-1.30
1524	x	x	propylamine	diethylether	107-10-8	60-29-7	-3.65	-1.71	-1.94
1525	x	x	1-propanol	carbontetrachloride	71-23-8	56-23-5	-3.64	-1.29	-2.35
1526	x	x	tetrahydrofuran	dimethylsulfoxide	109-99-9	67-68-5	-3.64	-2.13	-1.51
1527	x	x	aceticacid	carbontetrachloride	64-19-7	56-23-5	-3.64	-1.90	-1.74
1528	x	x	nitromethane	triethylamine	75-52-5	121-44-8	-3.63	-2.37	-1.26
1529	x	x	butylamine	n-hexane	109-73-9	110-54-3	-3.62	-0.87	-2.75

TABLE B.52: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1530	x	x	ethylethanoate	n-hexane	141-78-6	110-54-3	-3.62	-1.18	-2.44
1531	x	x	n-hexane	benzene	110-54-3	71-43-2	-3.62	-0.07	-3.55
1532	x	x	methanol	diethylether	67-56-1	60-29-7	-3.61	-2.06	-1.55
1533	x	x	diethylamine	cyclohexane	109-89-7	110-82-7	-3.61	-0.60	-3.01
1534	x	x	trimethylamine	n-octanol	75-50-3	111-87-5	-3.60	-1.07	-2.53
1535	x	x	z-1,2-dichloroethene	n-undecane	156-59-2	1120-21-4	-3.60	-0.82	-2.78
1536	x	x	acetonitrile	diethylether	75-05-8	60-29-7	-3.59	-3.43	-0.16
1537	x	x	propylamine	chlorobenzene	107-10-8	108-90-7	-3.59	-1.92	-1.67
1538	x	x	fluorobenzene	cyclohexane	462-06-6	110-82-7	-3.59	-0.64	-2.95
1539	x	x	methylethanoate	1-fluorooctane	79-20-9	463-11-6	-3.59	-2.27	-1.32
1540	x	x	ethanol	anisole	64-17-5	100-66-3	-3.59	-2.00	-1.59
1541	x		furan	furan	110-00-9	110-00-9	-3.59	-	-
1542	x		methylethanoate	1-2-3-trimethylbenzene	79-20-9	526-73-8	-3.58	-	-
1543	x	x	methylethanoate	tert-butylbenzene	79-20-9	98-06-6	-3.57	-1.62	-1.95
1544	x	x	1-propanol	xylene-mixture	71-23-8	1330-20-7	-3.57	-1.38	-2.19
1545	x	x	methylpropanoate	n-octane	554-12-1	111-65-9	-3.57	-1.21	-2.36
1546	x	x	butylamine	2,2,4-trimethylpentane	109-73-9	540-84-1	-3.57	-0.91	-2.66
1547	x	x	1-butanol	2,2,4-trimethylpentane	71-36-3	540-84-1	-3.56	-1.09	-2.47
1548	x	x	ethylethanoate	cyclohexane	141-78-6	110-82-7	-3.56	-1.28	-2.28
1549	x	x	butylamine	n-decane	109-73-9	124-18-5	-3.55	-0.94	-2.61
1550	x	x	butylamine	heptane	109-73-9	142-82-5	-3.55	-0.89	-2.66
1551	x	x	butylamine	n-nonane	109-73-9	111-84-2	-3.55	-0.93	-2.62
1552	x	x	butylamine	n-undecane	109-73-9	1120-21-4	-3.55	-0.94	-2.61
1553	x	x	methylethanoate	n-octanol	79-20-9	111-87-5	-3.54	-2.98	-0.56
1554	x	x	1-butanol	cyclohexane	71-36-3	110-82-7	-3.52	-1.15	-2.37
1555	x	x	1-hexene	n-hexadecane	592-41-6	544-76-3	-3.51	-0.28	-3.23
1556	x	x	ethanol	dibutylether	64-17-5	142-96-1	-3.51	-1.67	-1.84
1557	x	x	propylamine	toluene	107-10-8	108-88-3	-3.51	-1.15	-2.36
1558	x	x	ethylethanoate	heptane	141-78-6	142-82-5	-3.50	-1.20	-2.30
1559	x	x	methylpropanoate	decalin-mixture	554-12-1	91-17-8	-3.50	-1.40	-2.10

TABLE B.53: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1560	x	x	methylpropanoate	n-nonane	554-12-1	111-84-2	-3.50	-1.23	-2.27
1561	x	x	tert_butyl_methyl_ether	n-octanol	1634-04-4	111-87-5	-3.49	-1.45	-2.04
1562	x	x	formaldehyde	1-butanol	50-00-0	71-36-3	-3.49	-2.60	-0.89
1563	x	x	2-methylpentane	n-hexadecane	107-83-5	544-76-3	-3.48	-0.07	-3.41
1564	x	x	butanonitrile	n-hexadecane	109-74-0	544-76-3	-3.48	-1.94	-1.54
1565	x	x	2-propanol	benzene	67-63-0	71-43-2	-3.48	-1.25	-2.23
1566	x	x	methyl_ethyl_ketone	cyclohexane	78-93-3	110-82-7	-3.48	-1.31	-2.17
1567	x	x	formaldehyde	benzylalcohol	50-00-0	100-51-6	-3.48	-2.50	-0.98
1568	x	x	fluorobenzene	n-decane	462-06-6	124-18-5	-3.48	-0.63	-2.85
1569	x	x	methyl_ethyl_ketone	n-hexane	78-93-3	110-54-3	-3.48	-1.20	-2.28
1570	x	x	1-propanol	o-dichlorobenzene	71-23-8	95-50-1	-3.47	-2.56	-0.91
1571	x	x	2-nitropropane	n-hexadecane	79-46-9	544-76-3	-3.47	-1.74	-1.73
1572	x	x	1-butanol	n-dodecane	71-36-3	112-40-3	-3.47	-1.14	-2.33
1573	x	x	cyclohexane	n-octanol	110-82-7	111-87-5	-3.46	-0.09	-3.37
1574	x	x	2-pentene	carbontetrachloride	627-20-3	56-23-5	-3.46	-0.27	-3.19
1575	x	x	ethanol	fluorobenzene	64-17-5	462-06-6	-3.45	-2.21	-1.24
1576	x	x	acetone	1-chlorohexane	67-64-1	544-10-5	-3.45	-2.82	-0.63
1577	x	x	ethylethanoate	n-nonane	141-78-6	111-84-2	-3.45	-1.24	-2.21
1578	x		ethanol	ethoxybenzene	64-17-5	103-73-1	-3.45	-	-
1579	x	x	butanal	2,2,4-trimethylpentane	123-72-8	540-84-1	-3.45	-1.14	-2.31
1580	x	x	propylamine	o-dichlorobenzene	107-10-8	95-50-1	-3.44	-2.22	-1.22
1581	x	x	formaldehyde	1-pentanol	50-00-0	71-41-0	-3.44	-2.56	-0.88
1582	x	x	fluorobenzene	decalin-mixture	462-06-6	91-17-8	-3.44	-0.71	-2.73
1583	x	x	butylamine	n-octane	109-73-9	111-65-9	-3.44	-0.91	-2.53
1584	x	x	propylamine	ethylbenzene	107-10-8	100-41-4	-3.44	-1.18	-2.26
1585	x	x	hydrazine	n-octanol	302-01-2	111-87-5	-3.44	-3.90	0.46
1586	x	x	diethylether	diethylether	60-29-7	60-29-7	-3.44	-1.00	-2.44
1587	x	x	ethylethanoate	n-decane	141-78-6	124-18-5	-3.43	-1.26	-2.17
1588	x	x	ethanol	xylene-mixture	64-17-5	1330-20-7	-3.42	-1.38	-2.04
1589	x	x	1-hexyne	n-hexadecane	693-02-7	544-76-3	-3.42	-0.71	-2.71

TABLE B.54: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1590	x	x	ethanol	benzene	64-17-5	71-43-2	-3.42	-1.31	-2.11
1591	x	x	formaldehyde	1-hexanol	50-00-0	111-27-3	-3.42	-2.50	-0.92
1592	x	x	acetone	ethylbenzene	67-64-1	100-41-4	-3.41	-1.75	-1.66
1593	x	x	methyl_ethyl_ketone	2,2,4-trimethylpentane	78-93-3	540-84-1	-3.40	-1.24	-2.16
1594	x	x	diethylether	diethylether	60-29-7	60-29-7	-3.39	-1.00	-2.39
1595	x	x	methanol	ethylethanoate	67-56-1	141-78-6	-3.37	-2.34	-1.03
1596	x	x	ethylethanoate	n-pentadecane	141-78-6	629-62-9	-3.37	-1.30	-2.07
1597	x	x	methyl_ethyl_ketone	heptane	78-93-3	142-82-5	-3.36	-1.22	-2.14
1598	x	x	dimethylamine	xylene-mixture	124-40-3	1330-20-7	-3.36	-0.86	-2.50
1599	x	x	chloroform	1-iodohexadecane	67-66-3	544-77-4	-3.36	-1.21	-2.15
1600	x	x	acetone	carbontetrachloride	67-64-1	56-23-5	-3.35	-1.61	-1.74
1601	x	x	methylethanoate	1-bromooctane	79-20-9	111-83-1	-3.35	-2.52	-0.83
1602	x	x	methylpropanoate	n-pentadecane	554-12-1	629-62-9	-3.35	-1.29	-2.06
1603	x	x	acetamide	n-hexadecane	60-35-5	544-76-3	-3.33	-2.62	-0.71
1604	x	x	z-1,2-dichloroethene	n-hexadecane	156-59-2	544-76-3	-3.33	-0.85	-2.48
1605	x	x	ethanol	toluene	64-17-5	108-88-3	-3.33	-1.37	-1.96
1606	x	x	methylethanoate	p-isopropyltoluene	79-20-9	99-87-6	-3.32	-1.55	-1.77
1607	x	x	acetone	isopropylbenzene	67-64-1	98-82-8	-3.32	-1.71	-1.61
1608	x	x	methanol	chloroform	67-56-1	67-66-3	-3.32	-2.16	-1.16
1609	x	x	methyl_ethyl_ketone	n-decane	78-93-3	124-18-5	-3.30	-1.29	-2.01
1610	x	x	nitroethane	n-hexadecane	79-24-3	544-76-3	-3.29	-1.86	-1.43
1611	x	x	diethylamine	n-hexadecane	109-89-7	544-76-3	-3.27	-0.61	-2.66
1612	x	x	ethanol	bromobenzene	64-17-5	108-86-1	-3.26	-2.20	-1.06
1613	x	x	ethanol	bromoform	64-17-5	75-25-2	-3.24	-2.01	-1.23
1614	x	x	methyl_ethyl_ketone	n-octane	78-93-3	111-65-9	-3.24	-1.25	-1.99
1615	x	x	formaldehyde	n-octanol	50-00-0	111-87-5	-3.23	-2.41	-0.82
1616	x		2-methylbutane	2-methylbutane	78-78-4	78-78-4	-3.22	-	-
1617	x	x	methylcyclohexane	n-octanol	108-87-2	111-87-5	-3.21	-0.12	-3.09
1618	x	x	methylamine	xylene-mixture	74-89-5	1330-20-7	-3.20	-1.24	-1.96
1619	x	x	methyl_ethyl_ketone	n-nonane	78-93-3	111-84-2	-3.20	-1.27	-1.93

TABLE B.55: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1620	x	x	ethylamine	dichloroethane	75-04-7	107-06-2	-3.19	-2.22	-0.97
1621	x	x	nitroethane	n-hexane	79-24-3	110-54-3	-3.19	-1.68	-1.51
1622	x	x	ethanol	iodobenzene	64-17-5	591-50-4	-3.18	-2.07	-1.11
1623	x	x	methylamine	chloroform	74-89-5	67-66-3	-3.17	-1.90	-1.27
1624	x	x	2-propanol	carbontetrachloride	67-63-0	56-23-5	-3.15	-1.23	-1.92
1625	x	x	acetonitrile	n-octanol	75-05-8	111-87-5	-3.15	-4.30	1.15
1626	x	x	hydrogen_peroxide	carbondisulfide	7722-84-1	75-15-0	-3.14	-2.36	-0.78
1627	x	x	hydrogen_peroxide	carbontetrachloride	7722-84-1	56-23-5	-3.14	-2.04	-1.10
1628	x	x	hydrogen_peroxide	toluene	7722-84-1	108-88-3	-3.14	-2.17	-0.97
1629	x	x	acetone	carbondisulfide	67-64-1	75-15-0	-3.14	-1.86	-1.28
1630	x	x	methylethanoate	n-pentane	79-20-9	109-66-0	-3.13	-1.22	-1.91
1631	x	x	ethanethiol	2,2,4-trimethylpentane	75-08-1	540-84-1	-3.13	-0.86	-2.27
1632	x	x	propylamine	n-pentane	107-10-8	109-66-0	-3.13	-0.84	-2.29
1633	x	x	propylamine	n-hexane	107-10-8	110-54-3	-3.13	-0.87	-2.26
1634	x	x	methyl_ethyl_ketone	n-hexadecane	78-93-3	544-76-3	-3.12	-1.33	-1.79
1635	x	x	methylethanoate	n-hexane	79-20-9	110-54-3	-3.12	-1.27	-1.85
1636	x	x	methyl_ethyl_ketone	tetralin	78-93-3	119-64-2	-3.12	-1.79	-1.33
1637	x	x	1-propanethiol	cyclohexane	107-03-9	110-82-7	-3.12	-0.90	-2.22
1638	x		tetramethylsilane	tetramethylsilane	75-76-3	75-76-3	-3.11	-	-
1639	x	x	butanal	n-hexadecane	123-72-8	544-76-3	-3.10	-1.22	-1.88
1640	x	x	trimethylamine	carbontetrachloride	75-50-3	56-23-5	-3.09	-0.52	-2.57
1641	x	x	acetone	tetrachloroethene	67-64-1	127-18-4	-3.09	-1.64	-1.45
1642	x	x	n-hexane	2,2,4-trimethylpentane	110-54-3	540-84-1	-3.08	-0.06	-3.02
1643	x	x	hydrazine	diethylether	302-01-2	60-29-7	-3.08	-3.08	0.00
1644	x	x	methylethanoate	n-octane	79-20-9	111-65-9	-3.06	-1.32	-1.74
1645	x	x	methylethanoate	cyclohexane	79-20-9	110-82-7	-3.06	-1.38	-1.68
1646	x	x	dimethyldisulfide	n-hexadecane	75-18-3	544-76-3	-3.05	-0.93	-2.12
1647	x	x	methanol	butylethanoate	67-56-1	123-86-4	-3.04	-2.20	-0.84
1648	x	x	diethylether	cyclohexane	60-29-7	110-82-7	-3.03	-0.55	-2.48
1649	x	x	propylamine	heptane	107-10-8	142-82-5	-3.03	-0.89	-2.14

TABLE B.56: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1650	x	x	methylethanoate	n-nonane	79-20-9	111-84-2	-3.02	-1.34	-1.68
1651	x	x	dimethylamine	benzene	124-40-3	71-43-2	-3.01	-0.81	-2.20
1652	x	x	ethylamine	xylene-mixture	75-04-7	1330-20-7	-3.01	-1.16	-1.85
1653	x	x	n-hexane	n-octanol	110-54-3	111-87-5	-3.01	-0.14	-2.87
1654	x	x	1-propanol	heptane	71-23-8	142-82-5	-3.01	-1.07	-1.94
1655	x	x	propylamine	n-octane	107-10-8	111-65-9	-3.00	-0.91	-2.09
1656	x	x	1-propanol	2,2,4-trimethylpentane	71-23-8	540-84-1	-3.00	-1.09	-1.91
1657	x	x	methylethanoate	n-decane	79-20-9	124-18-5	-2.98	-1.36	-1.62
1658	x	x	propylamine	n-decane	107-10-8	124-18-5	-2.96	-0.94	-2.02
1659	x	x	ethanethiol	n-hexadecane	75-08-1	544-76-3	-2.96	-0.91	-2.05
1660	x	x	propylamine	n-nonane	107-10-8	111-84-2	-2.96	-0.92	-2.04
1661	x	x	ethanol	carbontetrachloride	64-17-5	56-23-5	-2.96	-1.29	-1.67
1662	x	x	propylamine	n-hexadecane	107-10-8	544-76-3	-2.92	-0.97	-1.95
1663	x	x	tetramethylsilane	n-hexadecane	75-76-3	544-76-3	-2.92	-0.14	-2.78
1664	x	x	formamide	n-hexadecane	75-12-7	544-76-3	-2.91	-2.62	-0.29
1665	x	x	ethanol	isopropylbenzene	64-17-5	98-82-8	-2.90	-1.37	-1.53
1666	x	x	methylethanoate	decalin-mixture	79-20-9	91-17-8	-2.90	-1.52	-1.38
1667	x	x	nitromethane	n-hexane	75-52-5	110-54-3	-2.90	-1.82	-1.08
1668	x	x	diethylether	n-octanol	60-29-7	111-87-5	-2.89	-1.32	-1.57
1669	x	x	ethylamine	diethylether	75-04-7	60-29-7	-2.89	-1.70	-1.19
1670	x	x	1-1-2-trichloro-1-2-2-trifluoroethane	n-hexadecane	76-13-1	544-76-3	-2.89	-0.22	-2.67
1671	x	x	3-chloro-1-propene	n-hexadecane	107-05-1	544-76-3	-2.88	-1.02	-1.86
1672	x	x	trimethylamine	4-methyl-2-pentanone	75-50-3	108-10-1	-2.86	-1.12	-1.74
1673	x	x	formaldehyde	2-butanol	50-00-0	78-92-2	-2.86	-2.58	-0.28
1674	x	x	1-chloropropane	n-hexadecane	540-54-5	544-76-3	-2.86	-0.86	-2.00
1675	x	x	nitromethane	cyclohexane	75-52-5	110-82-7	-2.86	-1.99	-0.87
1676	x	x	acetaldehyde	diethylether	75-07-0	60-29-7	-2.85	-2.25	-0.60
1677	x	x	propanonitrile	n-hexadecane	107-12-0	544-76-3	-2.84	-1.95	-0.89
1678	x	x	trimethylamine	chlorobenzene	75-50-3	108-90-7	-2.82	-0.93	-1.89
1679	x	x	methylethanoate	n-pentadecane	79-20-9	629-62-9	-2.82	-1.40	-1.42

TABLE B.57: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1680	x	x	methylmethanoate	n-octanol	107-31-3	111-87-5	-2.82	-2.87	0.05
1681	x	x	nitromethane	2,2,4-trimethylpentane	75-52-5	540-84-1	-2.82	-1.89	-0.93
1682	x	x	1-propanol	n-hexane	71-23-8	110-54-3	-2.81	-1.05	-1.76
1683	x	x	nitromethane	n-decane	75-52-5	124-18-5	-2.81	-1.95	-0.86
1684	x	x	diethylether	n-hexadecane	60-29-7	544-76-3	-2.81	-0.56	-2.25
1685	x	x	trimethylamine	benzene	75-50-3	71-43-2	-2.80	-0.53	-2.27
1686	x	x	1-pentene	n-hexadecane	109-67-1	544-76-3	-2.79	-0.27	-2.52
1687	x	x	1-propanol	n-hexadecane	71-23-8	544-76-3	-2.77	-1.16	-1.61
1688	x	x	ethylamine	carbontetrachloride	75-04-7	56-23-5	-2.77	-1.08	-1.69
1689	x	x	dichloromethane	n-hexadecane	75-09-2	544-76-3	-2.76	-1.02	-1.74
1690	x	x	dichloromethane	1-iodohexadecane	75-09-2	544-77-4	-2.76	-1.59	-1.17
1691	x	x	1-propanol	n-pentane	71-23-8	109-66-0	-2.76	-1.01	-1.75
1692	x	x	1-propanol	n-decane	71-23-8	124-18-5	-2.76	-1.13	-1.63
1693	x	x	dimethylamine	carbontetrachloride	124-40-3	56-23-5	-2.75	-0.80	-1.95
1694	x	x	dimethylamine	chlorobenzene	124-40-3	108-90-7	-2.75	-1.42	-1.33
1695	x	x	1-propanol	n-dodecane	71-23-8	112-40-3	-2.74	-1.14	-1.60
1696	x	x	1-pentyne	n-hexadecane	627-19-0	544-76-3	-2.74	-0.75	-1.99
1697	x	x	trimethylamine	diisopropylether	75-50-3	108-20-3	-2.74	-0.73	-2.01
1698	x	x	ethylamine	benzene	75-04-7	71-43-2	-2.73	-1.10	-1.63
1699	x	x	ethanol	n-hexane	64-17-5	110-54-3	-2.73	-1.05	-1.68
1700	x	x	1-propanol	cyclohexane	71-23-8	110-82-7	-2.73	-1.15	-1.58
1701	x	x	ethylamine	bromobenzene	75-04-7	108-86-1	-2.73	-1.88	-0.85
1702	x	x	ethylamine	iodobenzene	75-04-7	591-50-4	-2.73	-1.76	-0.97
1703	x	x	trimethylamine	toluene	75-50-3	108-88-3	-2.71	-0.55	-2.16
1704	x	x	2-chloropropane	n-hexadecane	75-29-6	544-76-3	-2.69	-0.87	-1.82
1705	x	x	methylpropanoate	n-hexadecane	554-12-1	544-76-3	-2.68	-1.29	-1.39
1706	x	x	dimethylamine	toluene	124-40-3	108-88-3	-2.68	-0.85	-1.83
1707	x	x	ethylamine	toluene	75-04-7	108-88-3	-2.67	-1.15	-1.52
1708	x	x	methylethanoate	n-hexadecane	79-20-9	544-76-3	-2.67	-1.40	-1.27
1709	x	x	acetone	cyclohexane	67-64-1	110-82-7	-2.67	-1.44	-1.23

TABLE B.58: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1710	x	x	cyclopentane	n-octanol	287-92-3	111-87-5	-2.65	-0.11	-2.54
1711	x	x	methylamine	toluene	74-89-5	108-88-3	-2.65	-1.23	-1.42
1712	x	x	fluorotrichloromethane	n-octanol	75-69-4	111-87-5	-2.63	-0.54	-2.09
1713	x	x	dimethylamine	diethylether	124-40-3	60-29-7	-2.63	-1.27	-1.36
1714	x	x	2-methylpropene	carbontetrachloride	115-11-7	56-23-5	-2.63	-0.31	-2.32
1715	x	x	trimethylamine	cyclohexane	75-50-3	110-82-7	-2.63	-0.46	-2.17
1716	x	x	fluorotrichloromethane	chloroform	75-69-4	67-66-3	-2.62	-0.44	-2.18
1717	x	x	ethylamine	ethylbenzene	75-04-7	100-41-4	-2.59	-1.18	-1.41
1718	x	x	ethylmethanoate	n-hexadecane	109-94-4	544-76-3	-2.59	-1.26	-1.33
1719	x	x	ethylamine	o-dichlorobenzene	75-04-7	95-50-1	-2.59	-2.21	-0.38
1720	x	x	nitromethane	n-hexadecane	75-52-5	544-76-3	-2.58	-2.02	-0.56
1721	x	x	methanol	benzene	67-56-1	71-43-2	-2.58	-1.36	-1.22
1722	x	x	chloroethane	n-octanol	75-00-3	111-87-5	-2.58	-1.87	-0.71
1723	x	x	acetone	tetralin	67-64-1	119-64-2	-2.54	-1.95	-0.59
1724	x	x	methanol	dichloroethane	67-56-1	107-06-2	-2.53	-2.64	0.11
1725	x	x	ethanol	ethylbenzene	64-17-5	100-41-4	-2.49	-1.40	-1.09
1726	x	x	propanal	n-hexadecane	123-38-6	544-76-3	-2.48	-1.29	-1.19
1727	x	x	2-2-dimethylpropane	n-hexadecane	463-82-1	544-76-3	-2.48	-0.07	-2.41
1728	x	x	1-butene	carbontetrachloride	106-98-9	56-23-5	-2.48	-0.29	-2.19
1729	x	x	2-propanol	n-hexadecane	67-63-0	544-76-3	-2.47	-1.11	-1.36
1730	x	x	acetone	n-decane	67-64-1	124-18-5	-2.47	-1.41	-1.06
1731	x	x	acetone	n-octane	67-64-1	111-65-9	-2.46	-1.37	-1.09
1732	x	x	n-pentane	n-octanol	109-66-0	111-87-5	-2.45	-0.12	-2.33
1733	x	x	ethanol	n-decane	64-17-5	124-18-5	-2.44	-1.13	-1.31
1734	x	x	ethanol	cyclohexane	64-17-5	110-82-7	-2.42	-1.15	-1.27
1735	x	x	acetonitrile	n-hexadecane	75-05-8	544-76-3	-2.37	-2.05	-0.32
1736	x	x	1-pentene	2,2,4-trimethylpentane	109-67-1	540-84-1	-2.36	-0.25	-2.11
1737	x	x	methanol	bromobenzene	67-56-1	108-86-1	-2.31	-2.26	-0.05
1738	x	x	acetone	n-hexadecane	67-64-1	544-76-3	-2.31	-1.46	-0.85
1739	x	x	dimethyl_ether	methanol	115-10-6	67-56-1	-2.31	-1.74	-0.57

TABLE B.59: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1740	x	x	ethylamine	n-hexadecane	75-04-7	544-76-3	-2.29	-0.97	-1.32
1741	x	x	chloroethane	n-hexadecane	75-00-3	544-76-3	-2.29	-0.87	-1.42
1742	x	x	1-butene	2,2,4-trimethylpentane	106-98-9	540-84-1	-2.26	-0.24	-2.02
1743	x	x	methanol	carbontetrachloride	67-56-1	56-23-5	-2.25	-1.33	-0.92
1744	x	x	trimethylamine	n-hexadecane	75-50-3	544-76-3	-2.21	-0.47	-1.74
1745	x	x	dimethylamine	n-hexadecane	124-40-3	544-76-3	-2.18	-0.72	-1.46
1746	x	x	methanol	iodobenzene	67-56-1	591-50-4	-2.18	-2.13	-0.05
1747	x	x	methanol	toluene	67-56-1	108-88-3	-2.18	-1.42	-0.76
1748	x	x	methylamine	chlorobenzene	74-89-5	108-90-7	-2.16	-2.03	-0.13
1749	x	x	ethanol	n-pentane	64-17-5	109-66-0	-2.15	-1.01	-1.14
1750	x	x	ethanol	heptane	64-17-5	142-82-5	-2.15	-1.07	-1.08
1751	x	x	ethanol	n-octane	64-17-5	111-65-9	-2.15	-1.09	-1.06
1752	x	x	ethanol	n-nonane	64-17-5	111-84-2	-2.15	-1.11	-1.04
1753	x	x	s-trans-1-3-butadiene	n-hexadecane	106-99-0	544-76-3	-2.10	-0.44	-1.66
1754	x	x	ethylamine	n-hexane	75-04-7	110-54-3	-2.09	-0.87	-1.22
1755	x	x	ethylamine	heptane	75-04-7	142-82-5	-2.09	-0.89	-1.20
1756	x	x	n-pentane	acetonitrile	109-66-0	75-05-8	-2.08	-0.14	-1.94
1757	x	x	1-butyne	n-hexadecane	107-00-6	544-76-3	-2.07	-0.75	-1.32
1758	x	x	acetonitrile	heptane	75-05-8	142-82-5	-2.06	-1.89	-0.17
1759	x	x	dimethyl_ether	n-octanol	115-10-6	111-87-5	-2.06	-1.54	-0.52
1760	x	x	ethanol	n-dodecane	64-17-5	112-40-3	-2.06	-1.14	-0.92
1761	x	x	ethylamine	n-octane	75-04-7	111-65-9	-2.04	-0.91	-1.13
1762	x	x	ethanol	n-hexadecane	64-17-5	544-76-3	-2.03	-1.16	-0.87
1763	x	x	1-butene	n-hexadecane	106-98-9	544-76-3	-2.03	-0.26	-1.77
1764	x	x	methylmethanoate	n-hexadecane	107-31-3	544-76-3	-1.99	-1.36	-0.63
1765	x	x	ethylamine	n-nonane	75-04-7	111-84-2	-1.98	-0.92	-1.06
1766	x	x	ethylamine	n-decane	75-04-7	124-18-5	-1.92	-0.94	-0.98
1767	x	x	2-methylpropane	n-hexadecane	75-28-5	544-76-3	-1.92	-0.06	-1.86
1768	x	x	1-1-difluoroethane	methanol	75-37-6	67-56-1	-1.90	-1.96	0.06
1769	x	x	acetonitrile	cyclohexane	75-05-8	110-82-7	-1.87	-2.02	0.15

TABLE B.60: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1770	x	x	n-butane	n-octanol	106-97-8	111-87-5	-1.86	-0.11	-1.75
1771	x	x	methanol	xylene-mixture	67-56-1	1330-20-7	-1.73	-1.42	-0.31
1772	x	x	water	toluene	7732-18-5	108-88-3	-1.69	-2.09	0.40
1773	x	x	acetaldehyde	n-hexadecane	75-07-0	544-76-3	-1.68	-1.33	-0.35
1774	x	x	propene	2,2,4-trimethylpentane	115-07-1	540-84-1	-1.61	-0.26	-1.35
1775	x	x	cyclopropane	n-octanol	75-19-4	111-87-5	-1.60	-0.51	-1.09
1776	x	x	water	xylene-mixture	7732-18-5	1330-20-7	-1.56	-2.10	0.54
1777	x	x	difluorodichloromethane	chloroform	75-71-8	67-66-3	-1.55	-0.34	-1.21
1778	x	x	water	ethylbenzene	7732-18-5	100-41-4	-1.51	-2.13	0.62
1779	x	x	dimethyl_ether	n-hexadecane	115-10-6	544-76-3	-1.49	-0.70	-0.79
1780	x	x	methanol	n-hexane	67-56-1	110-54-3	-1.49	-1.09	-0.40
1781	x	x	2-methylpropane	n-octanol	75-28-5	111-87-5	-1.45	-0.13	-1.32
1782	x	x	n-propane	n-hexadecane	74-98-6	544-76-3	-1.43	-0.04	-1.39
1783	x	x	water	isopropylbenzene	7732-18-5	98-82-8	-1.41	-2.08	0.67
1784	x	x	propyne	n-hexadecane	74-99-7	544-76-3	-1.40	-0.80	-0.60
1785	x	x	methanol	n-hexadecane	67-56-1	544-76-3	-1.32	-1.21	-0.11
1786	x	x	methanol	n-decane	67-56-1	124-18-5	-1.29	-1.17	-0.12
1787	x	x	methanol	heptane	67-56-1	142-82-5	-1.29	-1.11	-0.18
1788	x	x	methanol	n-octane	67-56-1	111-65-9	-1.29	-1.13	-0.16
1789	x	x	n-propane	n-octanol	74-98-6	111-87-5	-1.26	-0.09	-1.17
1790	x	x	1-1-difluoroethane	n-octanol	75-37-6	111-87-5	-1.13	-1.75	0.62
1791	x	x	formaldehyde	n-hexadecane	50-00-0	544-76-3	-0.99	-1.12	0.13
1792	x	x	fluoroethane	n-hexadecane	353-36-6	544-76-3	-0.76	-0.65	-0.11
1793	x	x	hydrogen_sulfide	n-hexadecane	7783-06-4	544-76-3	-0.72	-0.95	0.23
1794	x	x	ethane	n-octanol	74-84-0	111-87-5	-0.64	-0.08	-0.56
1795	x	x	ethene	n-hexadecane	74-85-1	544-76-3	-0.39	-0.28	-0.11
1796	x	x	water	n-hexadecane	7732-18-5	544-76-3	-0.35	-1.78	1.43
1797	x	x	ethyne	n-hexadecane	74-86-2	544-76-3	-0.20	-0.83	0.63
1798	x	x	methane	n-hexadecane	74-82-8	544-76-3	0.45	-0.04	0.49
1799	x	x	methane	n-octanol	74-82-8	111-87-5	0.51	-0.09	0.60

TABLE B.61: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1800	x	x	tetrafluoromethane	n-octanol	75-73-0	111-87-5	1.50	-0.23	1.73
1801	x	x	1,4-dioxane	2-methylpyridine	123-91-1	109-06-8	-5.01	-2.66	-2.35
1802	x	x	aceticacid	4-methyl-2-pentanone	64-19-7	108-10-1	-6.33	-3.76	-2.57
1803	x	x	diethylamine	4-methyl-2-pentanone	109-89-7	108-10-1	-3.63	-1.53	-2.10
1804	x	x	phenol	4-methyl-2-pentanone	108-95-2	108-10-1	-9.38	-3.45	-5.93
1805	x	x	1,4-dioxane	aceticacid	123-91-1	64-19-7	-5.80	-2.40	-3.40
1806	x	x	n-octane	aceticacid	111-65-9	64-19-7	-3.93	-0.14	-3.79
1807	x	x	toluene	aceticacid	108-88-3	64-19-7	-4.53	-1.19	-3.34
1808	x	x	1,4-dioxane	acetonitrile	123-91-1	75-05-8	-5.33	-3.01	-2.32
1809	x	x	n-octane	acetonitrile	111-65-9	75-05-8	-3.57	-0.17	-3.40
1810	x	x	toluene	acetonitrile	108-88-3	75-05-8	-4.68	-1.55	-3.13
1811	x	x	1,4-dioxane	acetophenone	123-91-1	98-86-2	-5.03	-2.87	-2.16
1812	x	x	hydrogen_peroxide	acetophenone	7722-84-1	98-86-2	-7.30	-4.33	-2.97
1813	x	x	n-octane	acetophenone	111-65-9	98-86-2	-4.24	-0.16	-4.08
1814	x	x	1,4-dioxane	aniline	123-91-1	62-53-3	-5.65	-2.46	-3.19
1815	x	x	hydrogen_peroxide	aniline	7722-84-1	62-53-3	-7.80	-3.72	-4.08
1816	x	x	n-octane	aniline	111-65-9	62-53-3	-3.48	-0.14	-3.34
1817	x	x	toluene	aniline	108-88-3	62-53-3	-4.57	-1.23	-3.34
1818	x	x	1,4-dioxane	anisole	123-91-1	100-66-3	-5.06	-2.08	-2.98
1819	x	x	1,4-dioxane	benzene	123-91-1	71-43-2	-5.21	-1.37	-3.84
1820	x	x	benzamide	benzene	55-21-0	71-43-2	-9.93	-2.85	-7.08
1821	x	x	cyclohexane	benzene	110-82-7	71-43-2	-4.05	-0.04	-4.01
1822	x	x	hydrazine	benzene	302-01-2	71-43-2	-4.02	-2.03	-1.99
1823	x	x	hydrogen_peroxide	benzene	7722-84-1	71-43-2	-4.77	-2.08	-2.69
1824	x	x	p_hydroxybenzaldehyde	benzene	123-08-0	71-43-2	-9.73	-2.85	-6.88
1825	x	x	pyridine	benzene	110-86-1	71-43-2	-5.28	-1.30	-3.98
1826	x	x	triethyl_phosphate	benzene	78-40-0	71-43-2	-8.58	-2.92	-5.66
1827	x	x	trimethyl_phosphate	benzene	512-56-1	71-43-2	-8.02	-3.10	-4.92
1828	x	x	1,4-dioxane	benzonitrile	123-91-1	100-47-0	-5.14	-2.96	-2.18
1829	x	x	n-octane	benzonitrile	111-65-9	100-47-0	-4.34	-0.17	-4.17

TABLE B.62: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1830	x	x	1,4-dioxane	benzylalcohol	123-91-1	100-51-6	-5.39	-2.76	-2.63
1831	x	x	aceticacid	benzylalcohol	64-19-7	100-51-6	-6.96	-3.75	-3.21
1832	x	x	n-octane	benzylalcohol	111-65-9	100-51-6	-3.77	-0.16	-3.61
1833	x	x	toluene	benzylalcohol	108-88-3	100-51-6	-4.46	-1.39	-3.07
1834	x	x	1,4-dioxane	bromobenzene	123-91-1	108-86-1	-5.02	-2.29	-2.73
1835	x	x	1,4-dioxane	1,2-dibromoethane	123-91-1	106-93-4	-5.39	-2.22	-3.17
1837	x	x	1,2-ethanediol	1-butanol	107-21-1	71-36-3	-8.69	-4.96	-3.73
1838	x	x	aceticacid	1-butanol	64-19-7	71-36-3	-6.81	-3.88	-2.93
1839	x	x	benzene	1-butanol	71-43-2	71-36-3	-3.39	-1.47	-1.92
1840	x	x	ethylbenzene	1-butanol	100-41-4	71-36-3	-4.46	-1.43	-3.03
1841	x	x	hydrogen_peroxide	1-butanol	7722-84-1	71-36-3	-7.94	-4.33	-3.61
1842	x	x	propanoicacid	1-butanol	79-09-4	71-36-3	-7.17	-3.77	-3.40
1843	x	x	toluene	1-butanol	108-88-3	71-36-3	-4.50	-1.46	-3.04
1844	x		1,4-dioxane	methyl_ethyl_ketone	123-91-1	78-93-3	-5.02	-	-
1845	x		aceticacid	methyl_ethyl_ketone	64-19-7	78-93-3	-6.88	-	-
1846	x		formaldehyde	methyl_ethyl_ketone	50-00-0	78-93-3	-1.77	-	-
1847	x	x	1,2-ethanediol	butylethanoate	107-21-1	123-86-4	-6.27	-3.84	-2.43
1848	x	x	o-cresol	butylethanoate	95-48-7	123-86-4	-8.90	-2.57	-6.33
1849	x	x	4-methylphenol	butylethanoate	106-44-5	123-86-4	-9.28	-2.70	-6.58
1850	x	x	hydrogen_peroxide	butylethanoate	7722-84-1	123-86-4	-6.76	-3.37	-3.39
1851	x	x	phenol	butylethanoate	108-95-2	123-86-4	-8.96	-2.74	-6.22
1852	x	x	water	butylethanoate	7732-18-5	123-86-4	-4.13	-3.21	-0.92
1853	x	x	1,4-dioxane	carbondisulfide	123-91-1	75-15-0	-4.67	-1.55	-3.12
1854	x	x	aceticacid	carbondisulfide	64-19-7	75-15-0	-2.98	-2.19	-0.79
1855	x	x	n-octane	carbondisulfide	111-65-9	75-15-0	-5.68	-0.09	-5.59
1856	x	x	1,4-dioxane	carbontetrachloride	123-91-1	56-23-5	-4.97	-1.34	-3.63
1857	x	x	benzamide	carbontetrachloride	55-21-0	56-23-5	-9.13	-2.79	-6.34
1858	x	x	bromobenzene	carbontetrachloride	108-86-1	56-23-5	-5.85	-0.83	-5.02
1859	x	x	t-butanol	carbontetrachloride	75-65-0	56-23-5	-3.40	-1.16	-2.24
1860	x	x	trimethyl_phosphate	carbontetrachloride	512-56-1	56-23-5	-7.24	-3.04	-4.20

TABLE B.63: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1861	x	x	water	carbontetrachloride	7732-18-5	56-23-5	-0.85	-1.96	1.11
1862	x	x	1,4-dioxane	chlorobenzene	123-91-1	108-90-7	-5.08	-2.33	-2.75
1863	x	x	1,2-ethanediol	chloroform	107-21-1	67-66-3	-5.98	-3.75	-2.23
1864	x	x	1,4-dioxane	chloroform	123-91-1	67-66-3	-6.21	-2.18	-4.03
1865	x	x	2,2,2-trifluoroethanol	chloroform	75-89-8	67-66-3	-3.03	-3.12	0.09
1866	x	x	2,6-dimethylpyridine	chloroform	108-48-5	67-66-3	-7.74	-1.78	-5.96
1867	x	x	2-ethylpyrazine	chloroform	13925-00-3	67-66-3	-7.72	-2.18	-5.54
1868	x	x	2-methylpyrazine	chloroform	109-08-0	67-66-3	-6.99	-2.31	-4.68
1869	x	x	2-methylpyridine	chloroform	109-06-8	67-66-3	-6.98	-1.93	-5.05
1870	x	x	3-methylpyridine	chloroform	108-99-6	67-66-3	-7.35	-2.17	-5.18
1871	x	x	4-methylpyridine	chloroform	108-89-4	67-66-3	-7.50	-2.23	-5.27
1872	x	x	acetamide	chloroform	60-35-5	67-66-3	-7.05	-4.65	-2.40
1873	x	x	acetonitrile	chloroform	75-05-8	67-66-3	-4.44	-3.57	-0.87
1874	x	x	benzamide	chloroform	55-21-0	67-66-3	-11.06	-4.58	-6.48
1875	x	x	bromobenzene	chloroform	108-86-1	67-66-3	-6.07	-1.37	-4.70
1876	x	x	chloroform	chloroform	67-66-3	67-66-3	-4.13	-1.39	-2.74
1877	x	x	cyclohexane	chloroform	110-82-7	67-66-3	-4.45	-0.07	-4.38
1878	x	x	diethylsulfide	chloroform	352-93-2	67-66-3	-6.40	-1.65	-4.75
1879	x	x	formaldehyde	chloroform	50-00-0	67-66-3	0.12	-1.99	2.11
1880	x	x	hydrazine	chloroform	302-01-2	67-66-3	-4.42	-3.21	-1.21
1881	x	x	hydrogen_sulfide	chloroform	7783-06-4	67-66-3	-0.51	-1.70	1.19
1882	x	x	morpholine	chloroform	110-91-8	67-66-3	-6.72	-2.26	-4.46
1883	x	x	n,n-dimethylacetamide	chloroform	127-19-5	67-66-3	-8.38	-3.47	-4.91
1884	x	x	propylethanoate	chloroform	109-60-4	67-66-3	-6.35	-2.30	-4.05
1885	x	x	p_hydroxybenzaldehyde	chloroform	123-08-0	67-66-3	-10.30	-4.55	-5.75
1886	x	x	piperidine	chloroform	110-89-4	67-66-3	-6.37	-1.15	-5.22
1887	x	x	pyridine	chloroform	110-86-1	67-66-3	-6.45	-2.10	-4.35
1888	x	x	quinoline	chloroform	91-22-5	67-66-3	-10.23	-2.33	-7.90
1889	x	x	tetrahydropyran	chloroform	142-68-7	67-66-3	-5.84	-1.27	-4.57
1890	x	x	thiophenol	chloroform	108-98-5	67-66-3	-7.61	-2.07	-5.54

TABLE B.64: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1891	x	x	triethyl_phosphate	chloroform	78-40-0	67-66-3	-10.90	-4.72	-6.18
1892	x	x	trimethyl_phosphate	chloroform	512-56-1	67-66-3	-9.74	-4.93	-4.81
1893	x	x	tripropyl_phosphate	chloroform	513-08-6	67-66-3	-11.11	-4.68	-6.43
1894	x	x	1,4-dioxane	cyclohexane	123-91-1	110-82-7	-4.17	-1.20	-2.97
1895	x	x	1-pentanol	cyclohexane	71-41-0	110-82-7	-3.61	-1.11	-2.50
1896	x	x	2,2,2-trifluoroethanol	cyclohexane	75-89-8	110-82-7	-1.53	-1.70	0.17
1897	x	x	2-propanol	cyclohexane	67-63-0	110-82-7	-2.37	-1.10	-1.27
1898	x	x	m-cresol	cyclohexane	108-39-4	110-82-7	-5.20	-1.43	-3.77
1899	x	x	4-methylphenol	cyclohexane	106-44-5	110-82-7	-5.89	-1.42	-4.47
1900	x	x	aceticacid	cyclohexane	64-19-7	110-82-7	-1.73	-1.71	-0.02
1901	x	x	benzamide	cyclohexane	55-21-0	110-82-7	-8.72	-2.49	-6.23
1902	x	x	bromobenzene	cyclohexane	108-86-1	110-82-7	-5.29	-0.74	-4.55
1903	x	x	cyclohexane	cyclohexane	110-82-7	110-82-7	-4.43	-0.04	-4.39
1904	x	x	methanol	cyclohexane	67-56-1	110-82-7	-1.29	-1.19	-0.10
1905	x	x	3-hydroxybenzaldehyde	cyclohexane	100-83-4	110-82-7	-6.88	-2.39	-4.49
1906	x	x	n-butane	cyclohexane	106-97-8	110-82-7	-2.86	-0.05	-2.81
1907	x	x	n-octane	cyclohexane	111-65-9	110-82-7	-5.63	-0.07	-5.56
1908	x	x	n-pentane	cyclohexane	109-66-0	110-82-7	-3.50	-0.05	-3.45
1909	x	x	n-propane	cyclohexane	74-98-6	110-82-7	-2.09	-0.04	-2.05
1910	x	x	p_hydroxybenzaldehyde	cyclohexane	123-08-0	110-82-7	-7.19	-2.49	-4.70
1911	x	x	t-butanol	cyclohexane	75-65-0	110-82-7	-2.93	-1.03	-1.90
1912	x	x	tetrahydropyran	cyclohexane	142-68-7	110-82-7	-4.41	-0.69	-3.72
1913	x	x	triethyl_phosphate	cyclohexane	78-40-0	110-82-7	-7.60	-2.55	-5.05
1914	x	x	water	cyclohexane	7732-18-5	110-82-7	-0.39	-1.76	1.37
1915	x	x	1,4-dioxane	cyclohexanone	123-91-1	108-94-1	-4.95	-2.83	-2.12
1916	x	x	hydrogen_peroxide	cyclohexanone	7722-84-1	108-94-1	-9.11	-4.28	-4.83
1917	x	x	m-cresol	decalin-mixture	108-39-4	91-17-8	-5.11	-1.58	-3.53
1918	x	x	4-methylphenol	decalin-mixture	106-44-5	91-17-8	-5.68	-1.57	-4.11
1919	x	x	bromobenzene	decalin-mixture	108-86-1	91-17-8	-5.25	-0.82	-4.43
1920	x	x	1,4-dioxane	n-decane	123-91-1	124-18-5	-3.97	-1.18	-2.79

TABLE B.65: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1921	x	x	acetamide	n-decane	60-35-5	124-18-5	-2.85	-2.53	-0.32
1922	x	x	bromobenzene	n-decane	108-86-1	124-18-5	-5.43	-0.73	-4.70
1924	x	x	n-octane	n-decane	111-65-9	124-18-5	-5.18	-0.06	-5.12
1925	x	x	phenol	1-decanol	108-95-2	112-30-1	-8.58	-3.12	-5.46
1926	x	x	ethanol	1,2-dibromoethane	64-17-5	106-93-4	-2.69	-2.13	-0.56
1927	x	x	methanol	1,2-dibromoethane	67-56-1	106-93-4	-2.38	-2.19	-0.19
1928	x	x	1,4-dioxane	dibutylether	123-91-1	142-96-1	-4.37	-1.74	-2.63
1929	x	x	hydrogen_peroxide	dibutylether	7722-84-1	142-96-1	-5.75	-2.64	-3.11
1930	x	x	benzamide	dichloroethane	55-21-0	107-06-2	-10.90	-5.64	-5.26
1931	x	x	ethanol	dichloroethane	64-17-5	107-06-2	-2.83	-2.57	-0.26
1932	x	x	p_hydroxybenzaldehyde	dichloroethane	123-08-0	107-06-2	-10.70	-5.59	-5.11
1933	x	x	triethyl_phosphate	dichloroethane	78-40-0	107-06-2	-9.59	-5.85	-3.74
1934	x	x	trimethyl_phosphate	dichloroethane	512-56-1	107-06-2	-8.55	-6.04	-2.51
1935	x	x	1,2-ethanediol	diethylether	107-21-1	60-29-7	-6.20	-3.59	-2.61
1936	x	x	1,4-dioxane	diethylether	123-91-1	60-29-7	-4.67	-2.09	-2.58
1937	x	x	acetamide	diethylether	60-35-5	60-29-7	-6.16	-4.45	-1.71
1938	x	x	aceticacid	diethylether	64-19-7	60-29-7	-6.26	-2.89	-3.37
1939	x	x	benzamide	diethylether	55-21-0	60-29-7	-10.60	-4.37	-6.23
1940	x	x	bromobenzene	diethylether	108-86-1	60-29-7	-5.99	-1.31	-4.68
1941	x	x	cyclopentanol	diethylether	96-41-3	60-29-7	-6.50	-1.88	-4.62
1942	x	x	formamide	diethylether	75-12-7	60-29-7	-5.97	-4.44	-1.53
1943	x	x	hydrogen_peroxide	diethylether	7722-84-1	60-29-7	-7.03	-3.16	-3.87
1944	x	x	hydrogen_sulfide	diethylether	7783-06-4	60-29-7	-0.60	-1.63	1.03
1945	x	x	3-hydroxybenzaldehyde	diethylether	100-83-4	60-29-7	-11.36	-4.20	-7.16
1946	x	x	n,n-dimethylformamide	diethylether	68-12-2	60-29-7	-5.31	-3.38	-1.93
1947	x	x	butanoicacid	diethylether	107-92-6	60-29-7	-7.32	-2.74	-4.58
1948	x	x	hexanoicacid	diethylether	142-62-1	60-29-7	-8.85	-2.60	-6.25
1949	x	x	pentanoicacid	diethylether	109-52-4	60-29-7	-7.87	-2.76	-5.11
1950	x	x	propanoicacid	diethylether	79-09-4	60-29-7	-6.75	-2.80	-3.95
1951	x	x	phenol	diethylether	108-95-2	60-29-7	-8.75	-2.56	-6.19

TABLE B.66: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1952	x	x	p_hydroxybenzaldehyde	diethylether	123-08-0	60-29-7	-12.07	-4.35	-7.72
1953	x	x	water	diethylether	7732-18-5	60-29-7	-3.85	-3.01	-0.84
1954	x	x	1,4-dioxane	diisopropylether	123-91-1	108-20-3	-4.42	-1.86	-2.56
1955	x	x	formaldehyde	diisopropylether	50-00-0	108-20-3	-1.04	-1.70	0.66
1956	x	x	hydrogen_peroxide	diisopropylether	7722-84-1	108-20-3	-6.72	-2.82	-3.90
1957	x	x	propanoicacid	diisopropylether	79-09-4	108-20-3	-6.37	-2.51	-3.86
1958	x	x	phenol	diisopropylether	108-95-2	108-20-3	-8.35	-2.26	-6.09
1959	x	x	p_hydroxybenzaldehyde	diisopropylether	123-08-0	108-20-3	-11.63	-3.87	-7.76
1960	x	x	water	diisopropylether	7732-18-5	108-20-3	-3.58	-2.69	-0.89
1962	x	x	methyl_ethyl_ketone	n,n-dimethylacetamide	78-93-3	127-19-5	-4.52	-3.41	-1.11
1963	x	x	n-octane	n,n-dimethylacetamide	111-65-9	127-19-5	-3.94	-0.17	-3.77
1964	x	x	toluene	n,n-dimethylacetamide	108-88-3	127-19-5	-4.94	-1.55	-3.39
1965	x	x	methyl_ethyl_ketone	n,n-dimethylformamide	78-93-3	68-12-2	-4.56	-3.40	-1.16
1966	x	x	n-octane	n,n-dimethylformamide	111-65-9	68-12-2	-3.77	-0.17	-3.60
1967	x	x	toluene	n,n-dimethylformamide	108-88-3	68-12-2	-4.88	-1.55	-3.33
1968	x	x	methyl_ethyl_ketone	dimethylsulfoxide	78-93-3	67-68-5	-4.23	-3.44	-0.79
1969	x	x	n-octane	dimethylsulfoxide	111-65-9	67-68-5	-2.84	-0.17	-2.67
1970	x	x	toluene	dimethylsulfoxide	108-88-3	67-68-5	-4.42	-1.57	-2.85
1972	x	x	chlorobenzene	ethanol	108-90-7	64-17-5	-3.30	-1.88	-1.42
1973	x	x	n-octane	ethanol	111-65-9	64-17-5	-4.23	-0.17	-4.06
1974	x	x	toluene	ethanol	108-88-3	64-17-5	-4.57	-1.51	-3.06
1975	x		1,4-dioxane	ethoxybenzene	123-91-1	103-73-1	-4.87	-	-
1976	x	x	1,2-ethanediol	ethylethanoate	107-21-1	141-78-6	-6.82	-4.08	-2.74
1977	x	x	1,4-dioxane	ethylethanoate	123-91-1	141-78-6	-5.03	-2.37	-2.66
1978	x	x	aceticacid	ethylethanoate	64-19-7	141-78-6	-6.46	-3.25	-3.21
1979	x	x	hydrogen_peroxide	ethylethanoate	7722-84-1	141-78-6	-7.60	-3.58	-4.02
1980	x	x	propanoicacid	ethylethanoate	79-09-4	141-78-6	-6.95	-3.15	-3.80
1981	x	x	phenol	ethylethanoate	108-95-2	141-78-6	-8.70	-2.92	-5.78
1982	x	x	water	ethylethanoate	7732-18-5	141-78-6	-4.26	-3.40	-0.86
1983	x	x	methanol	ethylbenzene	67-56-1	100-41-4	-1.43	-1.45	0.02

TABLE B.67: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
1984	x	x	1,4-dioxane	fluorobenzene	123-91-1	462-06-6	-5.18	-2.29	-2.89
1985	x	x	m-cresol	heptane	108-39-4	142-82-5	-5.01	-1.33	-3.68
1986	x	x	bromobenzene	heptane	108-86-1	142-82-5	-5.72	-0.69	-5.03
1987	x	x	heptane	heptane	142-82-5	142-82-5	-4.65	-0.06	-4.59
1988	x	x	trimethyl_phosphate	heptane	512-56-1	142-82-5	-5.59	-2.53	-3.06
1989	x	x	aceticacid	1-heptanol	64-19-7	111-70-6	-6.70	-3.70	-3.00
1990	x	x	ethylbenzene	1-heptanol	100-41-4	111-70-6	-4.58	-1.35	-3.23
1991	x	x	phenol	1-heptanol	108-95-2	111-70-6	-8.69	-3.39	-5.30
1992	x	x	toluene	1-heptanol	108-88-3	111-70-6	-4.33	-1.37	-2.96
1993	x	x	1-2-dimethoxyethane	n-hexadecane	110-71-4	544-76-3	-3.63	-1.00	-2.63
1994	x	x	1,4-dioxane	n-hexadecane	123-91-1	544-76-3	-3.82	-1.22	-2.60
1995	x	x	2,2,2-trifluoroethanol	n-hexadecane	75-89-8	544-76-3	-1.67	-1.73	0.06
1996	x	x	aceticacid	n-hexadecane	64-19-7	544-76-3	-2.39	-1.73	-0.66
1997	x	x	2-propen-1-ol	n-hexadecane	107-18-6	544-76-3	-2.73	-1.34	-1.39
1998	x	x	anthracene	n-hexadecane	120-12-7	544-76-3	-10.32	-1.01	-9.31
1999	x	x	bromobenzene	n-hexadecane	108-86-1	544-76-3	-5.51	-0.75	-4.76
2001	x	x	chloroform	n-hexadecane	67-66-3	544-76-3	-3.38	-0.76	-2.62
2002	x	x	cyclohexane	n-hexadecane	110-82-7	544-76-3	-4.04	-0.04	-4.00
2003	x	x	cyclopentane	n-hexadecane	287-92-3	544-76-3	-3.38	-0.05	-3.33
2004	x	x	cyclopropane	n-hexadecane	75-19-4	544-76-3	-1.78	-0.22	-1.56
2005	x	x	ethane	n-hexadecane	74-84-0	544-76-3	-0.67	-0.04	-0.63
2006	x	x	methylcyclohexane	n-hexadecane	108-87-2	544-76-3	-4.43	-0.05	-4.38
2007	x	x	naphthalene	n-hexadecane	91-20-3	544-76-3	-7.29	-0.79	-6.50
2008	x	x	n-butane	n-hexadecane	106-97-8	544-76-3	-2.20	-0.05	-2.15
2009	x	x	butanoicacid	n-hexadecane	107-92-6	544-76-3	-3.86	-1.62	-2.24
2010	x	x	heptane	n-hexadecane	142-82-5	544-76-3	-4.33	-0.06	-4.27
2012	x	x	n-hexane	n-hexadecane	110-54-3	544-76-3	-3.64	-0.06	-3.58
2013	x	x	n-octane	n-hexadecane	111-65-9	544-76-3	-5.02	-0.07	-4.95
2014	x	x	n-pentane	n-hexadecane	109-66-0	544-76-3	-2.95	-0.05	-2.90

TABLE B.68: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2015	x	x	propanoicacid	n-hexadecane	79-09-4	544-76-3	-3.12	-1.67	-1.45
2016	x	x	t-butanol	n-hexadecane	75-65-0	544-76-3	-2.74	-1.05	-1.69
2017	x	x	tetrachloroethene	n-hexadecane	127-18-4	544-76-3	-4.88	-0.31	-4.57
2018	x	x	tetrahydrofuran	n-hexadecane	109-99-9	544-76-3	-3.60	-0.84	-2.76
2019	x	x	tetrahydropyran	n-hexadecane	142-68-7	544-76-3	-4.08	-0.70	-3.38
2020	x	x	benzene	1-iodohexadecane	71-43-2	544-77-4	-3.71	-0.93	-2.78
2021	x	x	cyclohexane	1-iodohexadecane	110-82-7	544-77-4	-3.66	-0.06	-3.60
2022	x	x	methylcyclohexane	1-iodohexadecane	108-87-2	544-77-4	-4.07	-0.08	-3.99
2023	x	x	heptane	1-iodohexadecane	142-82-5	544-77-4	-3.90	-0.10	-3.80
2024	x	x	n-hexane	1-iodohexadecane	110-54-3	544-77-4	-3.26	-0.10	-3.16
2025	x	x	n-pentane	1-iodohexadecane	109-66-0	544-77-4	-2.59	-0.09	-2.50
2026	x	x	1,4-dioxane	n-hexane	123-91-1	110-54-3	-4.08	-1.10	-2.98
2027	x	x	aceticacid	n-hexane	64-19-7	110-54-3	-2.83	-1.56	-1.27
2028	x	x	bromobenzene	n-hexane	108-86-1	110-54-3	-5.66	-0.68	-4.98
2030	x	x	chloroform	n-hexane	67-66-3	110-54-3	-3.17	-0.69	-2.48
2031	x	x	n-hexane	n-hexane	110-54-3	110-54-3	-4.00	-0.05	-3.95
2032	x	x	n-octane	n-hexane	111-65-9	110-54-3	-5.46	-0.06	-5.40
2033	x	x	propanoicacid	n-hexane	79-09-4	110-54-3	-2.98	-1.51	-1.47
2034	x	x	p_hydroxybenzaldehyde	n-hexane	123-08-0	110-54-3	-9.18	-2.28	-6.90
2035	x	x	trimethyl_phosphate	n-hexane	512-56-1	110-54-3	-5.82	-2.48	-3.34
2036	x	x	4-methylphenol	1-hexanol	106-44-5	111-27-3	-9.21	-3.39	-5.82
2037	x	x	aceticacid	1-hexanol	64-19-7	111-27-3	-6.51	-3.75	-2.76
2038	x	x	ethylbenzene	1-hexanol	100-41-4	111-27-3	-4.54	-1.37	-3.17
2039	x	x	phenol	1-hexanol	108-95-2	111-27-3	-8.76	-3.44	-5.32
2040	x	x	toluene	1-hexanol	108-88-3	111-27-3	-4.27	-1.40	-2.87
2041	x	x	1,4-dioxane	iodobenzene	123-91-1	591-50-4	-4.94	-2.15	-2.79
2042	x	x	m-cresol	iodobenzene	108-39-4	591-50-4	-6.04	-2.62	-3.42
2043	x	x	aceticacid	2-methyl-1-propanol	64-19-7	78-83-1	-6.80	-3.87	-2.93
2044	x	x	hydrogen_peroxide	2-methyl-1-propanol	7722-84-1	78-83-1	-7.93	-4.32	-3.61
2045	x	x	methylamine	2-methyl-1-propanol	74-89-5	78-83-1	-4.56	-2.50	-2.06
2046	x	x	piperazine	2-methyl-1-propanol	110-85-0	78-83-1	-6.58	-3.27	-3.31

TABLE B.69: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2047	x	x	1,4-dioxane	2,2,4-trimethylpentane	123-91-1	540-84-1	-4.02	-1.14	-2.88
2048	x	x	benzene	2,2,4-trimethylpentane	71-43-2	540-84-1	-4.01	-0.53	-3.48
2049	x	x	chloroform	2,2,4-trimethylpentane	67-66-3	540-84-1	-3.06	-0.71	-2.35
2050	x	x	n-octane	2,2,4-trimethylpentane	111-65-9	540-84-1	-5.44	-0.06	-5.38
2051	x	x	n-pentane	2,2,4-trimethylpentane	109-66-0	540-84-1	-3.21	-0.05	-3.16
2052	x	x	methyl_ethyl_ketone	2-propanol	78-93-3	67-63-0	-4.07	-3.25	-0.82
2053	x	x	toluene	2-propanol	108-88-3	67-63-0	-4.38	-1.47	-2.91
2054	x	x	1,4-dioxane	m-cresol	123-91-1	108-39-4	-6.82	-2.76	-4.06
2055	x	x	ethanol	m-cresol	64-17-5	108-39-4	-5.58	-2.66	-2.92
2056	x	x	n-octane	m-cresol	111-65-9	108-39-4	-4.02	-0.16	-3.86
2058	x	x	phenol	mesitylene	108-95-2	108-67-8	-6.80	-1.64	-5.16
2059	x	x	methyl_ethyl_ketone	2-methoxyethanol	78-93-3	109-86-4	-4.28	-3.22	-1.06
2060	x	x	n-octane	2-methoxyethanol	111-65-9	109-86-4	-3.71	-0.16	-3.55
2061	x	x	toluene	2-methoxyethanol	108-88-3	109-86-4	-4.49	-1.46	-3.03
2062	x	x	1,4-dioxane	dichloromethane	123-91-1	75-09-2	-5.33	-2.61	-2.72
2063	x	x	methyl_ethyl_ketone	n-methylformamide-mixture	78-93-3	123-39-7	-4.34	-3.54	-0.80
2065	x	x	n-octane	n-methylformamide-mixture	111-65-9	123-39-7	-3.34	-0.18	-3.16
2066	x	x	toluene	n-methylformamide-mixture	108-88-3	123-39-7	-4.34	-1.62	-2.72
2067	x	x	m-cresol	nitrobenzene	108-39-4	98-95-3	-7.29	-3.75	-3.54
2068	x	x	aceticacid	nitrobenzene	64-19-7	98-95-3	-4.78	-4.06	-0.72
2069	x	x	methanol	nitrobenzene	67-56-1	98-95-3	-2.93	-2.97	0.04
2070	x	x	butanoicacid	nitrobenzene	107-92-6	98-95-3	-5.84	-3.90	-1.94
2071	x	x	hexanoicacid	nitrobenzene	142-62-1	98-95-3	-7.26	-3.73	-3.53
2072	x	x	pentanoicacid	nitrobenzene	109-52-4	98-95-3	-6.47	-3.93	-2.54
2073	x	x	propanoicacid	nitrobenzene	79-09-4	98-95-3	-5.38	-3.94	-1.44
2074	x	x	1,4-dioxane	nitroethane	123-91-1	79-24-3	-5.28	-2.98	-2.30
2075	x	x	n-octane	nitroethane	111-65-9	79-24-3	-3.89	-0.17	-3.72
2076	x	x	toluene	nitroethane	108-88-3	79-24-3	-4.88	-1.52	-3.36
2077	x	x	1,4-dioxane	nitromethane	123-91-1	75-52-5	-5.46	-3.02	-2.44
2078	x	x	n-octane	nitromethane	111-65-9	75-52-5	-3.15	-0.17	-2.98

TABLE B.70: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2079	x	x	toluene	nitromethane	108-88-3	75-52-5	-4.52	-1.55	-2.97
2081	x	x	phenol	1-nonanol	108-95-2	143-08-8	-8.61	-3.22	-5.39
2082	x	x	m-cresol	n-octane	108-39-4	111-65-9	-5.19	-1.36	-3.83
2083	x	x	n-octane	n-octane	111-65-9	111-65-9	-5.28	-0.06	-5.22
2084	x		1,1,1-trichloroethane	n-octanol	71-55-6	111-87-5	-3.69	-	-
2085	x	x	1-1-2-trichloro-1-2-2-trifluoroethane	n-octanol	76-13-1	111-87-5	-2.54	-0.50	-2.04
2086	x	x	1,1,2-trichloroethane	n-octanol	79-00-5	111-87-5	-4.53	-2.65	-1.88
2087	x	x	z-1,2-dichloroethene	n-octanol	156-59-2	111-87-5	-3.61	-1.90	-1.71
2088	x	x	1,2-ethanediol	n-octanol	107-21-1	111-87-5	-7.44	-4.59	-2.85
2089	x	x	1,4-dioxane	n-octanol	123-91-1	111-87-5	-4.89	-2.66	-2.23
2090	x	x	1-butene	n-octanol	106-98-9	111-87-5	-1.89	-0.59	-1.30
2091	x	x	1-chloropropane	n-octanol	540-54-5	111-87-5	-3.06	-1.84	-1.22
2092	x	x	1-hexene	n-octanol	592-41-6	111-87-5	-2.94	-0.63	-2.31
2093	x	x	1-hexyne	n-octanol	693-02-7	111-87-5	-3.43	-1.57	-1.86
2094	x	x	1-nitrobutane	n-octanol	627-05-4	111-87-5	-5.11	-3.94	-1.17
2095	x	x	1-pentyne	n-octanol	627-19-0	111-87-5	-2.79	-1.64	-1.15
2096	x	x	1-propanethiol	n-octanol	107-03-9	111-87-5	-3.52	-2.03	-1.49
2097	x	x	2-2-dimethylpropane	n-octanol	463-82-1	111-87-5	-1.74	-0.17	-1.57
2098	x	x	2-chloropropane	n-octanol	75-29-6	111-87-5	-2.84	-1.91	-0.93
2099	x	x	2-heptanone	n-octanol	110-43-0	111-87-5	-5.65	-3.00	-2.65
2100	x	x	o-nitrotoluene	n-octanol	88-72-2	111-87-5	-6.80	-3.55	-3.25
2101	x	x	2-methylpropene	n-octanol	115-11-7	111-87-5	-2.03	-0.64	-1.39
2102	x	x	2-nitropropane	n-octanol	79-46-9	111-87-5	-4.23	-3.69	-0.54
2103	x	x	2-pentanone	n-octanol	107-87-9	111-87-5	-4.35	-2.98	-1.37
2104	x	x	3-3-dimethylbutanone	n-octanol	75-97-8	111-87-5	-4.53	-2.80	-1.73
2105	x	x	aceticacid	n-octanol	64-19-7	111-87-5	-6.35	-3.62	-2.73
2106	x	x	acetone	n-octanol	67-64-1	111-87-5	-3.15	-3.17	0.02
2107	x	x	benzamide	n-octanol	55-21-0	111-87-5	-11.77	-5.62	-6.15
2108	x	x	bromobenzene	n-octanol	108-86-1	111-87-5	-5.46	-1.68	-3.78
2110	x	x	chloroform	n-octanol	67-66-3	111-87-5	-3.81	-1.72	-2.09

TABLE B.71: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2111	x	x	dichloromethane	n-octanol	75-09-2	111-87-5	-3.07	-2.22	-0.85
2112	x	x	diethylsulfide	n-octanol	352-93-2	111-87-5	-4.09	-2.04	-2.05
2113	x	x	difluorodichloromethane	n-octanol	75-71-8	111-87-5	-1.25	-0.42	-0.83
2114	x	x	dimethyl_disulfide	n-octanol	624-92-0	111-87-5	-4.24	-2.44	-1.80
2115	x	x	ethene	n-octanol	74-85-1	111-87-5	-0.27	-0.61	0.34
2116	x	x	ethyne	n-octanol	74-86-2	111-87-5	-0.51	-1.80	1.29
2117	x	x	formamide	n-octanol	75-12-7	111-87-5	-7.80	-5.62	-2.18
2118	x	x	4-butyrolactone	n-octanol	96-48-0	111-87-5	-6.83	-5.24	-1.59
2119	x	x	2-Hexanone	n-octanol	591-78-6	111-87-5	-5.02	-3.00	-2.02
2120	x	x	3-hydroxybenzaldehyde	n-octanol	100-83-4	111-87-5	-11.39	-5.39	-6.00
2121	x	x	n,n-dimethylacetamide	n-octanol	127-19-5	111-87-5	-7.48	-4.25	-3.23
2122	x	x	n,n-dimethylformamide	n-octanol	68-12-2	111-87-5	-6.14	-4.30	-1.84
2123	x	x	butanoicacid	n-octanol	107-92-6	111-87-5	-7.58	-3.46	-4.12
2124	x	x	hexanoicacid	n-octanol	142-62-1	111-87-5	-8.82	-3.30	-5.52
2125	x	x	nitroethane	n-octanol	79-24-3	111-87-5	-3.93	-4.00	0.07
2126	x	x	nitromethane	n-octanol	75-52-5	111-87-5	-3.51	-4.26	0.75
2128	x	x	pentanoicacid	n-octanol	109-52-4	111-87-5	-8.22	-3.49	-4.73
2129	x	x	propanoicacid	n-octanol	79-09-4	111-87-5	-6.86	-3.51	-3.35
2130	x	x	phenol	n-octanol	108-95-2	111-87-5	-8.69	-3.31	-5.38
2131	x	x	p_hydroxybenzaldehyde	n-octanol	123-08-0	111-87-5	-12.36	-5.57	-6.79
2132	x	x	propene	n-octanol	115-07-1	111-87-5	-1.14	-0.62	-0.52
2133	x	x	propyne	n-octanol	74-99-7	111-87-5	-1.59	-1.72	0.13
2134	x	x	s-trans-1-3-butadiene	n-octanol	106-99-0	111-87-5	-2.10	-1.01	-1.09
2135	x	x	tetramethylsilane	n-octanol	75-76-3	111-87-5	-1.79	-0.32	-1.47
2136	x	x	trichloroethene	n-octanol	79-01-6	111-87-5	-3.75	-1.16	-2.59
2137	x	x	triethyl_phosphate	n-octanol	78-40-0	111-87-5	-8.88	-5.82	-3.06
2138	x	x	trimethyl_phosphate	n-octanol	512-56-1	111-87-5	-7.81	-6.02	-1.79
2139	x	x	water	n-octanol	7732-18-5	111-87-5	-4.43	-3.80	-0.63
2140	x	x	1-butanol	o-dichlorobenzene	71-36-3	95-50-1	-3.90	-2.57	-1.33
2141	x	x	ethanol	o-dichlorobenzene	64-17-5	95-50-1	-2.34	-2.57	0.23

TABLE B.72: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2142	x	x	methanol	o-dichlorobenzene	67-56-1	95-50-1	-1.73	-2.63	0.90
2145	x	x	chloroform	n-pentane	67-66-3	109-66-0	-3.26	-0.66	-2.60
2146	x	x	methanol	n-pentane	67-56-1	109-66-0	-1.29	-1.05	-0.24
2147	x	x	n-pentane	n-pentane	109-66-0	109-66-0	-3.35	-0.05	-3.30
2148	x	x	aceticacid	1-pentanol	64-19-7	71-41-0	-6.65	-3.83	-2.82
2149	x	x	ethylbenzene	1-pentanol	100-41-4	71-41-0	-4.48	-1.41	-3.07
2150	x	x	hydrogen_peroxide	1-pentanol	7722-84-1	71-41-0	-7.46	-4.27	-3.19
2151	x	x	butanoicacid	1-pentanol	107-92-6	71-41-0	-7.74	-3.67	-4.07
2152	x	x	hexanoicacid	1-pentanol	142-62-1	71-41-0	-8.99	-3.50	-5.49
2153	x	x	propanoicacid	1-pentanol	79-09-4	71-41-0	-7.09	-3.72	-3.37
2154	x	x	phenol	1-pentanol	108-95-2	71-41-0	-8.55	-3.53	-5.02
2155	x	x	toluene	1-pentanol	108-88-3	71-41-0	-4.25	-1.43	-2.82
2156	x	x	toluene	1-propanol	108-88-3	71-23-8	-4.47	-1.48	-2.99
2157	x	x	1,4-dioxane	pyridine	123-91-1	110-86-1	-5.14	-2.77	-2.37
2158	x	x	aceticacid	2-butanol	64-19-7	78-92-2	-6.81	-3.85	-2.96
2159	x	x	water	2-butanol	7732-18-5	78-92-2	-5.71	-4.05	-1.66
2160	x	x	1,4-dioxane	tetrahydrofuran	123-91-1	109-99-9	-5.17	-2.51	-2.66
2161	x	x	1,4-dioxane	tetrahydrothiophene-s,s-dioxide	123-91-1	126-33-0	-4.90	-3.04	-1.86
2162	x	x	n-octane	tetrahydrothiophene-s,s-dioxide	111-65-9	126-33-0	-2.44	-0.17	-2.27
2163	x	x	3-3-dimethylbutanone	tetralin	75-97-8	119-64-2	-4.19	-1.70	-2.49
2164	x	x	ethanol	tetralin	64-17-5	119-64-2	-1.54	-1.56	0.02
2166	x	x	water	tetralin	7732-18-5	119-64-2	0.07	-2.37	2.44
2167	x	x	1,4-dioxane	toluene	123-91-1	108-88-3	-4.91	-1.43	-3.48
2168	x	x	methylbenzoate	toluene	93-58-3	108-88-3	-7.96	-1.70	-6.26
2169	x	x	1,4-dioxane	triethylamine	123-91-1	121-44-8	-4.41	-1.44	-2.97
2170	x	x	benzene	n-undecane	71-43-2	1120-21-4	-4.05	-0.55	-3.50
2172	x	x	chloroform	n-undecane	67-66-3	1120-21-4	-3.42	-0.74	-2.68
2174			1-1-2-2-tetrachloroethane	water	79-34-5	7732-18-5	-1.15	-3.40	2.25
2175			1-1-1-3-3-3-hexafluoro-2-propanol	water	920-66-1	7732-18-5	-3.77	-3.91	0.14
2176			1,1,1-trichloroethane	water	71-55-6	7732-18-5	-0.25	-	-

TABLE B.73: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2177			1-1-1-trifluoro-2-propanol	water	374-01-6	7732-18-5	-4.16	-3.19	-0.97
2178			1-1-2-trichloro-1-2-2-trifluoroethane	water	76-13-1	7732-18-5	1.77	-0.60	2.37
2179			1,1,2-trichloroethane	water	79-00-5	7732-18-5	-1.95	-3.09	1.14
2180			1-1-difluoroethane	water	75-37-6	7732-18-5	-0.11	-2.02	1.91
2181			1-2-diaminoethane	water	107-15-3	7732-18-5	-9.72	-4.88	-4.84
2182			o-dichlorobenzene	water	95-50-1	7732-18-5	-1.36	-2.30	0.94
2183			z-1,2-dichloroethene	water	156-59-2	7732-18-5	-0.76	-2.23	1.47
2184			z-1,2-dichloroethene	water	156-59-2	7732-18-5	-1.17	-2.23	1.06
2185			1-2-dimethoxyethane	water	110-71-4	7732-18-5	-4.84	-2.59	-2.25
2186			1,2-ethanediol	water	107-21-1	7732-18-5	-9.30	-5.36	-3.94
2187			1-4-dichlorobenzene	water	106-46-7	7732-18-5	-1.01	-2.07	1.06
2188			1,4-dioxane	water	123-91-1	7732-18-5	-5.05	-3.09	-1.96
2189			1-butanol	water	71-36-3	7732-18-5	-4.72	-3.00	-1.72
2190			1-butene	water	106-98-9	7732-18-5	1.38	-0.70	2.08
2191			1-butyne	water	107-00-6	7732-18-5	-0.16	-1.89	1.73
2192			1-chloropropane	water	540-54-5	7732-18-5	-0.27	-2.13	1.86
2193			1-heptanol	water	111-70-6	7732-18-5	-4.24	-2.92	-1.32
2194			1-hexanol	water	111-27-3	7732-18-5	-4.36	-3.03	-1.33
2195			1-hexene	water	592-41-6	7732-18-5	1.68	-0.75	2.43
2196			1-hexyne	water	693-02-7	7732-18-5	0.29	-1.83	2.12
2197			1-nitrobutane	water	627-05-4	7732-18-5	-3.08	-4.57	1.49
2198			1-nitropropane	water	108-03-2	7732-18-5	-3.34	-4.56	1.22
2199			n-octanol	water	111-87-5	7732-18-5	-4.09	-2.71	-1.38
2200			1-pentanol	water	71-41-0	7732-18-5	-4.47	-2.93	-1.54
2201			1-pentene	water	109-67-1	7732-18-5	1.66	-0.72	2.38
2202			1-pentyne	water	627-19-0	7732-18-5	0.01	-1.91	1.92
2203			1-propanethiol	water	107-03-9	7732-18-5	-1.05	-2.38	1.33
2204			1-propanol	water	71-23-8	7732-18-5	-4.83	-2.98	-1.85
2205			2,2,2-trifluoroethanol	water	75-89-8	7732-18-5	-4.31	-4.47	0.16
2206			2,2,4-trimethylpentane	water	540-84-1	7732-18-5	2.85	-0.26	3.11

TABLE B.74: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2207			2,2-dichlorobiphenyl	water	13029-08-8	7732-18-5	-2.73	-3.59	0.86
2208			dichlorphos	water	62-73-7	7732-18-5	-6.61	-6.05	-0.56
2209			2-2-dimethylpropane	water	463-82-1	7732-18-5	2.50	-0.20	2.70
2210			2-3-dichlorobiphenyl	water	16605-91-7	7732-18-5	-2.45	-3.14	0.69
2211			2,4-dimethylpentane	water	108-08-7	7732-18-5	2.88	-0.22	3.10
2212			2,4-dimethylpyridine	water	108-47-4	7732-18-5	-4.86	-2.98	-1.88
2213			2-5-dimethylpyridine	water	589-93-5	7732-18-5	-4.72	-2.90	-1.82
2214			2-6-dichlorobenzonitrile	water	1194-65-6	7732-18-5	-5.22	-5.12	-0.10
2215			2,6-dimethylpyridine	water	108-48-5	7732-18-5	-4.60	-2.63	-1.97
2216			2-chloropropane	water	75-29-6	7732-18-5	-0.25	-2.23	1.98
2218			2-ethylpyrazine	water	13925-00-3	7732-18-5	-5.51	-3.12	-2.39
2219			2-heptanone	water	110-43-0	7732-18-5	-3.04	-3.53	0.49
2220			2-methoxyethanol	water	109-86-4	7732-18-5	-6.77	-3.31	-3.46
2221			o-nitrotoluene	water	88-72-2	7732-18-5	-3.59	-4.13	0.54
2222			2-methylaniline	water	95-53-4	7732-18-5	-5.56	-3.98	-1.58
2223			2-methylpentane	water	107-83-5	7732-18-5	2.52	-0.18	2.70
2224			o-cresol	water	95-48-7	7732-18-5	-5.87	-3.68	-2.19
2225			2-methylpropane	water	75-28-5	7732-18-5	2.32	-0.15	2.47
2226			2-methylpropene	water	115-11-7	7732-18-5	1.16	-0.76	1.92
2227			2-methylpyrazine	water	109-08-0	7732-18-5	-5.57	-3.28	-2.29
2228			2-methylpyridine	water	109-06-8	7732-18-5	-4.63	-2.81	-1.82
2229			2-nitropropane	water	79-46-9	7732-18-5	-3.14	-4.26	1.12
2230			2-octanone	water	111-13-7	7732-18-5	-2.88	-3.57	0.69
2231			2-pentanone	water	107-87-9	7732-18-5	-3.53	-3.50	-0.03
2232			2-pentene	water	627-20-3	7732-18-5	1.34	-0.66	2.00
2233			2-propanol	water	67-63-0	7732-18-5	-4.76	-2.90	-1.86
2234			3-3-dimethylbutanone	water	75-97-8	7732-18-5	-2.89	-3.28	0.39
2235			3-chloro-1-propene	water	107-05-1	7732-18-5	-0.57	-2.58	2.01
2236			3-methylaniline	water	108-44-1	7732-18-5	-5.67	-4.10	-1.57

TABLE B.75: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2237			m-cresol	water	108-39-4	7732-18-5	-5.49	-3.86	-1.63
2238			3-methylpyridine	water	108-99-6	7732-18-5	-4.77	-3.11	-1.66
2239			3-pentanone	water	96-22-0	7732-18-5	-3.41	-3.29	-0.12
2240			4-heptanone	water	123-19-3	7732-18-5	-2.93	-3.30	0.37
2241			4-methylaniline	water	106-49-0	7732-18-5	-5.55	-4.06	-1.49
2242			4-methylphenol	water	106-44-5	7732-18-5	-6.14	-3.84	-2.30
2243			4-methylpyridine	water	108-89-4	7732-18-5	-4.94	-3.20	-1.74
2244			5-nonanone	water	502-56-7	7732-18-5	-2.67	-3.14	0.47
2245			acetaldehyde	water	75-07-0	7732-18-5	-3.50	-3.31	-0.19
2246			acetamide	water	60-35-5	7732-18-5	-9.71	-6.54	-3.17
2247			aceticacid	water	64-19-7	7732-18-5	-6.70	-4.16	-2.54
2248			acetone	water	67-64-1	7732-18-5	-3.85	-3.68	-0.17
2249			acetonitrile	water	75-05-8	7732-18-5	-3.89	-4.94	1.05
2250			acetophenone	water	98-86-2	7732-18-5	-4.58	-4.01	-0.57
2251			2-propen-1-ol	water	107-18-6	7732-18-5	-5.08	-3.52	-1.56
2252			aniline	water	62-53-3	7732-18-5	-5.49	-4.15	-1.34
2253			anisole	water	100-66-3	7732-18-5	-2.45	-2.68	0.23
2254			anthracene	water	120-12-7	7732-18-5	-4.23	-2.92	-1.31
2255			benzaldehyde	water	100-52-7	7732-18-5	-4.02	-3.80	-0.22
2256			benzamide	water	55-21-0	7732-18-5	-10.90	-6.58	-4.32
2257			benzene	water	71-43-2	7732-18-5	-0.87	-1.60	0.73
2258			benzonitrile	water	100-47-0	7732-18-5	-4.10	-4.69	0.59
2259			bromobenzene	water	108-86-1	7732-18-5	-1.46	-1.96	0.50
2260			butanal	water	123-72-8	7732-18-5	-3.18	-3.17	-0.01
2261			butanonitrile	water	109-74-0	7732-18-5	-3.64	-4.70	1.06
2262			butylamine	water	109-73-9	7732-18-5	-4.29	-2.62	-1.67
2263			chlorobenzene	water	108-90-7	7732-18-5	-1.12	-1.97	0.85
2264			chloroethane	water	75-00-3	7732-18-5	-0.63	-2.17	1.54

TABLE B.76: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2265			chloroform	water	67-66-3	7732-18-5	-1.07	-2.02	0.95
2266			cyclohexane	water	110-82-7	7732-18-5	1.23	-0.10	1.33
2267			cyclopentane	water	287-92-3	7732-18-5	1.20	-0.13	1.33
2268			cyclopentanol	water	96-41-3	7732-18-5	-5.49	-2.84	-2.65
2269			cyclopentanone	water	120-92-3	7732-18-5	-4.68	-3.67	-1.01
2270			cyclopropane	water	75-19-4	7732-18-5	0.75	-0.60	1.35
2271			dichloromethane	water	75-09-2	7732-18-5	-1.36	-2.58	1.22
2272			diethyl_disulfide	water	110-81-6	7732-18-5	-1.63	-2.66	1.03
2273			diethylether	water	60-29-7	7732-18-5	-1.76	-1.57	-0.19
2274			diethylsulfide	water	352-93-2	7732-18-5	-1.43	-2.41	0.98
2275			diethylamine	water	109-89-7	7732-18-5	-4.07	-1.75	-2.32
2276			difluorodichloromethane	water	75-71-8	7732-18-5	1.69	-0.50	2.19
2277			fenchlorphos	water	299-84-3	7732-18-5	-5.06	-5.90	0.84
2278			dimethyl_disulfide	water	624-92-0	7732-18-5	-1.83	-2.85	1.02
2279			dimethyl_ether	water	115-10-6	7732-18-5	-1.92	-1.79	-0.13
2280			dimethyldisulfide	water	75-18-3	7732-18-5	-1.54	-2.40	0.86
2281			dimethylamine	water	124-40-3	7732-18-5	-4.29	-1.93	-2.36
2282			dipropyl_sulfide	water	111-47-7	7732-18-5	-1.27	-2.11	0.84
2283			dipropylamine	water	142-84-7	7732-18-5	-3.66	-1.73	-1.93
2284			ethane	water	74-84-0	7732-18-5	1.83	-0.09	1.92
2285			ethanethiol	water	75-08-1	7732-18-5	-1.30	-2.39	1.09
2286			ethanol	water	64-17-5	7732-18-5	-5.01	-2.99	-2.02
2287			ethene	water	74-85-1	7732-18-5	1.27	-0.72	1.99
2288			ethylethanoate	water	141-78-6	7732-18-5	-3.10	-3.24	0.14
2289			ethylmethanoate	water	109-94-4	7732-18-5	-2.65	-3.12	0.47
2290			ethoxybenzene	water	103-73-1	7732-18-5	-2.22	-2.59	0.37
2291			ethylamine	water	75-04-7	7732-18-5	-4.50	-2.60	-1.90
2292			ethylbenzene	water	100-41-4	7732-18-5	-0.80	-1.57	0.77
2293			ethyne	water	74-86-2	7732-18-5	-0.01	-2.09	2.08
2294			fluorobenzene	water	462-06-6	7732-18-5	-0.78	-1.69	0.91

TABLE B.77: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2295			fluorotrichloromethane	water	75-69-4	7732-18-5	0.82	-0.64	1.46
2296			hydrazine	water	302-01-2	7732-18-5	-6.26	-4.51	-1.75
2297			hydrogen_peroxide	water	7722-84-1	7732-18-5	-8.58	-4.67	-3.91
2298			hydrogen_sulfide	water	7783-06-4	7732-18-5	-0.70	-2.42	1.72
2299			diisopropylether	water	108-20-3	7732-18-5	-0.53	-1.67	1.14
2300			methane	water	74-82-8	7732-18-5	2.00	-0.11	2.11
2301			methanol	water	67-56-1	7732-18-5	-5.11	-3.05	-2.06
2302			fenthion	water	55-38-9	7732-18-5	-6.92	-6.87	-0.05
2303			methylethanoate	water	79-20-9	7732-18-5	-3.32	-3.44	0.12
2304			methylbenzoate	water	93-58-3	7732-18-5	-3.91	-3.69	-0.22
2305			methylbutanoate	water	623-42-7	7732-18-5	-2.83	-3.29	0.46
2306			2-Hexanone	water	591-78-6	7732-18-5	-3.29	-3.53	0.24
2307			methyl_ethyl_ketone	water	78-93-3	7732-18-5	-3.64	-3.49	-0.15
2308			methylmethanoate	water	107-31-3	7732-18-5	-2.78	-3.31	0.53
2309			methyl_hexanoate	water	106-70-7	7732-18-5	-2.49	-3.33	0.84
2310			2-methoxypropane	water	598-53-8	7732-18-5	-2.01	-1.72	-0.29
2311			methyl_pentanoate	water	624-24-8	7732-18-5	-2.57	-3.32	0.75
2312			methylpropanoate	water	554-12-1	7732-18-5	-2.93	-3.28	0.35
2313			1-methoxypropane	water	557-17-5	7732-18-5	-1.66	-1.68	0.02
2314			methylamine	water	74-89-5	7732-18-5	-4.56	-2.72	-1.84
2315			methylcyclohexane	water	108-87-2	7732-18-5	1.71	-0.14	1.85
2316			methylhydrazine	water	60-34-4	7732-18-5	-5.31	-4.08	-1.23
2317			3-hydroxybenzaldehyde	water	100-83-4	7732-18-5	-9.51	-6.31	-3.20
2318			morpholine	water	110-91-8	7732-18-5	-7.17	-3.26	-3.91
2319			m-xylene	water	108-38-3	7732-18-5	-0.84	-1.57	0.73
2320			N-N-dimethylaniline	water	121-69-7	7732-18-5	-3.58	-2.75	-0.83
2321			naphthalene	water	91-20-3	7732-18-5	-2.39	-2.27	-0.12
2322			n-butane	water	106-97-8	7732-18-5	2.08	-0.13	2.21
2323			butanoicacid	water	107-92-6	7732-18-5	-6.36	-4.00	-2.36

TABLE B.78: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2324			butylethanoate	water	123-86-4	7732-18-5	-2.55	-3.24	0.69
2325			heptane	water	142-82-5	7732-18-5	2.62	-0.18	2.80
2326			n-hexane	water	110-54-3	7732-18-5	2.49	-0.16	2.65
2327			hexanoicacid	water	142-62-1	7732-18-5	-6.21	-3.83	-2.38
2328			nitrobenzene	water	98-95-3	7732-18-5	-4.12	-4.31	0.19
2329			nitroethane	water	79-24-3	7732-18-5	-3.71	-4.63	0.92
2330			nitromethane	water	75-52-5	7732-18-5	-3.95	-4.91	0.96
2331			N-methylaniline	water	100-61-8	7732-18-5	-4.68	-3.40	-1.28
2332			n-octane	water	111-65-9	7732-18-5	2.89	-0.18	3.07
2333			n-pentane	water	109-66-0	7732-18-5	2.33	-0.14	2.47
2334			pentanoicacid	water	109-52-4	7732-18-5	-6.16	-4.04	-2.12
2335			n-propane	water	74-98-6	7732-18-5	1.96	-0.11	2.07
2336			propanoicacid	water	79-09-4	7732-18-5	-6.47	-4.04	-2.43
2337			propylethanoate	water	109-60-4	7732-18-5	-2.86	-3.23	0.37
2338			octanal	water	124-13-0	7732-18-5	-2.29	-3.21	0.92
2339			o-xylene	water	95-47-6	7732-18-5	-0.90	-1.64	0.74
2340			pentanal	water	110-62-3	7732-18-5	-3.03	-2.97	-0.06
2341			pentylethanoate	water	628-63-7	7732-18-5	-2.45	-3.25	0.80
2342			pentylamine	water	110-58-7	7732-18-5	-4.10	-2.64	-1.46
2343			phenol	water	108-95-2	7732-18-5	-6.62	-3.89	-2.73
2344			p_hydroxybenzaldehyde	water	123-08-0	7732-18-5	-10.48	-6.49	-3.99
2345			piperazine	water	110-85-0	7732-18-5	-7.40	-3.57	-3.83
2346			piperidine	water	110-89-4	7732-18-5	-5.11	-1.71	-3.40
2347			propanal	water	123-38-6	7732-18-5	-3.44	-3.20	-0.24
2348			propene	water	115-07-1	7732-18-5	1.27	-0.72	1.99
2349			propanonitrile	water	107-12-0	7732-18-5	-3.85	-4.72	0.87
2350			propylamine	water	107-10-8	7732-18-5	-4.39	-2.61	-1.78
2351			propyne	water	74-99-7	7732-18-5	-0.31	-2.00	1.69
2352			p-xylene	water	106-42-3	7732-18-5	-0.81	-1.56	0.75
2353			pyridine	water	110-86-1	7732-18-5	-4.70	-3.02	-1.68

TABLE B.79: Experimental database containing solute and solvent names and CAS numbers, with $\Delta G^{\circ}_{s,i,j}$, $\Delta E^{el}_{i,j}$ and $G^{CDS}_{i,j}$ energies. The solute and solvent names are the inputs for the solvent names in the Gaussian09 model suite (Frisch et al., 2016). Further information regarding the energies can be found in chapters 2 and 3. The crosses in the $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ set columns indicate if a data point belongs to the nonaqueous original $\Delta G^{\circ}_{s,i,j}$ data set or the nonaqueous reduced $\Delta G^{\circ}_{s,i,j}$ and $G^{CDS}_{i,j}$ data sets found in chapter 3.

Number	$\Delta G^{\circ}_{s,i,j}$ set	$G^{CDS}_{i,j}$ set	Solute name	Solvent name	Solute CAS	Solvent CAS	$\Delta G^{\circ}_{s,i,j}$ / kcal mol^{-1}	$\Delta E^{el}_{i,j}$ / kcal mol^{-1}	$G^{CDS}_{i,j}$ / kcal mol^{-1}
2354			pyrrolidine	water	123-75-1	7732-18-5	-5.48	-20.88	15.40
2355			s-trans-1-3-butadiene	water	106-99-0	7732-18-5	0.61	-1.19	1.80
2356			t-butanol	water	75-65-0	7732-18-5	-4.51	-2.80	-1.71
2357			tert_butyl_methyl_ether	water	1634-04-4	7732-18-5	-2.21	-1.74	-0.47
2358			tetrachloroethene	water	127-18-4	7732-18-5	0.05	-0.84	0.89
2359			tetrafluoromethane	water	75-73-0	7732-18-5	3.16	-0.27	3.43
2360			tetrahydrofuran	water	109-99-9	7732-18-5	-3.47	-2.16	-1.31
2361			tetrahydropyran	water	142-68-7	7732-18-5	-3.12	-1.82	-1.30
2362			tetramethylsilane	water	75-76-3	7732-18-5	3.04	-0.39	3.43
2363			thioanisole	water	100-68-5	7732-18-5	-2.73	-3.62	0.89
2364			thiophene	water	110-02-1	7732-18-5	-1.42	-1.88	0.46
2365			thiophenol	water	108-98-5	7732-18-5	-2.55	-3.04	0.49
2366			toluene	water	108-88-3	7732-18-5	-0.89	-1.59	0.70
2367			trichloroethene	water	79-01-6	7732-18-5	-0.39	-1.36	0.97
2368			triethyl_phosphate	water	78-40-0	7732-18-5	-7.80	-6.86	-0.94
2369			trimethyl_phosphate	water	512-56-1	7732-18-5	-8.70	-7.01	-1.69
2370			trimethylamine	water	75-50-3	7732-18-5	-3.23	-1.27	-1.96
2371			tripropyl_phosphate	water	513-08-6	7732-18-5	-6.10	-5.91	-0.19
2372	x	x	n-butane	n-decane	106-97-8	124-18-5	-2.41	-0.05	-2.36
2373	x	x	n-propane	n-decane	74-98-6	124-18-5	-1.65	-0.04	-1.61
2374	x	x	n-butane	n-dodecane	106-97-8	112-40-3	-2.28	-0.05	-2.24
2375	x	x	n-propane	n-dodecane	74-98-6	112-40-3	-1.55	-0.04	-1.51
2376	x		s-trans-1-3-butadiene	styrene	106-99-0	100-42-5	-2.36	-	-
2377	x	x	n-butane	n-nonane	106-97-8	111-84-2	-2.45	-0.05	-2.41
2378	x	x	n-propane	n-nonane	74-98-6	111-84-2	-1.70	-0.04	-1.66
2379	x	x	n-butane	n-octane	106-97-8	111-65-9	-2.45	-0.04	-2.41
2380	x	x	n-propane	n-octane	74-98-6	111-65-9	-1.71	-0.04	-1.67
2381	x	x	n-butane	n-pentadecane	106-97-8	629-62-9	-2.24	-0.05	-2.20
2382	x	x	n-propane	n-pentadecane	74-98-6	629-62-9	-1.48	-0.04	-1.43

Appendix C

Extra information for the PLS, QPLS and ALAMO data-driven models

C.1 Tables for the determining of number of splits analysis

TABLE C.1: Table containing the R^2 values of the ALAMO models per split found in Figure 3.6

	R^2					
	2	3	5	10	15	20
A	0.80	0.79	0.79	0.80	0.79	0.79
B	0.79	0.79	0.80	0.79	0.79	0.79
C	0.79	0.79	0.79	0.79	0.79	0.78
D	-31.7	0.86	0.88	0.89	0.88	0.88
E	0.71	0.74	0.83	0.88	0.8	0.86
F	-471	0.86	0.87	0.88	0.89	0.88
G	-56.6	0.74	0.83	0.44	0.72	0.07

TABLE C.2: Table containing the average RMSE values of the ALAMO models per split found in Figure 3.6

	RMSE (kcal mol)-1					
	2	3	5	10	15	20
A	0.83	0.83	0.83	0.83	0.83	0.82
B	0.83	0.83	0.83	0.83	0.83	0.82
C	0.84	0.85	0.84	0.84	0.84	0.83
D	7.55	0.67	0.62	0.61	0.62	0.62
E	0.96	0.9	0.74	0.62	0.75	0.64
F	27.8	0.69	0.65	0.64	0.61	0.62
G	10.0	0.9	0.75	1.05	0.84	1.00

Table C.3: Table containing the average bias values of the ALAMO models per split found in Figure 3.6

	R^2					
	2	3	5	10	15	20
A	-0.00271	0.00406	-0.00186	-0.000768	0.00117	0.000997
B	-0.00634	0.00397	-0.00129	0.00322	0.0015	-0.00235
C	-0.00349	-0.000667	-0.00144	0.000369	0.00176	-0.0021
D	-0.42200	0.000201	-0.00484	-0.00405	0.00186	-0.00399
E	-0.00958	-0.0236	0.00829	-0.0123	-0.0162	0.0000774
F	-1.59000	-0.00947	0.00231	0.00738	-0.00332	-0.00886
G	0.17200	0.00892	-0.000675	-0.0422	-0.00721	-0.0417

Table C.4: Table containing the average model size of the ALAMO models found in Figure 3.6

	R^2					
	2	3	5	10	15	20
A	21.50	25.67	23.40	27.80	27.87	27.90
B	21.50	26.33	26.60	28.90	27.53	27.60
C	22.00	23.66	23.40	25.30	25.47	25.50
D	85.00	105.0	118.2	121.9	130.3	131.7
E	120.0	142.7	149.6	150.6	157.7	155.9
F	91.50	98.30	114.8	124.7	122.5	124.3
G	132.0	155.0	179.4	188.4	189.7	190.5

C.2 Breakdown of ALAMO-type models found in chapter 3.

This section contains the model constituents for the A-D10-10, HA-G10-5 and HA-E3-1 ALAMO models. The models are listed with a set of basis functions and corresponding coefficients. For the HA-E3-1 model, the model is regressed against the G_{CDS} data set found in section 4.1.4 and appendix B. In constrast, the A-D10-10 and HA-G10-5 models are regressed against the free energy of solvation data set found in sections 3.3 and 4.1.4 and appendix B.

C.2.1 ALAMO model A-D10-10

The A-D10-10 model is composed of 132 basis functions which include linear, quadratic, and bilinear terms. Table C.5 contains the basis functions and their corresponding coefficients.

C.2.2 ALAMO model HA-G10-5

The HA-G10-5 model is composed of 179 basis functions which include linear, inverse, and bilinear terms. Tables C.6 and C.7 contains the basis functions and their corresponding

coefficients.

C.2.3 ALAMO model HA-E3-1

The HA-E3-1 model is composed of 118 basis functions which include linear, inverse, bilinear and, inverse bilinear terms. Table C.8 contains the basis functions and their corresponding coefficients.

TABLE C.5: Table containing the functions and corresponding coefficients for the A-D10-10 model.

Coefficient	Function	Coefficient	Function	Coefficient	Function
-0.129079037	T_b	-1.43E-02	$T_b * \varepsilon_H$	-1.539144699	$\mu^j * q^+$
-1.27E-02	M_w	2.00E-02	$T_b * \varepsilon_L$	7.62E-03	$\mu^j * V_{mc}$
0.179549924	T_c	4.16E-04	$T_b * \mu_i$	-4.72E-03	$\mu^j * S$
-1.272096124	P_c	-5.09E-02	$T_b * q^+$	5.01E-03	$\epsilon * \sigma$
1.15E-02	V_c	-7.07E-03	$M_w * P_c$	1.54E-04	$\epsilon * \pi$
-22.60432484	V_m	6.43E-04	$M_w * \epsilon$	-3.91E-02	$\epsilon * \varepsilon_L$
-2.385604624	μ^j	2.25E-02	$M_w * n_D$	6.66E-02	$\epsilon * q^+$
0.10255058	ϵ	2.99E-04	$M_w * \sigma$	0.52222805	$n_D * \sigma$
-24.89370467	n_D	-9.87E-04	$M_w * \Delta H_v$	8.635912064	$n_D * \varepsilon_L$
-6.70E-02	σ	-4.55E-05	$M_w * \pi$	-0.820814972	$n_D * \mu_i$
0.169232911	ΔH_v	-1.45E-03	$M_w * \mu_i$	-1.68E-02	$n_D * V_{mc}$
0.22633315	$\log K_{ow}$	1.20E-02	$M_w * q^+$	0.461885344	$\sigma * \varepsilon_H$
-8.31E-02	π	6.49E-04	$M_w * q^-$	0.293514462	$\sigma * q^+$
17.34593709	ε_H	2.35E-06	$M_w * E$	-0.112323555	$\Delta H_v * \varepsilon_L$
-23.97362464	ε_L	-1.33E-02	$T_c * \mu^j$	-5.69E-03	$\Delta H_v * \mu_i$
1.315531603	μ_i	1.51E-04	$T_c * \epsilon$	-9.17E-02	$\Delta H_v * q^+$
-24.21967314	q^+	-4.27E-02	$T_c * n_D$	3.81E-02	$\Delta H_v * q^-$
5.876254708	q^-	-1.09E-03	$T_c * \sigma$	-3.23E-04	$\Delta H_v * S$
9.61E-03	V_{mc}	2.53E-04	$T_c * \Delta H_v$	-3.62E-03	$\log K_{ow} * \pi$
4.33E-03	E	-3.74E-02	$T_c * \varepsilon_H$	-0.878193453	$\log K_{ow} * \varepsilon_L$
5.85E-02	S	1.43E-03	$T_c * \mu_i$	2.031631015	$\log K_{ow} * q^+$
-13208.26488	$(T_b)^{-1}$	2.49E-02	$T_c * q^+$	4.17E-03	$\log K_{ow} * S$
35.44873647	$(M_w)^{-1}$	19.64783338	$P_c * V_m$	7.39E-02	$\pi * \varepsilon_L$
36073.07191	$(T_c)^{-1}$	-0.155276247	$P_c * \mu^j$	-6.65E-03	$\pi * \mu_i$
12.46781798	$(P_c)^{-1}$	4.51E-03	$P_c * \epsilon$	-5.69E-02	$\pi * q^+$
0.192853573	$(V_c)^{-1}$	0.896681311	$P_c * n_D$	-7.12E-02	$\pi * q^-$
8.50E-02	$(V_m)^{-1}$	-2.12E-02	$P_c * \sigma$	-3.34E-05	$\pi * V_{mc}$
1.883415713	$(\epsilon)^{-1}$	-3.96E-02	$P_c * \Delta H_v$	-2.94E-05	$\pi * E$
-58.36984607	$(n_D)^{-1}$	3.36E-04	$P_c * \pi$	1.91E-04	$\pi * S$
3.442298202	$(\Delta H)_v^{-1}$	-2.576062702	$P_c * \varepsilon_H$	-36.89150543	$\varepsilon_H * \varepsilon_L$
-6.031746752	$(\pi)^{-1}$	-1.47300261	$P_c * \varepsilon_L$	27.80740485	$\varepsilon_H * q^+$
-1.19167735	$(\varepsilon_H)^{-1}$	-3.36E-02	$P_c * \mu_i$	0.175280112	$\varepsilon_L * \mu_i$
-2.95E-05	$(\varepsilon_L)^{-1}$	0.939831411	$P_c * q^+$	57.45388076	$\varepsilon_L * q^+$
0.177899236	$(q^-)^{-1}$	0.768963307	$P_c * q^-$	8.44E-03	$\varepsilon_L * E$
-88.12728704	$(E)^{-1}$	-8.25E-02	$V_c * V_m$	-0.752742748	$\mu_i * q^+$
-86.03718143	$(S)^{-1}$	4.29E-05	$V_c * q^+$	0.642450294	$\mu_i * q^-$
1.11E-04	$T_b * M_w$	-10.69826406	$V_m * \mu^j$	1.06E-02	$\mu_i * V_{mc}$
5.08E-03	$T_b * P_c$	0.134181659	$V_m * q^+$	8.48E-05	$\mu_i * E$
1.53E-02	$T_b * \mu^j$	4.17627538	$\mu^j * n_D$	-1.63E-02	$\mu_i * S$
-1.02E-03	$T_b * \epsilon$	-1.02E-02	$\mu^j * \pi$	-31.08962556	$q^+ * q^-$
1.41E-04	$T_b * \sigma$	-1.292308385	$\mu^j * \varepsilon_H$	-1.39E-02	$q^+ * E$
-1.50E-03	$T_b * \log K_{ow}$	0.213663628	$\mu^j * \varepsilon_L$	4.71E-02	$q^- * V_{mc}$
9.93E-05	$T_b * \pi$	-6.86E-02	$\mu^j * \mu_i$	-2.27E-03	$q^- * E$
				1.63E-05	$V_{mc} * E$

TABLE C.6: Table containing the functions and corresponding coefficients of the HA-G10-5 model.

Coefficient	Function	Coefficient	Function	Coefficient	Function
-6.32E-02	T_b	1.73E-05	$T_b\pi$	-0.560634706	$P_c\varepsilon_L$
8.17E-02	M_w	5.46E-03	$T_b\varepsilon_H$	-0.335029095	P_cq^+
9.29E-03	T_c	8.68E-02	$T_b\varepsilon_L$	0.947123051	P_cq^-
-4.748204017	P_c	2.02E-03	$T_b\mu_i$	-0.58668682	$V_c\epsilon$
-21.60523915	V_c	-2.60E-02	T_bq^+	-3.826302095	$V_c\log K_{ow}$
45.57638747	V_m	2.07E-02	T_bq^-	-6.81E-02	V_cq^+
-7.101555104	μ^j	5.11E-06	T_bE	1.93E-03	$V_m\mu_i$
0.253224036	ϵ	3.87E-04	M_wT_c	0.165618221	V_mq^+
51.11167251	n_D	6.33E-04	M_wP_c	5.55E-02	$\mu^j\epsilon$
-8.99E-02	σ	-1.85E-02	M_wV_c	8.374256815	μ^jn_D
-9.00E-03	ΔH_v	1.29E-03	$M_w\epsilon$	-1.05E-02	$\mu^j\Delta H_v$
0.421969672	$\log K_{ow}$	1.35E-03	$M_w\Delta H_v$	-0.102549415	$\mu^j\log K_{ow}$
-0.1426266	π	-1.50E-04	$M_w\pi$	2.82E-03	$\mu^j\pi$
81.62115	ε_H	-2.21E-02	$M_w\varepsilon_H$	0.219838212	$\mu^j\varepsilon_H$
-23.46407304	ε_L	5.02E-03	$M_w\varepsilon_L$	0.415034208	$\mu^j\varepsilon_L$
2.746269007	μ_i	-3.30E-03	$M_w\mu_i$	-5.99E-02	$\mu^j\mu_i$
-45.31629341	q^+	-0.114561056	M_wq^+	-1.081732532	μ^jq^+
48.70877536	q^-	-0.458323762	T_cV_m	-1.48E-03	μ^jV_{mc}
0.133668602	V_{mc}	-9.83E-03	$T_c\mu^j$	2.49E-03	$\epsilon\sigma$
2.32E-02	E	-2.33E-03	$T_c\epsilon$	4.73E-04	$\epsilon\pi$
3.78E-02	S	-3.58E-02	T_cn_D	-1.82E-02	$\epsilon\varepsilon_L$
-9.14E-04	T_bM_w	-1.53E-03	$T_c\log K_{ow}$	2.20E-02	ϵq^+
2.11E-04	T_bT_c	-5.36E-02	$T_c\varepsilon_H$	24.57560919	n_Dq^+
1.30E-02	T_bP_c	4.28E-03	T_cq^+	-26.25095385	n_Dq^-
8.51E-02	T_bV_c	1.17E-03	T_cq^-	-1.37E-02	n_DE
0.558544568	T_bV_m	3.99E-06	T_cE	-0.270787572	n_DS
6.48E-03	$T_b\mu^j$	-0.233965079	$P_c\mu^j$	-4.27E-03	$\sigma\Delta H_v$
2.34E-03	$T_b\epsilon$	-0.233498712	$P_c\log K_{ow}$	1.272251443	σq^+
-0.102814613	T_bn_D	-1.73E-03	$P_c\pi$	8.74E-04	σV_{mc}
1.09E-02	$T_b\log K_{ow}$	-2.824372401	$P_c\varepsilon_H$		

TABLE C.7: Table containing the functions and corresponding coefficients of the HA-G10-5 model.

Coefficient	Function	Coefficient	Function	Coefficient	Function
1.27E-04	σE	-0.222159314	$\varepsilon_L S$	-8.65E-04	$(P_c\mu_i)^2$
-2.37E-04	σS	-4.310447189	$\mu_i q^+$	4.66E-06	$(V_c\Delta H_v)^2$
-3.75E-02	$\Delta H_v \log$	-1.222678435	$\mu_i q^-$	3.80E-04	$(V_c q^+)^2$
4.31E-04	$\Delta H_v \pi$	1.43E-02	$\mu_i V_{mc}$	-9.15E-05	$(\mu^j \epsilon)^2$
-0.153917243	$\Delta H_v \varepsilon_H$	9.63E-05	$\mu_i E$	-1.10E-05	$(\mu^j \pi)^2$
-0.295495712	$\Delta H_v \varepsilon_L$	-2.93E-02	$\mu_i S$	-2.057849051	$(\mu^j \varepsilon_L)^2$
-1.23E-03	$\Delta H_v \mu_i$	-21.57578797	$q^+ q^-$	1.36E-06	$(\mu^j V_{mc})^2$
0.201296926	$\Delta H_v q^+$	-1.88E-02	$q^+ V_{mc}$	5.26E-04	$(\epsilon n_D)^2$
-3.65E-05	$\Delta H_v E$	6.39E-04	$q^+ E$	2.81E-04	$(\epsilon \log K_{ow})^2$
-1.45E-03	$\log K_{ow} \pi$	0.366474946	$q^+ S$	-1.52E-08	$(\epsilon \pi)^2$
-1.226158946	$\log K_{ow} \varepsilon_L$	0.222980809	$q^- V_{mc}$	6.60E-05	$(n_D \pi)^2$
1.579993124	$\log K_{ow} q^+$	-3.44E-03	$q^- E$	-3.80E-02	$(n_D \mu_i)^2$
4.86E-03	$\pi \varepsilon_H$	-0.348912416	$q^- S$	-28.90972492	$(n_D q^+)^2$
-0.143159928	$\pi \varepsilon_L$	4.95E-05	$V_{mc} E$	-7.55809495	$(n_D q^-)^2$
-1.26E-02	$\pi \mu_i$	-9.62E-05	ES	-2.83E-02	$(\sigma q^+)^2$
-0.204533891	πq^+	-6.08E-04	$(T_b \varepsilon_L)^2$	-2.59E-08	$(\Delta H_v \pi)^2$
-0.310107113	πq^-	-3.00E-04	$(T_b q^+)^2$	1.53E-02	$(\Delta H_v \varepsilon_L)^2$
-3.78E-05	πE	1.72E-05	$(T_b q^-)^2$	-1.51E-05	$(\Delta H_v \mu_i)^2$
-85.09169252	$\varepsilon_H \varepsilon_L$	-7.15E-06	$(M_w P_c)^2$	-5.50E-03	$(\Delta H_v q^+)^2$
7.328824297	$\varepsilon_H \mu_i$	6.52E-05	$(M_w n_D)^2$	0.307545191	$(\log K_{ow} q^+)^2$
47.16405015	$\varepsilon_H q^+$	2.61E-09	$(M_w \pi)^2$	6.03E-06	$(\pi \mu_i)^2$
47.21829396	$\varepsilon_H q^-$	-3.17E-04	$(M_w \varepsilon_H)^2$	-4.19E-05	$(\pi q^+)^2$
0.273035553	$\varepsilon_H V_{mc}$	2.03E-06	$(M_w \mu_i)^2$	-4.41E-04	$(\pi q^-)^2$
2.79E-02	$\varepsilon_H E$	1.08E-03	$(M_w q^+)^2$	2.376922977	$(\varepsilon_H \mu_i)^2$
-0.735020042	$\varepsilon_H S$	-6.18E-05	$(T_c \varepsilon_L)^2$	-3.023301955	$(\varepsilon_L \mu_i)^2$
-1.038448733	$\varepsilon_L \mu_i$	1.09E-04	$(T_c q^+)^2$	2003.767928	$(\varepsilon_L q^+)^2$
-78.13313996	$\varepsilon_L q^+$	-3.92E-06	$(T_c q^-)^2$	-1.08E-04	$(\varepsilon_L E)^2$
-30.3964217	$\varepsilon_L q^-$	5.46E-06	$(P_c \epsilon)^2$	0.717158121	$(\mu_i q^+)^2$
0.192357933	$\varepsilon_L V_{mc}$	3.75E-02	$(P_c n_D)^2$	-0.50679265	$(\mu_i q^-)^2$
1.84E-02	$\varepsilon_L E$	-1.104432064	$(P_c \varepsilon_L)^2$	98.36402146	$(q^+ q^-)^2$

TABLE C.8: Table containing the functions and corresponding coefficients of the HA-E3-1 model.

Coefficient	Function	Coefficient	Function	Coefficient	Function
1.09E-02	T_b	18.61480333	$(M_w * \log K_{ow})^{-1}$	-9.90E-04	$M_w * \mu_i$
-9.54E-02	M_w	-682.5625301	$(M_w * \pi)^{-1}$	1.46E-02	$M_w * q^+$
4.51E-02	T_c	650.0330319	$(M_w * E)^{-1}$	-2.43E-04	$T_c * \mu_i$
-0.255316392	P_c	-13346.82853	$(T_c * \epsilon)^{-1}$	4.75E-03	$T_c * q^+$
21.20260245	V_m	-76234.18669	$(T_c * \pi)^{-1}$	-0.405299413	$P_c * \varepsilon_L$
1.504521651	μ^j	-409.8486802	$(T_c * \varepsilon_H)^{-1}$	0.803457395	$P_c * q^+$
-5.12E-02	ϵ	-9.525549454	$(T_c * q^-)^{-1}$	0.680888966	$P_c * q^-$
46.21280021	n_D	127972.0485	$(T_c * V_{mc})^{-1}$	-7.62E-05	$P_c * E$
-6.22E-02	σ	-69.13059473	$(P_c * \pi)^{-1}$	-4.403959042	$V_m * \mu^j$
8.79E-03	ΔH_v	836.6675514	$(P_c * V_{mc})^{-1}$	-0.613851526	$\mu^j * q^+$
0.234403777	$\log K_{ow}$	-299.1137301	$(P_c * S)^{-1}$	-0.690009414	$\mu^j * q^-$
5.46E-02	π	0.291985659	$(V_m * \epsilon)^{-1}$	7.30E-02	$\epsilon * q^+$
20.67765278	ε_H	-10.3602285	$(V_m * S)^{-1}$	0.383299782	$\sigma * q^+$
-2.575167035	ε_L	55.86008356	$(\epsilon * n_D)^{-1}$	-0.111253045	$\Delta H_v * \varepsilon_L$
-0.470125953	μ_i	-6.36283623	$(\epsilon * \pi)^{-1}$	3.60E-03	$\Delta H_v * \mu_i$
-8.842983297	q^+	1.90E-02	$(\epsilon * \varepsilon_H)^{-1}$	-9.53E-02	$\Delta H_v * q^+$
-6.061061901	q^-	0.192101807	$(\epsilon * q^-)^{-1}$	2.126162897	$\log K_{ow} * q^+$
-8.14E-02	V_{mc}	82.63119038	$(n_D * \sigma)^{-1}$	0.605572619	$\pi * \varepsilon_H$
-0.540512207	S	-0.600703232	$(n_D * q^-)^{-1}$	1.84E-03	$\pi * \mu_i$
-13045.17753	T_b^{-1}	1662.967829	$(n_D * E)^{-1}$	-2.70E-02	$\pi * q^+$
-876.8539142	M_w^{-1}	1.157321887	$(\sigma * \log K_{ow})^{-1}$	-3.16E-04	$\pi * V_{mc}$
18808.08809	T_c^{-1}	4.20E-03	$(\log K_{ow} * q^-)^{-1}$	1.26E-03	$\pi * S$
8.5016492	P_c^{-1}	43.56859415	$(\pi * \varepsilon_H)^{-1}$	-2.278517881	$\varepsilon_H * \mu_i$
0.532341578	V_m^{-1}	3.381112299	$(\pi * q^-)^{-1}$	-0.302434468	$\varepsilon_H * V_{mc}$
-38.53588608	ϵ^{-1}	-6959.915038	$(\pi * V_{mc})^{-1}$	27.20340602	$\varepsilon_L * q^+$
49.07015673	n_D^{-1}	-40182.35752	$(\pi * E)^{-1}$	11.66231199	$\varepsilon_L * q^-$
-72.08997219	σ^{-1}	-32670.08221	$(\pi * S)^{-1}$	-7.63E-03	$\varepsilon_L * E$
-0.345873128	$\log K_{ow}^{-1}$	-2.47E-04	$(\varepsilon_H * \varepsilon_L)^{-1}$	-0.433531674	$\mu_i * q^+$
707.1911787	π^{-1}	2.65E-02	$(\varepsilon_H * q^-)^{-1}$	-2.04E-03	$q^+ * E$
2.64784396	ε_H^{-1}	34.86960687	$(\varepsilon_H * V_{mc})^{-1}$	4.57E-04	$V_{mc} * S$
-1.06E-03	ε_L^{-1}	62.9823968	$(\varepsilon_H * E)^{-1}$		
0.357277176	$(q^-)^{-1}$	-130.4007391	$(\varepsilon_H * S)^{-1}$		
-2639.89733	V_{mc}^{-1}	2.526613083	$(q^- * V_{mc})^{-1}$		
-952.2684396	E^{-1}	234440.5252	$(V_{mc} * E)^{-1}$		
-4027.31551	S^{-1}	204393.6599	$(V_{mc} * S)^{-1}$		
229315.1508	$(T_b * M_w)^{-1}$	-92170.57384	$(E * S)^{-1}$		
-7755.400403	$(T_b * P_c)^{-1}$	1.87E-04	$T_b * M_w$		
-157.6681244	$(T_b * V_m)^{-1}$	-3.52E-03	$T_b * \mu^j$		
10663.68123	$(T_b * \epsilon)^{-1}$	-5.38E-02	$T_b * n_D$		
13717.5428	$(T_b * \pi)^{-1}$	1.86E-02	$T_b * \varepsilon_L$		
200518.4639	$(T_b * V_{mc})^{-1}$	-3.58E-02	$T_b * q^+$		
-32286.78478	$(T_b * E)^{-1}$	8.01E-03	$T_b * q^-$		
11.7614249	$(M_w * V_m)^{-1}$	8.86E-04	$M_w * \epsilon$		
-804.8085712	$(M_w * \epsilon)^{-1}$	-4.81E-03	$M_w * \log K_{ow}$		

Appendix D

List of molecules used in the experimental subsets found in chapter 5

TABLE D.1: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-dodecanol	n-dodecanol	112-53-8	112-53-8	alcohol	alcohol	SA	SA	×	
1-nonanol	1-nonanol	143-08-8	143-08-8	alcohol	alcohol	SA	SA	×	×
1-octanol	1-octanol	111-87-5	111-87-5	alcohol	alcohol	SA	SA	×	×
1-octanol	benzene	111-87-5	71-43-2	alcohol	aromatic	SA	NA	×	×
n-pentanoic acid	1-butanol	109-52-4	71-36-3	acid	alcohol	SA	SA	×	×
1-heptanol	1-heptanol	111-70-6	111-70-6	alcohol	alcohol	SA	SA	×	×
1-hexanol	1-hexanol	111-27-3	111-27-3	alcohol	alcohol	SA	SA	×	×
n-propanoic acid	2-methyl-1-propanol	79-09-4	78-83-1	acid	alcohol	SA	SA	×	×
n-hexanoic acid	n-heptane	142-62-1	142-82-5	acid	alkane	SA	NA	×	×
1-butanol	1-butanol	71-36-3	71-36-3	alcohol	alcohol	SA	SA	×	×
m-xylene	m-xylene	108-38-3	108-38-3	aromatic	aromatic	NA	NA	×	×
ethylbenzene	ethylbenzene	100-41-4	100-41-4	aromatic	aromatic	NA	NA	×	×
pentyl acetate	cyclohexane	628-63-7	110-82-7	ester	c-alkane	NA	NA	×	×
pentyl acetate	n-pentane	628-63-7	109-66-0	ester	alkane	NA	NA	×	×
n-butyl acetate	n-butyl acetate	123-86-4	123-86-4	ester	ester	NA	NA	×	×
1-heptanol	n-octane	111-70-6	111-65-9	alcohol	alkane	SA	NA	×	×
pentyl acetate	n-heptane	628-63-7	142-82-5	ester	alkane	NA	NA	×	×
n-octane	toluene	111-65-9	108-88-3	alkane	aromatic	NA	NA	×	×
o-xylene	n-hexadecane	95-47-6	544-76-3	aromatic	alkane	NA	NA	×	×
n-hexanoic acid	n-hexadecane	142-62-1	544-76-3	acid	alkane	SA	NA	×	×
pentyl acetate	n-nonane	628-63-7	111-84-2	ester	alkane	NA	NA	×	×
toluene	benzene	108-88-3	71-43-2	aromatic	aromatic	NA	NA	×	×
1-hexanol	2,2,4-trimethylpentane	111-27-3	540-84-1	alcohol	alkane	SA	NA	×	×
ethanol	1-butanol	64-17-5	71-36-3	alcohol	alcohol	SA	SA	×	×
p-xylene	n-hexane	106-42-3	110-54-3	aromatic	alkane	NA	NA	×	×
n-butyl acetate	n-hexane	123-86-4	110-54-3	ester	alkane	NA	NA	×	×
toluene	n-hexane	108-88-3	110-54-3	aromatic	alkane	NA	NA	×	×
2-propanol	2-propanol	67-63-0	67-63-0	alcohol	alcohol	SA	SA	×	×
1-pentanol	ethylbenzene	71-41-0	100-41-4	alcohol	aromatic	SA	NA	×	×
n-butyl acetate	n-nonane	123-86-4	111-84-2	ester	alkane	NA	NA	×	×
n-butyl acetate	n-decane	123-86-4	124-18-5	ester	alkane	NA	NA	×	×
n-butyl acetate	n-hexadecane	123-86-4	544-76-3	ester	alkane	NA	NA	×	×
ethylbenzene	1-nonanol	100-41-4	143-08-8	aromatic	alcohol	NA	SA	×	×
toluene	1-octanol	108-88-3	111-87-5	aromatic	alcohol	NA	SA	×	×

TABLE D.2: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-octane	2-propanol	111-65-9	67-63-0	alkane	alcohol	NA	SA	×	×
n-butyl acetate	n-pentadecane	123-86-4	629-62-9	ester	alkane	NA	NA	×	×
ethyl acetate	ethyl acetate	141-78-6	141-78-6	ester	ester	NA	NA	×	×
n-octane	1-butanol	111-65-9	71-36-3	alkane	alcohol	NA	SA	×	×
ethanol	1-octanol	64-17-5	111-87-5	alcohol	alcohol	SA	SA	×	×
n-propyl acetate	cyclohexane	109-60-4	110-82-7	ester	c-alkane	NA	NA	×	×
1-hexanol	n-dodecane	111-27-3	112-40-3	alcohol	alkane	SA	NA	×	×
1-pentanol	2,2,4-trimethylpentane	71-41-0	540-84-1	alcohol	alkane	SA	NA	×	×
methyl acetate	methyl acetate	79-20-9	79-20-9	ester	ester	NA	NA	×	×
n-propyl acetate	n-octane	109-60-4	111-65-9	ester	alkane	NA	NA	×	×
1-pentanol	n-pentane	71-41-0	109-66-0	alcohol	alkane	SA	NA	×	×
1-pentanol	n-decane	71-41-0	124-18-5	alcohol	alkane	SA	NA	×	×
n-propyl acetate	n-pentadecane	109-60-4	629-62-9	ester	alkane	NA	NA	×	×
1-butanol	ethylbenzene	71-36-3	100-41-4	alcohol	aromatic	SA	NA	×	×
benzene	1-heptanol	71-43-2	111-70-6	aromatic	alcohol	NA	SA	×	×
methyl propanoate	cyclohexane	554-12-1	110-82-7	ester	c-alkane	NA	NA	×	×
methyl propanoate	n-heptane	554-12-1	142-82-5	ester	alkane	NA	NA	×	×
1-butanol	n-hexadecane	71-36-3	544-76-3	alcohol	alkane	SA	NA	×	×
benzene	1-pentanol	71-43-2	71-41-0	aromatic	alcohol	NA	SA	×	×
methyl propanoate	n-decane	554-12-1	124-18-5	ester	alkane	NA	NA	×	×
ethyl acetate	n-octane	141-78-6	111-65-9	ester	alkane	NA	NA	×	×
ethyl acetate	n-hexadecane	141-78-6	544-76-3	ester	alkane	NA	NA	×	×
n-pentane	benzene	109-66-0	71-43-2	alkane	aromatic	NA	NA	×	×
methyl acetate	n-heptane	79-20-9	142-82-5	ester	alkane	NA	NA	×	×
1-propanol	n-octane	71-23-8	111-65-9	alcohol	alkane	SA	NA	×	×
1-propanol	n-nonane	71-23-8	111-84-2	alcohol	alkane	SA	NA	×	×
acetone	n-heptane	67-64-1	142-82-5	acetone	alkane	E	NA	×	×
acetone	n-hexane	67-64-1	110-54-3	acetone	alkane	E	NA	×	×
ethanol	2,2,4-trimethylpentane	64-17-5	540-84-1	alcohol	alkane	SA	NA	×	×
ethylamine	n-pentane	75-04-7	109-66-0	amino	alkane	SA	NA	×	×
ethylamine	cyclohexane	75-04-7	110-82-7	amino	c-alkane	SA	NA	×	×
water	benzene	7732-18-5	71-43-2	water	aromatic	SA	NA	×	
methanol	n-nonane	67-56-1	111-84-2	alcohol	alkane	SA	NA	×	×
propene	n-hexadecane	115-07-1	544-76-3	alkene	alkane	NA	NA	×	×

TABLE D.3: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-heptanoic acid	n-heptanoic acid	111-14-8	111-14-8	acid	acid	SA	SA	×	
1-decanol	1-octanol	112-30-1	111-87-5	alcohol	alcohol	SA	SA	×	×
1-decanol	1-decanol	112-30-1	112-30-1	alcohol	alcohol	SA	SA	×	×
n-hexanoic acid	2-methyl-1-propanol	142-62-1	78-83-1	acid	alcohol	SA	SA	×	×
n-hexanoic acid	1-butanol	142-62-1	71-36-3	acid	alcohol	SA	SA	×	×
n-pentanoic acid	1-pentanol	109-52-4	71-41-0	acid	alcohol	SA	SA	×	×
n-pentanoic acid	n-pentanoic acid	109-52-4	109-52-4	acid	acid	SA	SA	×	×
n-pentanoic acid	2-methyl-1-propanol	109-52-4	78-83-1	acid	alcohol	SA	SA	×	×
1-heptanol	1-octanol	111-70-6	111-87-5	alcohol	alcohol	SA	SA	×	×
1-decanol	n-hexadecane	112-30-1	544-76-3	alcohol	alkane	SA	NA	×	×
n-butanoic acid	1-butanol	107-92-6	71-36-3	acid	alcohol	SA	SA	×	×
n-butanoic acid	2-methyl-1-propanol	107-92-6	78-83-1	acid	alcohol	SA	SA	×	×
n-butanoic acid	n-butanoic acid	107-92-6	107-92-6	acid	acid	SA	SA	×	×
1-hexanol	1-octanol	111-27-3	111-87-5	alcohol	alcohol	SA	SA	×	×
n-hexanoic acid	benzene	142-62-1	71-43-2	acid	aromatic	SA	NA	×	×
butylbenzene	butylbenzene	104-51-8	104-51-8	aromatic	aromatic	NA	NA	×	×
1-heptanol	benzene	111-70-6	71-43-2	alcohol	aromatic	SA	NA	×	×
1-heptanol	toluene	111-70-6	108-88-3	alcohol	aromatic	SA	NA	×	×
1-heptanol	ethylbenzene	111-70-6	100-41-4	alcohol	aromatic	SA	NA	×	×
1-pentanol	1-pentanol	71-41-0	71-41-0	alcohol	alcohol	SA	SA	×	×
3-methyl-1-butanol	3-methyl-1-butanol	123-51-3	123-51-3	alcohol	alcohol	SA	SA	×	
n-propanoic acid	n-propanoic acid	79-09-4	79-09-4	acid	acid	SA	SA	×	×
1,2,3-trimethylbenzene (hemellitene)	1,2,3-trimethylbenzene (hemellitene)	526-73-8	526-73-8	aromatic	aromatic	NA	NA	×	
trimethylbenzene (hemellitene)	1,2,3-trimethylbenzene (hemellitene)	95-63-6	526-73-8	aromatic	aromatic	NA	NA	×	
1-pentanol	1-octanol	71-41-0	111-87-5	alcohol	alcohol	SA	SA	×	×
1-octanol	n-hexadecane	111-87-5	544-76-3	alcohol	alkane	SA	NA	×	×
1-hexanol	benzene	111-27-3	71-43-2	alcohol	aromatic	SA	NA	×	×
1-hexanol	toluene	111-27-3	108-88-3	alcohol	aromatic	SA	NA	×	×
isopropylbenzene	isopropylbenzene	98-82-8	98-82-8	aromatic	aromatic	NA	NA	×	×
n-pentanoic acid	benzene	109-52-4	71-43-2	acid	aromatic	SA	NA	×	×
o-xylene	o-xylene	95-47-6	95-47-6	aromatic	aromatic	NA	NA	×	×
1,2-diaminoethane	1,2-diaminoethane	107-15-3	107-15-3	amino	amino	SA	SA	×	
2-methyl-1-propanol	2-methyl-1-propanol	78-83-1	78-83-1	alcohol	alcohol	SA	SA	×	×
p-xylene	p-xylene	106-42-3	106-42-3	aromatic	aromatic	NA	NA	×	×

TABLE D.4: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
1-heptanol	n-hexane	111-70-6	110-54-3	alcohol	alkane	SA	NA	×	×
1-butanol	1-octanol	71-36-3	111-87-5	alcohol	alcohol	SA	SA	×	×
1-hexanol	ethylbenzene	111-27-3	100-41-4	alcohol	aromatic	SA	NA	×	×
m-xylene	n-heptane	108-38-3	142-82-5	aromatic	alkane	NA	NA	×	×
1-heptanol	n-pentane	111-70-6	109-66-0	alcohol	alkane	SA	NA	×	×
1-heptanol	n-decane	111-70-6	124-18-5	alcohol	alkane	SA	NA	×	×
1-heptanol	n-nonane	111-70-6	111-84-2	alcohol	alkane	SA	NA	×	×
1-heptanol	n-hexadecane	111-70-6	544-76-3	alcohol	alkane	SA	NA	×	×
1-heptanol	n-heptane	111-70-6	142-82-5	alcohol	alkane	SA	NA	×	×
t-amyl alcohol	t-amyl alcohol	75-85-4	75-85-4	alcohol	alcohol	SA	SA	×	
1-propanol	1-propanol	71-23-8	71-23-8	alcohol	alcohol	SA	SA	×	×
p-xylene	n-heptane	106-42-3	142-82-5	aromatic	alkane	NA	NA	×	×
pentyl acetate	n-hexane	628-63-7	110-54-3	ester	alkane	NA	NA	×	×
o-xylene	n-heptane	95-47-6	142-82-5	aromatic	alkane	NA	NA	×	×
2-butanol	2-butanol	78-92-2	78-92-2	alcohol	alcohol	SA	SA	×	×
ethylbenzene	n-undecane	100-41-4	1120-21-4	aromatic	alkane	NA	NA	×	×
1-heptanol	n-dodecane	111-70-6	112-40-3	alcohol	alkane	SA	NA	×	×
pentyl acetate	n-octane	628-63-7	111-65-9	ester	alkane	NA	NA	×	×
n-octane	benzene	111-65-9	71-43-2	alkane	aromatic	NA	NA	×	×
pentyl acetate	n-decane	628-63-7	124-18-5	ester	alkane	NA	NA	×	×
n-butanoic acid	benzene	107-92-6	71-43-2	acid	aromatic	SA	NA	×	×
ethylbenzene	n-decane	100-41-4	124-18-5	aromatic	alkane	NA	NA	×	×
m-xylene	1-octanol	108-38-3	111-87-5	aromatic	alcohol	NA	SA	×	×
m-xylene	n-hexadecane	108-38-3	544-76-3	aromatic	alkane	NA	NA	×	×
p-xylene	n-hexadecane	106-42-3	544-76-3	aromatic	alkane	NA	NA	×	×
n-pentanoic acid	n-heptane	109-52-4	142-82-5	acid	alkane	SA	NA	×	×
o-xylene	n-hexane	95-47-6	110-54-3	aromatic	alkane	NA	NA	×	×
pentyl acetate	n-hexadecane	628-63-7	544-76-3	ester	alkane	NA	NA	×	×
p-xylene	1-octanol	106-42-3	111-87-5	aromatic	alcohol	NA	SA	×	×
pentyl acetate	n-pentadecane	628-63-7	629-62-9	ester	alkane	NA	NA	×	×
1-pentanol	toluene	71-41-0	108-88-3	alcohol	aromatic	SA	NA	×	×
toluene	toluene	108-88-3	108-88-3	aromatic	aromatic	NA	NA	×	×
ethylbenzene	n-hexadecane	100-41-4	544-76-3	aromatic	alkane	NA	NA	×	×

TABLE D.5: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
1-hexanol	n-hexane	111-27-3	110-54-3	alcohol	alkane	SA	NA	×	×
1-pentanol	benzene	71-41-0	71-43-2	alcohol	aromatic	SA	NA	×	×
ethylbenzene	1-octanol	100-41-4	111-87-5	aromatic	alcohol	NA	SA	×	×
ethanol	ethanol	64-17-5	64-17-5	alcohol	alcohol	SA	SA	×	×
o-xylene	1-octanol	95-47-6	111-87-5	aromatic	alcohol	NA	SA	×	×
n-butanoic acid	n-heptane	107-92-6	142-82-5	acid	alkane	SA	NA	×	×
1-propanol	1-octanol	71-23-8	111-87-5	alcohol	alcohol	SA	SA	×	×
n-propyl acetate	n-propyl acetate	109-60-4	109-60-4	ester	ester	NA	NA	×	×
ethanol	1-propanol	64-17-5	71-23-8	alcohol	alcohol	SA	SA	×	×
t-butanol	t-butanol	75-65-0	75-65-0	alcohol	alcohol	SA	SA	×	
m-xylene	n-hexane	108-38-3	110-54-3	aromatic	alkane	NA	NA	×	×
ethylbenzene	n-hexane	100-41-4	110-54-3	aromatic	alkane	NA	NA	×	×
1-hexanol	n-nonane	111-27-3	111-84-2	alcohol	alkane	SA	NA	×	×
1-hexanol	n-pentane	111-27-3	109-66-0	alcohol	alkane	SA	NA	×	×
1-hexanol	n-decane	111-27-3	124-18-5	alcohol	alkane	SA	NA	×	×
ethyl propanoate	ethyl propanoate	105-37-3	105-37-3	ester	ester	NA	NA	×	
n-butyl acetate	cyclohexane	123-86-4	110-82-7	ester	c-alkane	NA	NA	×	×
n-butanoic acid	isopropylbenzene	107-92-6	98-82-8	acid	aromatic	SA	NA	×	×
1-hexanol	n-hexadecane	111-27-3	544-76-3	alcohol	alkane	SA	NA	×	×
1-hexanol	n-heptane	111-27-3	142-82-5	alcohol	alkane	SA	NA	×	×
n-butyl acetate	n-pentane	123-86-4	109-66-0	ester	alkane	NA	NA	×	×
1-hexanol	n-octane	111-27-3	111-65-9	alcohol	alkane	SA	NA	×	×
n-butyl acetate	n-heptane	123-86-4	142-82-5	ester	alkane	NA	NA	×	×
toluene	n-octane	108-88-3	111-65-9	aromatic	alkane	NA	NA	×	×
toluene	n-undecane	108-88-3	1120-21-4	aromatic	alkane	NA	NA	×	×
n-butyl acetate	n-octane	123-86-4	111-65-9	ester	alkane	NA	NA	×	×
toluene	n-heptane	108-88-3	142-82-5	aromatic	alkane	NA	NA	×	×
n-propanoic acid	benzene	79-09-4	71-43-2	acid	aromatic	SA	NA	×	×
n-octane	ethyl acetate	111-65-9	141-78-6	alkane	ester	NA	NA	×	×
toluene	n-decane	108-88-3	124-18-5	aromatic	alkane	NA	NA	×	×
n-pentanoic acid	n-hexadecane	109-52-4	544-76-3	acid	alkane	SA	NA	×	×
methyl propanoate	methyl propanoate	554-12-1	554-12-1	ester	ester	NA	NA	×	×
2,2,4-trimethylpentane	2,2,4-trimethylpentane	540-84-1	540-84-1	alkane	alkane	NA	NA	×	×
toluene	n-hexadecane	108-88-3	544-76-3	aromatic	alkane	NA	NA	×	×
butylamine	butylamine	109-73-9	109-73-9	amino	amino	SA	SA	×	×

TABLE D.6: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
1-butanol	benzene	71-36-3	71-43-2	alcohol	aromatic	SA	NA	×	×
n-octane	1-propanol	111-65-9	71-23-8	alkane	alcohol	NA	SA	×	×
1-pentanol	n-hexane	71-41-0	110-54-3	alcohol	alkane	SA	NA	×	×
toluene	1-nonanol	108-88-3	143-08-8	aromatic	alcohol	NA	SA	×	×
1-butanol	toluene	71-36-3	108-88-3	alcohol	aromatic	SA	NA	×	×
pentylamine	n-hexadecane	110-58-7	544-76-3	amino	alkane	SA	NA	×	×
1-pentanol	n-hexadecane	71-41-0	544-76-3	alcohol	alkane	SA	NA	×	×
2,2,4-trimethylpentane	n-hexadecane	540-84-1	544-76-3	alkane	alkane	NA	NA	×	×
n-propanoic acid	isopropylbenzene	79-09-4	98-82-8	acid	aromatic	SA	NA	×	×
n-propyl acetate	n-pentane	109-60-4	109-66-0	ester	alkane	NA	NA	×	×
n-octane	1-octanol	111-65-9	111-87-5	alkane	alcohol	NA	SA	×	×
n-propyl acetate	n-hexane	109-60-4	110-54-3	ester	alkane	NA	NA	×	×
1-pentanol	n-octane	71-41-0	111-65-9	alcohol	alkane	SA	NA	×	×
1-pentanol	n-heptane	71-41-0	142-82-5	alcohol	alkane	SA	NA	×	×
1-pentanol	n-dodecane	71-41-0	112-40-3	alcohol	alkane	SA	NA	×	×
n-propyl acetate	n-heptane	109-60-4	142-82-5	ester	alkane	NA	NA	×	×
n-propyl acetate	n-nonane	109-60-4	111-84-2	ester	alkane	NA	NA	×	×
n-propanoic acid	n-heptane	79-09-4	142-82-5	acid	alkane	SA	NA	×	×
n-propyl acetate	n-decane	109-60-4	124-18-5	ester	alkane	NA	NA	×	×
methyl butanoate	n-hexadecane	623-42-7	544-76-3	ester	alkane	NA	NA	×	×
benzene	n-heptane	71-43-2	142-82-5	aromatic	alkane	NA	NA	×	×
benzene	n-hexane	71-43-2	110-54-3	aromatic	alkane	NA	NA	×	×
n-propyl acetate	n-hexadecane	109-60-4	544-76-3	ester	alkane	NA	NA	×	×
1-pentanol	n-nonane	71-41-0	111-84-2	alcohol	alkane	SA	NA	×	×
2,4-dimethylpentane	n-hexadecane	108-08-7	544-76-3	alkane	alkane	NA	NA	×	×
1-propanol	benzene	71-23-8	71-43-2	alcohol	aromatic	SA	NA	×	×
benzene	1-nonanol	71-43-2	143-08-8	aromatic	alcohol	NA	SA	×	×
benzene	n-hexadecane	71-43-2	544-76-3	aromatic	alkane	NA	NA	×	×
benzene	n-decane	71-43-2	124-18-5	aromatic	alkane	NA	NA	×	×
acetone	benzene	67-64-1	71-43-2	acetone	aromatic	E	NA	×	×
1-butanol	n-pentane	71-36-3	109-66-0	alcohol	alkane	SA	NA	×	×
1-butanol	n-decane	71-36-3	124-18-5	alcohol	alkane	SA	NA	×	×
1-butanol	n-hexane	71-36-3	110-54-3	alcohol	alkane	SA	NA	×	×
1-butanol	n-nonane	71-36-3	111-84-2	alcohol	alkane	SA	NA	×	×

TABLE D.7: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-heptane	1-octanol	142-82-5	111-87-5	alkane	alcohol	NA	SA	×	×
benzene	1-octanol	71-43-2	111-87-5	aromatic	alcohol	NA	SA	×	×
1-propanol	toluene	71-23-8	108-88-3	alcohol	aromatic	SA	NA	×	×
1-propanol	ethylbenzene	71-23-8	100-41-4	alcohol	aromatic	SA	NA	×	×
methyl propanoate	n-pentane	554-12-1	109-66-0	ester	alkane	NA	NA	×	×
ethyl acetate	n-pentane	141-78-6	109-66-0	ester	alkane	NA	NA	×	×
1-butanol	n-octane	71-36-3	111-65-9	alcohol	alkane	SA	NA	×	×
benzene	1-hexanol	71-43-2	111-27-3	aromatic	alcohol	NA	SA	×	×
1-butanol	n-heptane	71-36-3	142-82-5	alcohol	alkane	SA	NA	×	×
methyl propanoate	n-hexane	554-12-1	110-54-3	ester	alkane	NA	NA	×	×
butylamine	n-hexane	109-73-9	110-54-3	amino	alkane	SA	NA	×	×
ethyl acetate	n-hexane	141-78-6	110-54-3	ester	alkane	NA	NA	×	×
n-hexane	benzene	110-54-3	71-43-2	alkane	aromatic	NA	NA	×	×
methyl propanoate	n-octane	554-12-1	111-65-9	ester	alkane	NA	NA	×	×
1-butanol	2,2,4-trimethylpentane	71-36-3	540-84-1	alcohol	alkane	SA	NA	×	×
ethyl acetate	cyclohexane	141-78-6	110-82-7	ester	c-alkane	NA	NA	×	×
butylamine	n-decane	109-73-9	124-18-5	amino	alkane	SA	NA	×	×
butylamine	n-heptane	109-73-9	142-82-5	amino	alkane	SA	NA	×	×
butylamine	n-nonane	109-73-9	111-84-2	amino	alkane	SA	NA	×	×
butylamine	n-undecane	109-73-9	1120-21-4	amino	alkane	SA	NA	×	×
1-hexene	n-hexadecane	592-41-6	544-76-3	alkene	alkane	NA	NA	×	×
ethyl acetate	n-heptane	141-78-6	142-82-5	ester	alkane	NA	NA	×	×
methyl propanoate	n-nonane	554-12-1	111-84-2	ester	alkane	NA	NA	×	×
2-methylpentane	n-hexadecane	107-83-5	544-76-3	alkane	alkane	NA	NA	×	×
1-butanol	n-dodecane	71-36-3	112-40-3	alcohol	alkane	SA	NA	×	×
ethyl acetate	n-nonane	141-78-6	111-84-2	ester	alkane	NA	NA	×	×
butylamine	n-octane	109-73-9	111-65-9	amino	alkane	SA	NA	×	×
ethyl acetate	n-decane	141-78-6	124-18-5	ester	alkane	NA	NA	×	×
ethanol	benzene	64-17-5	71-43-2	alcohol	aromatic	SA	NA	×	×
acetone	ethylbenzene	67-64-1	100-41-4	acetone	aromatic	E	NA	×	×
ethyl acetate	n-pentadecane	141-78-6	629-62-9	ester	alkane	NA	NA	×	×
methyl propanoate	n-pentadecane	554-12-1	629-62-9	ester	alkane	NA	NA	×	×
ethanol	toluene	64-17-5	108-88-3	alcohol	aromatic	SA	NA	×	×
acetone	isopropylbenzene	67-64-1	98-82-8	acetone	aromatic	E	NA	×	×

TABLE D.8: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
2-methylbutane	2-methylbutane	78-78-4	78-78-4	alkane	alkane	NA	NA	×	
methyl acetate	n-pentane	79-20-9	109-66-0	ester	alkane	NA	NA	×	×
propylamine	n-pentane	107-10-8	109-66-0	amino	alkane	SA	NA	×	×
propylamine	n-hexane	107-10-8	110-54-3	amino	alkane	SA	NA	×	×
methyl acetate	n-hexane	79-20-9	110-54-3	ester	alkane	NA	NA	×	×
n-hexane	2,2,4-trimethylpentane	110-54-3	540-84-1	alkane	alkane	NA	NA	×	×
methyl acetate	n-octane	79-20-9	111-65-9	ester	alkane	NA	NA	×	×
methyl acetate	cyclohexane	79-20-9	110-82-7	ester	c-alkane	NA	NA	×	×
propylamine	n-heptane	107-10-8	142-82-5	amino	alkane	SA	NA	×	×
methyl acetate	n-nonane	79-20-9	111-84-2	ester	alkane	NA	NA	×	×
n-hexane	1-octanol	110-54-3	111-87-5	alkane	alcohol	NA	SA	×	×
1-propanol	n-heptane	71-23-8	142-82-5	alcohol	alkane	SA	NA	×	×
propylamine	n-octane	107-10-8	111-65-9	amino	alkane	SA	NA	×	×
1-propanol	2,2,4-trimethylpentane	71-23-8	540-84-1	alcohol	alkane	SA	NA	×	×
methyl acetate	n-decane	79-20-9	124-18-5	ester	alkane	NA	NA	×	×
propylamine	n-decane	107-10-8	124-18-5	amino	alkane	SA	NA	×	×
propylamine	n-nonane	107-10-8	111-84-2	amino	alkane	SA	NA	×	×
propylamine	n-hexadecane	107-10-8	544-76-3	amino	alkane	SA	NA	×	×
ethanol	isopropylbenzene	64-17-5	98-82-8	alcohol	aromatic	SA	NA	×	×
methyl acetate	n-pentadecane	79-20-9	629-62-9	ester	alkane	NA	NA	×	×
1-propanol	n-hexane	71-23-8	110-54-3	alcohol	alkane	SA	NA	×	×
1-pentene	n-hexadecane	109-67-1	544-76-3	alkene	alkane	NA	NA	×	×
1-propanol	n-hexadecane	71-23-8	544-76-3	alcohol	alkane	SA	NA	×	×
1-propanol	n-pentane	71-23-8	109-66-0	alcohol	alkane	SA	NA	×	×
1-propanol	n-decane	71-23-8	124-18-5	alcohol	alkane	SA	NA	×	×
1-propanol	n-dodecane	71-23-8	112-40-3	alcohol	alkane	SA	NA	×	×
ethanol	n-hexane	64-17-5	110-54-3	alcohol	alkane	SA	NA	×	×
methyl propanoate	n-hexadecane	554-12-1	544-76-3	ester	alkane	NA	NA	×	×
methyl acetate	n-hexadecane	79-20-9	544-76-3	ester	alkane	NA	NA	×	×
ethanol	ethylbenzene	64-17-5	100-41-4	alcohol	aromatic	SA	NA	×	×
2,2-dimethylpropane	n-hexadecane	463-82-1	544-76-3	alkane	alkane	NA	NA	×	×
2-propanol	n-hexadecane	67-63-0	544-76-3	alcohol	alkane	SA	NA	×	×
acetone	n-decane	67-64-1	124-18-5	acetone	alkane	E	NA	×	×
acetone	n-octane	67-64-1	111-65-9	acetone	alkane	E	NA	×	×

TABLE D.9: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-pentane	1-octanol	109-66-0	111-87-5	alkane	alcohol	NA	SA	×	×
ethanol	n-decane	64-17-5	124-18-5	alcohol	alkane	SA	NA	×	×
acetone	n-hexadecane	67-64-1	544-76-3	acetone	alkane	E	NA	×	×
ethylamine	n-hexadecane	75-04-7	544-76-3	amino	alkane	SA	NA	×	×
ethanol	n-pentane	64-17-5	109-66-0	alcohol	alkane	SA	NA	×	×
ethanol	n-heptane	64-17-5	142-82-5	alcohol	alkane	SA	NA	×	×
ethanol	n-octane	64-17-5	111-65-9	alcohol	alkane	SA	NA	×	×
ethanol	n-nonane	64-17-5	111-84-2	alcohol	alkane	SA	NA	×	×
s-trans-1,3-butadiene	n-hexadecane	106-99-0	544-76-3	alkene	alkane	NA	NA	×	×
ethylamine	n-hexane	75-04-7	110-54-3	amino	alkane	SA	NA	×	×
ethylamine	n-heptane	75-04-7	142-82-5	amino	alkane	SA	NA	×	×
ethanol	n-dodecane	64-17-5	112-40-3	alcohol	alkane	SA	NA	×	×
ethylamine	n-octane	75-04-7	111-65-9	amino	alkane	SA	NA	×	×
ethanol	n-hexadecane	64-17-5	544-76-3	alcohol	alkane	SA	NA	×	×
1-butene	n-hexadecane	106-98-9	544-76-3	alkene	alkane	NA	NA	×	×
ethylamine	n-nonane	75-04-7	111-84-2	amino	alkane	SA	NA	×	×
ethylamine	n-decane	75-04-7	124-18-5	amino	alkane	SA	NA	×	×
2-methylpropane	n-hexadecane	75-28-5	544-76-3	alkane	alkane	NA	NA	×	×
n-butane	1-octanol	106-97-8	111-87-5	alkane	alcohol	NA	SA	×	×
water	toluene	7732-18-5	108-88-3	water	aromatic	SA	NA	×	
water	ethylbenzene	7732-18-5	100-41-4	water	aromatic	SA	NA	×	
methanol	n-hexane	67-56-1	110-54-3	alcohol	alkane	SA	NA	×	×
2-methylpropane	1-octanol	75-28-5	111-87-5	alkane	alcohol	NA	SA	×	×
n-propane	n-hexadecane	74-98-6	544-76-3	alkane	alkane	NA	NA	×	×
water	isopropylbenzene	7732-18-5	98-82-8	water	aromatic	SA	NA	×	
methanol	n-hexadecane	67-56-1	544-76-3	alcohol	alkane	SA	NA	×	×
methanol	n-decane	67-56-1	124-18-5	alcohol	alkane	SA	NA	×	×
methanol	n-heptane	67-56-1	142-82-5	alcohol	alkane	SA	NA	×	×
methanol	n-octane	67-56-1	111-65-9	alcohol	alkane	SA	NA	×	×
n-propane	1-octanol	74-98-6	111-87-5	alkane	alcohol	NA	SA	×	×
ethane	1-octanol	74-84-0	111-87-5	alkane	alcohol	NA	SA	×	×
ethene	n-hexadecane	74-85-1	544-76-3	alkene	alkane	NA	NA	×	×
water	n-hexadecane	7732-18-5	544-76-3	water	alkane	SA	NA	×	
methane	n-hexadecane	74-82-8	544-76-3	alkane	alkane	NA	NA	×	×

TABLE D.10: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
benzene	1-butanol	71-43-2	71-36-3	aromatic	alcohol	NA	SA	×	×
ethylbenzene	1-butanol	100-41-4	71-36-3	aromatic	alcohol	NA	SA	×	×
n-propanoic acid	1-butanol	79-09-4	71-36-3	acid	alcohol	SA	SA	×	×
toluene	1-butanol	108-88-3	71-36-3	aromatic	alcohol	NA	SA	×	×
2-propanol	cyclohexane	67-63-0	110-82-7	alcohol	c-alkane	SA	NA	×	×
n-butane	cyclohexane	106-97-8	110-82-7	alkane	c-alkane	NA	NA	×	×
n-octane	cyclohexane	111-65-9	110-82-7	alkane	c-alkane	NA	NA	×	×
n-pentane	cyclohexane	109-66-0	110-82-7	alkane	c-alkane	NA	NA	×	×
n-propane	cyclohexane	74-98-6	110-82-7	alkane	c-alkane	NA	NA	×	×
t-butanol	cyclohexane	75-65-0	110-82-7	alcohol	c-alkane	SA	NA	×	×
n-octane	n-decane	111-65-9	124-18-5	alkane	alkane	NA	NA	×	×
n-octane	ethanol	111-65-9	64-17-5	alkane	alcohol	NA	SA	×	×
toluene	ethanol	108-88-3	64-17-5	aromatic	alcohol	NA	SA	×	×
n-heptane	n-heptane	142-82-5	142-82-5	alkane	alkane	NA	NA	×	×
ethylbenzene	1-heptanol	100-41-4	111-70-6	aromatic	alcohol	NA	SA	×	×
toluene	1-heptanol	108-88-3	111-70-6	aromatic	alcohol	NA	SA	×	×
cyclohexane	n-hexadecane	110-82-7	544-76-3	c-alkane	alkane	NA	NA	×	×
cyclopentane	n-hexadecane	287-92-3	544-76-3	c-alkane	alkane	NA	NA	×	×
ethane	n-hexadecane	74-84-0	544-76-3	alkane	alkane	NA	NA	×	×
n-butane	n-hexadecane	106-97-8	544-76-3	alkane	alkane	NA	NA	×	×
n-butanoic acid	n-hexadecane	107-92-6	544-76-3	acid	alkane	SA	NA	×	×
n-heptane	n-hexadecane	142-82-5	544-76-3	alkane	alkane	NA	NA	×	×
n-hexane	n-hexadecane	110-54-3	544-76-3	alkane	alkane	NA	NA	×	×
n-octane	n-hexadecane	111-65-9	544-76-3	alkane	alkane	NA	NA	×	×
n-pentane	n-hexadecane	109-66-0	544-76-3	alkane	alkane	NA	NA	×	×
n-propanoic acid	n-hexadecane	79-09-4	544-76-3	acid	alkane	SA	NA	×	×
t-butanol	n-hexadecane	75-65-0	544-76-3	alcohol	alkane	SA	NA	×	×
n-hexane	n-hexane	110-54-3	110-54-3	alkane	alkane	NA	NA	×	×
n-octane	n-hexane	111-65-9	110-54-3	alkane	alkane	NA	NA	×	×
n-propanoic acid	n-hexane	79-09-4	110-54-3	acid	alkane	SA	NA	×	×
ethylbenzene	1-hexanol	100-41-4	111-27-3	aromatic	alcohol	NA	SA	×	×
toluene	1-hexanol	108-88-3	111-27-3	aromatic	alcohol	NA	SA	×	×
n-octane	2,2,4-trimethylpentane	111-65-9	540-84-1	alkane	alkane	NA	NA	×	×
n-pentane	2,2,4-trimethylpentane	109-66-0	540-84-1	alkane	alkane	NA	NA	×	×

TABLE D.11: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
n-octane	n-octane	111-65-9	111-65-9	alkane	alkane	NA	NA	×	×
2,2-dimethylpropane	1-octanol	463-82-1	111-87-5	alkane	alcohol	NA	SA	×	×
n-butanoic acid	1-octanol	107-92-6	111-87-5	acid	alcohol	SA	SA	×	×
n-hexanoic acid	1-octanol	142-62-1	111-87-5	acid	alcohol	SA	SA	×	×
n-pentanoic acid	1-octanol	109-52-4	111-87-5	acid	alcohol	SA	SA	×	×
n-propanoic acid	1-octanol	79-09-4	111-87-5	acid	alcohol	SA	SA	×	×
water	1-octanol	7732-18-5	111-87-5	water	alcohol	SA	SA	×	
methanol	n-pentane	67-56-1	109-66-0	alcohol	alkane	SA	NA	×	×
n-pentane	n-pentane	109-66-0	109-66-0	alkane	alkane	NA	NA	×	×
ethylbenzene	1-pentanol	100-41-4	71-41-0	aromatic	alcohol	NA	SA	×	×
n-butanoic acid	1-pentanol	107-92-6	71-41-0	acid	alcohol	SA	SA	×	×
n-hexanoic acid	1-pentanol	142-62-1	71-41-0	acid	alcohol	SA	SA	×	×
n-propanoic acid	1-pentanol	79-09-4	71-41-0	acid	alcohol	SA	SA	×	×
toluene	1-pentanol	108-88-3	71-41-0	aromatic	alcohol	NA	SA	×	×
toluene	1-propanol	108-88-3	71-23-8	aromatic	alcohol	NA	SA	×	×
benzene	n-undecane	71-43-2	1120-21-4	aromatic	alkane	NA	NA	×	×
1,2-diaminoethane	water	107-15-3	7732-18-5	amino	water	SA	SA	×	
1-butanol	water	71-36-3	7732-18-5	alcohol	water	SA	SA	×	
1-heptanol	water	111-70-6	7732-18-5	alcohol	water	SA	SA	×	
1-hexanol	water	111-27-3	7732-18-5	alcohol	water	SA	SA	×	
1-octanol	water	111-87-5	7732-18-5	alcohol	water	SA	SA	×	
1-pentanol	water	71-41-0	7732-18-5	alcohol	water	SA	SA	×	
1-propanol	water	71-23-8	7732-18-5	alcohol	water	SA	SA	×	
2,2,4-trimethylpentane	water	540-84-1	7732-18-5	alkane	water	NA	SA	×	
2,2-dimethylpropane	water	463-82-1	7732-18-5	alkane	water	NA	SA	×	
2,4-dimethylpentane	water	108-08-7	7732-18-5	alkane	water	NA	SA	×	
2-methylpentane	water	107-83-5	7732-18-5	alkane	water	NA	SA	×	
2-methylpropane	water	75-28-5	7732-18-5	alkane	water	NA	SA	×	
acetone	water	67-64-1	7732-18-5	acetone	water	E	SA	×	
benzene	water	71-43-2	7732-18-5	aromatic	water	NA	SA	×	
butylamine	water	109-73-9	7732-18-5	amino	water	SA	SA	×	
ethane	water	74-84-0	7732-18-5	alkane	water	NA	SA	×	
ethanol	water	64-17-5	7732-18-5	alcohol	water	SA	SA	×	
ethylamine	water	75-04-7	7732-18-5	amino	water	SA	SA	×	

TABLE D.12: List of molecules with corresponding CAS numbers, classes, types and respective experimental data subsets

Solute name	Solvent name	Solute CAS number	Solvent CAS number	Solute class	Solvent class	Solute interaction	Solvent interaction	Subset 1	Subset 2
ethylbenzene	water	100-41-4	7732-18-5	aromatic	water	NA	SA	×	
methane	water	74-82-8	7732-18-5	alkane	water	NA	SA	×	
methanol	water	67-56-1	7732-18-5	alcohol	water	SA	SA	×	
methylamine	water	74-89-5	7732-18-5	amino	water	SA	SA	×	
m-xylene	water	108-38-3	7732-18-5	aromatic	water	NA	SA	×	
n-butane	water	106-97-8	7732-18-5	alkane	water	NA	SA	×	
n-butanoic acid	water	107-92-6	7732-18-5	acid	water	SA	SA	×	
n-heptane	water	142-82-5	7732-18-5	alkane	water	NA	SA	×	
n-hexane	water	110-54-3	7732-18-5	alkane	water	NA	SA	×	
n-hexanoic acid	water	142-62-1	7732-18-5	acid	water	SA	SA	×	
n-octane	water	111-65-9	7732-18-5	alkane	water	NA	SA	×	
n-pentane	water	109-66-0	7732-18-5	alkane	water	NA	SA	×	
n-pentanoic acid	water	109-52-4	7732-18-5	acid	water	SA	SA	×	
n-propane	water	74-98-6	7732-18-5	alkane	water	NA	SA	×	
n-propanoic acid	water	79-09-4	7732-18-5	acid	water	SA	SA	×	
o-xylene	water	95-47-6	7732-18-5	aromatic	water	NA	SA	×	
pentylamine	water	110-58-7	7732-18-5	amino	water	SA	SA	×	
propylamine	water	107-10-8	7732-18-5	amino	water	SA	SA	×	
p-xylene	water	106-42-3	7732-18-5	aromatic	water	NA	SA	×	
toluene	water	108-88-3	7732-18-5	aromatic	water	NA	SA	×	
n-butane	n-decane	106-97-8	124-18-5	alkane	alkane	NA	NA	×	×
n-propane	n-decane	74-98-6	124-18-5	alkane	alkane	NA	NA	×	×
n-butane	n-dodecane	106-97-8	112-40-3	alkane	alkane	NA	NA	×	×
n-propane	n-dodecane	74-98-6	112-40-3	alkane	alkane	NA	NA	×	×
n-butane	n-nonane	106-97-8	111-84-2	alkane	alkane	NA	NA	×	×
n-propane	n-nonane	74-98-6	111-84-2	alkane	alkane	NA	NA	×	×
n-butane	n-octane	106-97-8	111-65-9	alkane	alkane	NA	NA	×	×
n-propane	n-octane	74-98-6	111-65-9	alkane	alkane	NA	NA	×	×
n-butane	n-pentadecane	106-97-8	629-62-9	alkane	alkane	NA	NA	×	×
n-propane	n-pentadecane	74-98-6	629-62-9	alkane	alkane	NA	NA	×	×

Appendix E

Metrics for the box plots found in chapter 5

E.1 Activity coefficient model metrics

E.1.1 Type of bonding interaction metrics

This section contains the box plot metrics per bonding interaction type for the activity coefficient models test found in chapter 5.

Table E.1: Table showing the minimum unsigned errors of the activity coefficient models per type of interaction.

Solute, solvent	Count	Minimum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
SA-SA	66	0.004	0.006	0.006
SA-NA	128	0.081	0.015	0.003
NA-NA	146	0.004	0.001	0.014
NA-SA	55	0.002	0.024	0.006
E-NA	8	0.063	0.073	0.109
E-SA	1	0.086	0.311	0.099

Table E.2: Table showing the maximum unsigned errors of the activity coefficient models per type of interaction.

Solute, solvent	Count	Maximum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
SA-SA	66	2.089	1.198	1.792
SA-NA	128	3.519	2.276	2.165
NA-NA	146	1.391	1.575	1.060
NA-SA	55	6.462	2.649	2.845
E-NA	8	0.845	0.413	0.416
E-SA	1	0.086	0.311	0.099

Table E.3: Table showing the median unsigned error of the activity coefficient models per type of interaction

Solute, solvent	Count	Median unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
SA-SA	66	0.201	0.200	0.184
SA-NA	128	1.127	0.617	0.335
NA-NA	146	0.262	0.293	0.205
NA-SA	55	0.533	0.546	0.401
E-NA	8	0.365	0.323	0.322
E-SA	1	0.086	0.311	0.099

E.1.2 Solute class metrics

This section contains the box plot metrics per solute class for the activity coefficient models test found in chapter 5.

Table E.4: Table showing the minimum unsigned errors of the activity coefficient models per solute class.

Solute class	Count	Minimum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acetone	9	0.063	0.073	0.099
acid	39	0.022	0.038	0.036
alcohol	119	0.004	0.006	0.003
alkane	69	0.004	0.001	0.014
alkene	6	0.036	0.267	0.199
amino	30	0.019	0.070	0.070
aromatic	63	0.002	0.002	0.006
c-alkane	2	0.089	0.173	0.036
ester	61	0.048	0.049	0.046
water	6	0.085	0.015	0.530

Table E.5: Table showing the maximum unsigned errors of the activity coefficient models per solute class.

Solute class	Count	Maximum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acetone	9	0.845	0.413	0.416
acid	39	3.519	2.276	2.165
alcohol	119	2.854	1.367	1.198
alkane	69	6.462	2.649	2.845
alkene	6	0.122	0.940	0.583
amino	30	2.089	0.956	1.792
aromatic	63	1.180	0.757	0.711
c-alkane	2	0.122	0.214	0.049
ester	61	1.391	1.173	0.901
water	6	0.438	0.579	1.749

Table E.6: Table showing the median unsigned error of the activity coefficient models per type of interaction

Solute class	Count	Median unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acetone	9	0.355	0.311	0.308
acid	39	0.548	0.469	0.669
alcohol	119	0.964	0.564	0.237
alkane	69	0.221	0.380	0.253
alkene	6	0.093	0.374	0.248
amino	30	0.728	0.315	0.407
aromatic	63	0.288	0.187	0.171
c-alkane	2	0.105	0.193	0.042
ester	61	0.841	0.347	0.250
water	6	0.193	0.153	0.886

E.1.3 Solvent class metrics

This section contains the box plot metrics per solvent class for the activity coefficient models test found in chapter 5.

Table E.7: Table showing the minimum unsigned errors of the activity coefficient models per solvent class.

Solvent class	Count	Minimum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acid	4	0.037	0.038	0.038
alcohol	78	0.004	0.006	0.006
alkane	215	0.008	0.002	0.003
amino	2	0.071	0.070	0.070
aromatic	47	0.022	0.011	0.014
c-alkane	13	0.004	0.001	0.035
ester	7	0.048	0.049	0.050
water	38	0.002	0.012	0.006

Table E.8: Table showing the maximum unsigned errors of the activity coefficient models per solvent class.

Solvent class	Count	Maximum unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acid	4	0.368	0.372	0.372
alcohol	78	1.853	1.198	1.198
alkane	215	3.519	2.276	2.165
amino	2	0.196	0.197	0.197
aromatic	47	2.854	1.519	1.632
c-alkane	13	1.130	0.742	0.393
ester	7	0.228	0.146	0.146
water	38	6.462	2.649	2.845

Table E.9: Table showing the median unsigned error of the activity coefficient models per solvent class.

Solvent class	Count	Median unsigned error / kcal mol^{-1}		
		NRTL	UNIFAC	modUNIFAC (Do)
acid	4	0.072	0.074	0.074
alcohol	78	0.378	0.325	0.239
alkane	215	0.821	0.419	0.272
amino	2	0.133	0.133	0.134
aromatic	47	0.294	0.316	0.152
c-alkane	13	0.498	0.247	0.133
ester	7	0.080	0.079	0.079
water	38	0.162	0.447	0.637

E.2 Data-driven model metrics

E.2.1 Type of bonding interaction metrics

This section contains the box plot metrics per bonding interaction type for the data-driven models test found in chapter 5.

Table E.10: Table showing the minimum unsigned errors of the data-driven models per type of interaction.

Solute, solvent	Count	Minimum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
SA-SA	41	0.005	0.003	0.004
SA-NA	123	0.000	0.005	0.001
NA-NA	142	0.001	0.002	0.003
NA-SA	36	0.005	0.001	0.008
E-NA	8	0.007	0.019	0.025

TABLE E.11: Table showing the maximum unsigned errors of the data-driven models per type of interaction.

Solute, solvent	Count	Maximum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
SA-SA	41	1.226	1.294	1.396
SA-NA	123	1.225	1.069	0.870
NA-NA	142	1.199	1.028	0.770
NA-SA	36	0.829	0.806	0.908
E-NA	8	0.333	0.393	0.435

TABLE E.12: Table showing the median unsigned errors of the data-driven models per type of interaction.

Solute, solvent	Count	Median unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
SA-SA	41	0.222	0.212	0.240
SA-NA	123	0.156	0.177	0.156
NA-NA	142	0.166	0.110	0.179
NA-SA	36	0.253	0.344	0.192
E-NA	8	0.194	0.113	0.193

E.2.2 Solute class metrics

This section contains the box plot metrics per solute class for the data-driven models test found in chapter 5.

TABLE E.13: Table showing the minimum unsigned errors of the data-driven models per solute class.

Solute class	Count	Minimum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acetone	8	0.007	0.019	0.025
acid	34	0.021	0.003	0.004
alcohol	107	0.00	0.005	0.001
alkane	55	0.003	0.001	0.003
alkene	6	0.023	0.031	0.038
amino	23	0.037	0.010	0.015
aromatic	55	0.001	0.008	0.008
c-alkane	2	0.327	0.072	0.010
ester	60	0.001	0.003	0.010

E.2.3 Solvent class metrics

This section contains the box plot metrics per solvent class for the data-driven models test found in chapter 5.

Table E.14: Table showing the maximum unsigned errors of the data-driven models per solute class.

Solute class	Count	Maximum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acetone	8	0.333	0.393	0.435
acid	34	1.225	1.069	0.829
alcohol	107	1.226	1.294	1.396
alkane	55	1.199	1.028	0.908
alkene	6	0.219	0.680	0.642
amino	23	0.708	0.308	0.653
aromatic	55	0.624	0.806	0.547
c-alkane	2	1.073	0.307	0.356
ester	60	0.666	0.652	0.673

Table E.15: Table showing the median unsigned errors of the data-driven models per solute class.

Solute class	Count	Median unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acetone	8	0.194	0.113	0.193
acid	34	0.307	0.245	0.256
alcohol	107	0.156	0.152	0.168
alkane	55	0.250	0.207	0.238
alkene	6	0.095	0.080	0.103
amino	23	0.101	0.134	0.111
aromatic	55	0.295	0.240	0.206
c-alkane	2	0.700	0.189	0.183
ester	60	0.124	0.096	0.153

Table E.16: Table showing the minimum unsigned errors of the data-driven models per solvent class.

Solvent class	Count	Minimum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acid	3	0.101	0.212	0.046
alcohol	73	0.005	0.001	0.004
alkane	213	0.00	0.003	0.001
amino	1	0.708	0.149	0.653
aromatic	41	0.007	0.005	0.007
c-alkane	13	0.032	0.004	0.048
ester	6	0.004	0.002	0.027

E.3 Nonaqueous comparison metrics

E.3.1 Type of bonding interaction metrics

This section contains the box plot metrics per bonding interaction type for the nonaqueous test found in chapter 5.

Table E.17: Table showing the maximum unsigned errors of the data-driven models per solvent class.

Solvent class	Count	Maximum unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acid	3	0.385	0.559	0.370
alcohol	73	1.226	1.294	1.396
alkane	213	1.225	1.069	0.854
amino	1	0.708	0.149	0.653
aromatic	41	0.937	0.873	0.870
c-alkane	13	0.832	0.487	0.699
ester	6	0.329	0.439	0.499

Table E.18: Table showing the median unsigned errors of the data-driven models per solvent class.

Solvent class	Count	Median unsigned error / kcal mol^{-1}		
		A-D10-10	HA-G10-5	HA-E3-1
acid	3	0.355	0.434	0.125
alcohol	73	0.222	0.275	0.232
alkane	213	0.159	0.121	0.157
amino	1	0.708	0.149	0.653
aromatic	41	0.217	0.127	0.204
c-alkane	13	0.247	0.191	0.264
ester	6	0.118	0.386	0.158

Table E.19: Table showing the minimum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Minimum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	41	0.004	0.001	0.036	0.037
SA-NA	123	0.001	0.003	0.003	0.000
NA-NA	142	0.003	0.005	0.014	0.002
NA-SA	36	0.008	0.109	0.006	0.230
E-NA	8	0.025	0.080	0.109	0.355

Table E.20: Table showing the maximum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Maximum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	41	1.396	1.172	1.198	1.471
SA-NA	123	0.870	1.392	2.165	1.403
NA-NA	142	0.770	1.044	1.060	1.020
NA-SA	36	0.908	0.753	0.711	1.110
E-NA	8	0.435	0.692	0.416	0.646

Table E.21: Table showing the median unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Median unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	41	0.24	0.190	0.172	0.195
SA-NA	123	0.156	0.482	0.323	0.227
NA-NA	142	0.179	0.211	0.205	0.242
NA-SA	36	0.192	0.366	0.310	0.558
E-NA	8	0.193	0.173	0.322	0.555

E.3.2 Solute class metrics

This section contains the box plot metrics per solute class for the nonaqueous test found in chapter 5.

Table E.22: Table showing the minimum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solute class.

Solute class	Count	Minimum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	8	0.025	0.080	0.109	0.355
acid	34	0.004	0.001	0.036	0.036
alcohol	107	0.001	0.025	0.003	0.000
alkane	55	0.003	0.007	0.014	0.025
alkene	6	0.038	0.078	0.199	0.167
amino	23	0.015	0.003	0.107	0.003
aromatic	55	0.008	0.012	0.006	0.002
c-alkane	2	0.010	0.014	0.036	0.171
ester	60	0.010	0.005	0.046	0.048

Table E.23: Table showing the maximum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solute class

Solute class	Count	Maximum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	8	0.435	0.692	0.416	0.646
acid	34	0.829	1.392	2.165	1.471
alcohol	107	1.396	1.172	1.198	1.196
alkane	55	0.908	1.044	1.060	1.020
alkene	6	0.642	0.712	0.583	0.284
amino	23	0.653	0.517	0.588	0.458
aromatic	55	0.547	0.661	0.711	1.110
c-alkane	2	0.356	0.048	0.049	0.182
ester	60	0.673	0.818	0.901	0.897

Table E.24: Table showing the median unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solute class.

Solute class	Count	Median unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	8	0.193	0.173	0.322	0.555
acid	34	0.256	0.349	0.811	0.638
alcohol	107	0.168	0.447	0.258	0.210
alkane	55	0.238	0.250	0.232	0.275
alkene	6	0.103	0.232	0.248	0.219
amino	23	0.111	0.310	0.337	0.105
aromatic	55	0.206	0.241	0.178	0.230
c-alkane	2	0.183	0.031	0.042	0.176
ester	60	0.153	0.222	0.253	0.268

E.3.3 Solvent class metrics

This section contains the box plot metrics per solvent class for the nonaqueous test found in chapter 5.

Table E.25: Table showing the minimum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solute class.

Solvent class	Count	Minimum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	3	0.046	0.026	0.038	0.037
alcohol	73	0.004	0.001	0.006	0.054
alkane	213	0.001	0.003	0.003	0.000
amino	1	0.653	0.186	0.197	0.196
aromatic	41	0.007	0.016	0.014	0.002
c-alkane	13	0.048	0.007	0.035	0.021
ester	6	0.027	0.011	0.050	0.048

Table E.26: Table showing the maximum unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solute class

Solvent class	Count	Maximum unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	3	0.370	0.345	0.107	0.105
alcohol	73	1.396	1.172	1.198	1.471
alkane	213	0.854	1.363	2.165	1.403
amino	1	0.653	0.186	0.197	0.196
aromatic	41	0.870	1.392	1.632	1.236
c-alkane	13	0.699	0.979	0.393	0.420
ester	6	0.499	0.521	0.146	0.286

Table E.27: Table showing the median unsigned errors of the HA-E3-1, modUNIFAC (Do), SAFT-γ Mie, and COSMO-SAC models per solvent class.

Solvent class	Count	Median unsigned error / kcal mol^{-1}			
		HA-E3-1	modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	3	0.125	0.202	0.042	0.039
alcohol	73	0.232	0.266	0.253	0.435
alkane	213	0.157	0.293	0.271	0.245
amino	1	0.653	0.186	0.197	0.196
aromatic	41	0.204	0.293	0.136	0.188
c-alkane	13	0.264	0.136	0.133	0.176
ester	6	0.158	0.120	0.085	0.097

E.4 Nonaqueous and aqueous comparison metrics

E.4.1 Type of bonding interaction metrics

This section contains the box plot metrics per bonding interaction type for the nonaqueous and aqueous test found in chapter 5.

Table E.28: Table showing the minimum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Minimum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	66	0.001	0.006	0.004
SA-NA	128	0.003	0.003	0.000
NA-NA	146	0.005	0.014	0.002
NA-SA	55	0.000	0.006	0.230
E-NA	8	0.08	0.109	0.355
E-SA	1	0.005	0.099	0.221

Table E.29: Table showing the maximum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Maximum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	66	2.077	1.792	1.471
SA-NA	128	1.392	2.165	1.446
NA-NA	146	1.044	1.060	1.020
NA-SA	55	0.753	2.845	2.084
E-NA	8	0.692	0.416	0.646
E-SA	1	0.005	0.099	0.221

Table E.30: Table showing the median unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per type of interaction.

Solute, solvent	Count	Median unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
SA-SA	66	0.206	0.184	0.195
SA-NA	128	0.469	0.335	0.242
NA-NA	146	0.204	0.205	0.242
NA-SA	55	0.253	0.401	0.700
E-NA	8	0.173	0.322	0.555
E-SA	1	0.005	0.099	0.221

E.4.2 Solute class metrics

This section contains the box plot metrics per solute class for the nonaqueous and aqueous test found in chapter 5.

Table E.31: Table showing the minimum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solute class.

Solute class	Count	Minimum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	9	0.005	0.099	0.221
acid	39	0.001	0.036	0.036
alcohol	119	0.025	0.003	0.000
alkane	69	0.005	0.014	0.025
alkene	6	0.078	0.199	0.167
amino	30	0.003	0.070	0.003
aromatic	63	0.000	0.006	0.002
c-alkane	2	0.014	0.036	0.171
ester	61	0.005	0.046	0.048
water	6	0.033	0.530	0.059

Table E.32: Table showing the maximum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solute class.

Solute class	Count	Maximum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	9	0.692	0.416	0.646
acid	39	1.392	2.165	1.471
alcohol	119	1.172	1.198	1.196
alkane	69	1.044	2.845	2.084
alkene	6	0.712	0.583	0.284
amino	30	2.077	1.792	0.710
aromatic	63	0.661	0.711	1.507
c-alkane	2	0.048	0.049	0.182
ester	61	0.818	0.901	0.897
water	6	0.252	1.749	1.446

Table E.33: Table showing the median unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solute class.

Solute class	Count	Median unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acetone	9	0.151	0.308	0.550
acid	39	0.324	0.669	0.635
alcohol	119	0.431	0.237	0.210
alkane	69	0.232	0.253	0.312
alkene	6	0.232	0.248	0.219
amino	30	0.342	0.407	0.124
aromatic	63	0.205	0.171	0.245
c-alkane	2	0.031	0.042	0.176
ester	61	0.217	0.250	0.265
water	6	0.192	0.886	1.189

E.4.3 Solvent class metrics

This section contains the box plot metrics per solvent class for the nonaqueous and aqueous test found in chapter 5.

Table E.34: Table showing the minimum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solvent class.

Solvent class	Count	Minimum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	4	0.026	0.038	0.037
alcohol	78	0.001	0.006	0.004
alkane	215	0.003	0.003	0.000
amino	2	0.186	0.070	0.071
aromatic	47	0.016	0.014	0.002
c-alkane	13	0.007	0.035	0.021
ester	7	0.011	0.050	0.048
water	38	0.000	0.006	0.046

TABLE E.35: Table showing the maximum unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solvent class.

Solvent class	Count	Maximum unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	4	0.494	0.372	0.368
alcohol	78	1.172	1.198	1.471
alkane	215	1.363	2.165	1.446
amino	2	0.271	0.197	0.196
aromatic	47	1.392	1.632	1.236
c-alkane	13	0.979	0.393	0.420
ester	7	0.521	0.146	0.286
water	38	2.077	2.845	2.084

TABLE E.36: Table showing the median unsigned errors of the modUNIFAC (Do), SAFT-γ Mie and COSMO-SAC models per solvent class.

Solvent class	Count	Median unsigned error / kcal mol^{-1}		
		modUNIFAC (Do)	SAFT-γ Mie	COSMO-SAC
acid	4	0.273	0.074	0.072
alcohol	78	0.263	0.239	0.412
alkane	215	0.293	0.272	0.246
amino	2	0.229	0.134	0.133
aromatic	47	0.260	0.152	0.206
c-alkane	13	0.136	0.133	0.176
ester	7	0.128	0.079	0.080
water	38	0.159	0.637	0.848