5-2021

Predicting Henry's Law constants of volatile organic compounds present in bourbon using molecular simulations.

Christopher A Abney
University of Louisville

PREDICTING HENRY'S LAW CONSTANTS OF VOLATILE ORGANIC COMPOUNDS PRESENT IN BOURBON USING MOLECULAR SIMULATIONS

By

Christopher Addison Abney
B.S., University of Louisville, 2020

A Thesis
Submitted to the Faculty of the
J.B. Speed School of Engineering of the University of Louisville
in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering (M.Eng.)
in Chemical Engineering

Department of Chemical Engineering
University of Louisville
Louisville, Kentucky

April 2021

Predicting Henry's Law Constants of Volatile Organic Compounds Present in Bourbon with MD Simulations

Submitted by:

Christopher Addison Abney

A Thesis Approved on:

4/12/2021

By the following Reading and Examination Committee:

Vance Jaeger, Thesis Chair

Eric Berson, Committee Member

Srinivasan Rasipuram, Committee Member

ACKNOWLEDGMENTS

I would like to thank Dr. Vance Jaeger for his endless help, guidance, and knowledge—I am forever grateful for his mentorship and am happy to call him a friend. I would also like to thank Dr. Eric Berson and Dr. Srinivasan Rasipuram for serving on my thesis committee and for their valuable feedback. Lastly, to all the faculty members of the department of Chemical Engineering, I want to thank you for the joys, the frustrations, and the moments of pride that have contributed towards making my time at the J.B. Speed School of Engineering an invaluable experience.

ABSTRACT

PREDICTING HENRY'S LAW CONSTANTS OF VOLATILE ORGANIC COMPOUNDS PRESENT IN BOURBON USING MOLECULAR SIMULATIONS

Chris Abney

April 12, 2021

Henry's Law describes the partitioning of molecules into liquid and gas phases at low concentrations. Henry's Law, which is based upon a species-dependent constant and the gas phase partial pressure, is useful for predicting phase behavior of dilute solutes. However, Henry's Law constants are difficult to measure experimentally or to predict using structure-property or thermodynamic models. Herein, molecular simulations were used to calculate Henry's Law constants for 18 volatile organic compounds (VOCs) present in bourbon. The novel simulations analyzed solvation thermodynamics of small organic molecules in 120 proof ethanol. A fast-growth non-equilibrium free energy method was used in which the VOC of interest was removed or added, thus affecting the overall thermodynamic properties of the system. Work distributions for forward and reverse transitions were analyzed. The Gibbs free energy of solvation for each VOC was thus estimated, which is directly related to the chemical potential of the VOC, thus providing access to Henry's law constants. Results of models were compared to values of aqueous solvation from literature. The results of the simulations were precise over multiple iterations, but a lack of experimental data with respect to solvation in ethanol-water solutions presents difficulties in assessing the accuracy of presented models.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

Henry's Law, in the simplest sense, is a proportionality between the amount of gas dissolved in a liquid and its corresponding partial pressure above the liquid. Henry's Law holds for dilute solutes. When taught in early science courses, carbonated beverages are used as an example to demonstrate the law—the solubility of carbon dioxide increases with pressure inside its container. When opened and exposed to atmospheric pressure, the solubility of carbon dioxide decreases, and gas bubbles are forced out of the liquid.

Henry's Law constants are the proportionality factors between the aqueous phase concentration and gaseous phase partial pressure and can be classified into two fundamental types. Henry's Law solubility constants, H, relate the proportionality when referring to the aqueous phase in the numerator and the gaseous phase in the denominator. Conversely, Henry's Law volatility constants, K_H, refer to the proportionality where the gaseous phase is in the numerator, with the aqueous phase in the denominator[1].

In the context of the distilled spirits industry, and more directly with bourbon, Henry's Law is intimately connected with the volatile organic compounds (VOCs) present within the bourbon and the headspace in the bourbon barrel. These compounds, in addition with non-volatile compounds determine the flavor and aroma profiles of bourbons. VOCs are also the principal emissions from bourbon production that occur primarily during the aging process where barrels are stored in warehouses for at least three years.

Focusing exclusively on bourbon, the industry generates $8.6 billion and provides more than 20,000 jobs with an annual payroll of $1 billion. For every distilling job in Kentucky, three more are created down the line from utilities to logistics. Additionally, distilling ranks second in terms of the state's share of national employment and manufacturing output with more than $2.3 billion in capital projects completed or planned through 2022. At any one time, the Commonwealth of Kentucky has a total inventory of nearly 10 million barrels of bourbon and other spirits, which accounts for almost 2 barrels for every single person living in Kentucky[2].

Additionally, Kentucky exported over $570 million in bourbon and other spirits in 2019, with top markets in Japan, Spain, Canada, and Australia. Visitors to Kentucky made 1.7 million stops at Kentucky distilleries in 2019, with the Kentucky Bourbon Trail attracting 1.3 million visits and the Kentucky Bourbon Trail Craft Tour distilleries hosting over 440,000 visits. The bourbon industry is booming with no signs of slowing any time soon, despite a global pandemic and a restless political environment. One would think that with the prevalence of alcohol in the world, scientific data involving distilled spirits would be ubiquitous[2].

However, even with alcoholic beverages playing such a prominent part of worldwide social culture, data for some, if not all, of the VOCs present within the bourbon headspace is difficult to find. In fact, the chemical components of distilled liquors are still insufficiently understood, with new 'unidentified' compounds continually discovered today. In a 2018 study, using mass spectrometry-based metabolomic approaches, a total of 879 VOCs were identified in just 24 distilled liquor samples of various types[3].

Several analytical techniques have been proposed and employed to identify the VOCs in distilled spirits[4]. The majority of studies use chromatographic methods to analyze VOCs. For identifying and quantifying the concentration of amino acids, phenolic compounds, glycerol, and ethanol, liquid chromatography is often used. For higher alcohols, esters, aldehydes, methanol, and volatile acids, gas chromatography with flame ionization is used. Additionally, some studies have used gas chromatography with mass spectroscopy (GC-MS) to determine compounds used as markers for liquor aging[4]. The issue with chromatographic methods for determining VOCs, is that, while compounds can be independently identified due to varying residence times, to allow concentration measurements, identification of said compounds depends on having known values for residence time. Several studies routinely identify 'unknown' compounds within the spirit, highlighting the imperfect nature of chromatography for this purpose.

For example, in a 2008 study of characterizing odor-active compounds in American bourbon whisky, Poisson et al. used aroma extract dilution analysis (AEDA) on a volatile fraction of an unrevealed bourbon. To isolate the VOCs in the whisky, 1:1 extraction by dilution with tap water was dried and then concentrated using a column. Then, the nonvolatile compounds were removed via high vacuum distillation, with the distillate being concentrated. By treating the distillate with sodium bicarbonate, the distillate was fractioned into the neutral/basic and the acidic volatiles. The neutral/basic fraction was concentrated and fed into a water-cooled column to yield five fractions of increasing polarity, which was then separated using ether mixtures and dried. High resolution gas chromatography—olfactometry (HRGC-O) and mass spectrometry were performed.

While the gas chromatography was being performed, a panelist was present to smell the outgoing aroma of each VOC. Linear retention indices were calculated and mass spectra were recorded. Using diethyl ether, both the neutral/basic and acidic volatiles were diluted until sets from 1:1 up to 1:4096 were obtained. Panelists performed sensory tests on these portions until no aroma could be detected by GC-O[5].

This lengthy process for determining the concentrations and aroma intensities of the various compounds is slow and scientifically imprecise. Once again, this method does not determine the identity of unknown compounds. In fact, of the 45 VOCs tabulated in Poisson's study, four were newly identified unknown compounds. Additionally, some of the identified VOCs have no measured Henry's law constant data reported for them[5]. Of these 45 compounds, 18 were chosen for this thesis based on their flavor dilution factor, as well as the availability of experimental data.

Experimental data regarding Henry's law constants are also more or less exclusively tabulated with water as the solvent. Literature data involving Henry's law constants for any solvent system other than water are few and far between, with notable exceptions such as water and methanol solutions. This lack of data is unfortunate, because Henry's law can be incredibly useful in many applications for the distilled spirts industry, where the solvent is water and ethanol. This solvent system behaves differently than pure water, despite forming an azeotrope with water. Also, some notable VOCs are entirely miscible in alcohol but insoluble in water, some are entirely miscible in water and not alcohol, and some have moderate solubility in either.

Henry's law constants in literature are also often expressed with differing unit systems, which can make comparing experimental values to literature tedious. This is

partly due to the numerator/denominator convention discussed previously, as well as different methods to calculate Henry's law constants. While the method in this thesis involves using the Gibbs free energy of solvation, Henry's Law constants can be calculated from the chemical potential, from the fugacity of the component, or from simply taking the partial pressure and concentration of the solute and solvent and dividing the two. In addition to the experimental methods mentioned previously, Henry's law constants have been estimated using molecular dynamics (MD) simulations and through group contribution methods.

The group contribution method allows structure-property relationships dependent on molecular structure influence to predict Henry's law constants. Group contribution methods assume that any given functional group makes a constant contribution to the Henry's law constant. For example, -OH groups would have different contributions than -CH_3 groups. However, use of this method is limited by the availability of literature data for any given group that has been determined in the past for a specific solvent[6].

In addition to group contribution methods, there are methods based on bond contributions, which have much wider application in calculating Henry's law constants. This is due to the fact that there are much fewer types of bonds than types of functional groups. A downside to these methods is that they are less specific than any group contribution method, so it is expected that values calculated using this method are less accurate[6].

Using the vapor pressure and aqueous solubility (concentration) of a specific chemical in each solvent is also a reasonable method for calculating a Henry's law constant, especially for compounds with low solubility. This method is straightforward

and has been used in literature to calculate values for more exotic compounds, where often this is the only method with published data. However, calculating a Henry's law constant from these two data sets can be problematic. Even when determined experimentally, error is introduced when measuring and calculating the vapor pressure of solutes and the solubilities of these solutes in any given solvent system. Thus, the final calculation for the constant relies on two values with inherent error. The resultant calculation, therefore, has magnified error[6].

The other method for calculating Henry's constants is by using MD simulations, of which there are many variations. In MD simulations of fluids, the equations of motion for a collection of molecules are solved using numerical integration over time.

During MD simulations, certain physical parameters are held constant. Simulations produce an ensemble of states that represent a sample of all possible states. Ensembles, then, are differentiated by which physical parameters are held constant. The two most common ensembles are the canonical ensemble and the isothermal-isobaric ensemble. In both systems, the number of particles, N, as well as the temperature are assumed to remain constant. With the canonical ensemble, the volume is held constant, and the temperature is controlled by modifying kinetic energy using a mathematical thermostat. For the isothermal-isobaric ensemble, as the name would suggest, pressure is controlled by modifying the box volume using a mathematical barostat, and temperature is likewise controlled with a thermostat. Thus, the canonical ensemble is often referred to as the NVT ensemble and the isothermal-isobaric ensemble is often referred to as the NPT ensemble. During an MD simulation, snapshots of the system properties including atomic positions are recorded, which, when combined, constitute a trajectory. While the

trajectory can have fluctuating values for some of the physical constants mentioned above, it is assumed that these values, when averaged, are constant and unchanging[7].

MD simulations are commonly used in industries outside of food science such as the biological field, the medical field, and in materials research. It has been demonstrated that MD simulations are capable of predicting structures of complex macromolecules with accuracy that rivals experiment[8].

In materials engineering, there has been an increasing number of articles and journals related to MD simulations. This reflects the growing desire (and capability) of understanding microscopic physical and chemical processes, which underlie the macroscopic performance of construction materials. MD allows for fundamental descriptions of physical material properties, especially in nano-engineering, where it is hard to experimentally ascertain material quality characteristics[9].

However, certain real-life events simply are not feasible to produce experimentally. Whether this is due to cost, the required repeatability, or if the work is purely theoretical in nature, simulation allows for the circumvention of these issues. These difficulties become apparent when considering Henry's Law constants with bourbon or other spirits. As stated previously, experimental methods for measuring Henry's Law constants are tedious and costly in time and resources. Moreover, it is not ideal to tamper with potential product—whether for food-safety standards, or for simply maximizing profit. Molecular simulations present the possibility of modeling the aging process of bourbon without disturbing the product and without waiting several years to collect data. Specifically, if Henry's Law constants for all VOCs in bourbon could be

accurately estimated, the product will not need to be disturbed, and only sampling of the barrel's headspace will be required.

With the justification for MD simulations laid bare, and with how ubiquitous they are in other industries, it begs the question: "Why aren't simulations more common in food science?"

While MD algorithms are well established and founded on universally accepted scientific principles, there is a knowledge barrier for entry with respect to running a simulation. While various proprietary and free open-source software packages exist, learning the underlying theory and practice behind these simulations takes time and effort. In addition to a large time commitment (months or more), many of the most advanced packages, such as GROMACS (GROningen Machine for Chemical Simulations) run on Linux (or Unix) operating systems. Learning how to navigate a new operating system, while using a terminal and keyboard commands as opposed to using Microsoft Windows with neat graphical user interfaces presents an additional learning curve. In order to run the number of simulations necessary to collect a sufficient amount of data, scripts have to be created in order to submit queued jobs to a research cluster or a distributed computing environment.

The results presented herein were the fruit of over 1,800 MD simulations of various VOCs solvated in a bourbon solution. In total, these simulations took roughly two days of processing time spread over a portion of the large Cardinal Research Cluster at the University of Louisville, which is made up of hundreds of nodes, having 16 nodes completely dedicated to running these specific calculations. Each of these 16 nodes has a high-performance graphics processing unit (GPU). Just one of the *current* best in class

GPUs (Nvidia GeForce RTX 3090) costs anywhere from $1500 to $2000 due to manufacturing shortages, but also because GPUs are generally expensive. Creating a cluster of supercomputers capable of the computational throughput requires a large capital investment. Alternatively, estimates could be made on a slower basis using less expensive computers or on cloud computing resources such as Amazon Web Services or Microsoft Azure, but these options also present disadvantages.

The accuracy of MD simulations when calculating Henry's law constants has not been established, especially for non-aqueous solvents. For instance, the compendium produced by Sander indicates that Henry's law constants range from a magnitude of 10^{40} to 10^{-14}. Additionally, with some compounds there is a large degree of variation between the values reported from different studies. For example, with (2,4-dichlorophenoxy)-ethanoic acid, there are two measured values: 1.2 and 0.14 mol/m^3-Pa, while the calculated values range from 1.8 to 5.5x10^6 mol/m^3-Pa. It is not uncommon to see independent studies differ by several orders of magnitude for a given compound. Because of the large variations reported in experimental values, it is hard to say whether the values produced by MD are accurate and precise compared to literature. However, studies by Mobley et al. prove the feasibility of using MD for calculating free energy of hydration[10, 11].

The ambiguity of currently available experimental and theoretical methods calls for a robust, reproducible theoretical framework by which researchers can estimate Henry's Law constants. Not only can theoretical estimates be made more quickly than experimental results can be measured, but theoretical estimates based on fundamental chemical physics can help to determine whether a measured experimental value is

reasonable. I envision a future in which many (thousands or more) Henry's Law constants are estimated and tabulated via MD simulations, thus providing preliminary estimates of VOC volatility in spirits.

Statistically, MD simulations are incredibly accurate provided that the interaction models used to describe atomic interactions are representative of real behaviors. For the properties needed to estimate Henry's law constants, MD systems are quick to converge to a statistically correct answer, with relatively low error.

While many publications have in the past have focused on the identification of new VOCs in whiskeys, very few have put effort into obtaining quantitative data for these compounds[5, 12]. While Salo et al. were the first group of researchers to determine odor thresholds based on quantitative data, the activity values that were calculated were actually determined from other authors' quantitative values[13].

An artificial whiskey model was created by the team and the aroma compounds were characterized based on sensory tests in omission experiments. While this study identified carbonyl compounds and straight chain ester compounds to be particularly important whiskey compounds, a separate, following study, determined that various phenols exceeded the thresholds of the carbonyl/ester compounds in a water/ethanol mixture[13].

However, the studies by Salo et al. did not use GC-Olfactometry analysis during their identification experiments. Therefore, the actual selection of aroma compounds was arbitrary and had no bearing on the aroma intensity within the artificial whiskey[5, 12].

A second 2008 study, again conducted by Poisson and Schieberle, sought to quantify the aroma compounds that the pair previously identified as being the most

important to bourbon whiskey in the prior study. The pair's previous study identified 45 odor-active areas within the bourbon, allowing for the identification of 42 unique odor compounds. This new study had the goal of quantifying the aroma compounds previously identified with the highest flavor dilution factors using stable isotope dilution assays. Subsequent goals involved calculating the odor activity values on the basis of odor threshold in water/ethanol, as well as verifying the experimental results using aroma recombination and omission experiments, similar to the studies performed by Salo et al. [5, 12].

The original identification study sorted and characterized 45 of the most odor-active volatile constituents, as well as some compounds present in the barrel headspace, with a threshold of their flavor dilution factors being >32. This created a necessary cutoff that eliminated the arbitrary nature that the two were critical of when it came to the studies by Salo et al[12].

In addition to identifying the compounds and their respective flavor dilution values, the researchers also characterized their odor quality perceived at the sniffing port, using adjectives such as 'fruity, soapy, earthy, coconut-like, phenolic, etc.' The researchers also indicated the fraction of the gas chromatograph column in which the odorant was detected (A through E), with some compounds present in a combination of two fractions (B+C, C+D, etc.). Retention indices were calculated and reported for each compound. Of particular interest, the team noted whether or not each compound had been previously identified as a whiskey VOC in literature, with some having been reported two times, and some with no reports at all. Seventeen of these compounds were newly identified in this study, with four compounds denoted as being completely unknown[12].

In this thesis, 18 of these 45 compounds were chosen for simulation and calculation of Henry's law constants under the criteria of higher flavor dilution indices. Compounds in this study were chosen based on the availability of literature values for Henry's law constants for each compound, however, as noted previously, some compounds had no literature values at all.

In addition to applications within the bourbon industry, using MD simulations to predict Henry's law constants could prove useful to industries that are adjacent to the bourbon industry. When it comes specifically to champagne, Henry's law is directly applicable to the carbonation of the beverage. While the solubility of carbon dioxide is partially due to the temperature that champagne is stored in, Henry's law itself absolutely dominates the contribution towards carbon dioxide solubility. In fact, a team of researchers built a multiparameter model in order to investigate providing the dissolved carbon dioxide content in champagne through the entire aging period. The researchers were able to demonstrate a clear correlation with the aging process and the losses of dissolved carbon dioxide, in that the longer a bottle sat corked over time, carbon dioxide was increasingly lost. The team interpreted these losses as the diffusion of gases through the cork stoppers. It was the combination of principles of diffusion with Henry's law that allowed the team to construct their model[14]. Molecular dynamics has been used in the past to study the interplay of carbon dioxide diffusion and ethanol diffusion in champagne wines. Bonhommeau et al. found that there was excellent agreement between theoretical and experimental diffusion coefficients calculated using MD and NMR. The team specifically notes the reliability of their approach and the benefit of using this

method for physical chemists aiming to model transport phenomena in water and alcohol mixtures[15].

Methods

There are two commonly used, well-established classes of simulations for calculating free energy differences between thermodynamic states by using energy gradients—equilibrium and nonequilibrium methods. The most prominent equilibrium method involves free energy perturbation (FEP) developed by Zwanzig[16]. A perturbation theory was developed in which two systems are compared. The first system has thermodynamic properties related to those of the second system, which are encapsulated in a difference between intermolecular potential energies of the two systems. Differences in the interatomic potentials between the two neighboring systems needs to be small such that thermodynamic fluctuations in each of the neighboring states allows for the observation of overlapping phase space. Ultimately, a series of many neighboring systems can be constructed by adding more nearby neighbors to span distant thermodynamic states. The FEP method unfortunately suffers in accuracy due to the very high amount of sampling needed. This is due to the exponential growth of statistical uncertainty with decreasing phase space density overlap[17]. Therefore, many closely neighboring states need to be sampled or long simulation times are needed to observe consistent overlap in phase space.

A second technique in calculating free energy differences with energy gradients is called thermodynamic integration (TI) — with three common variations: slow-growth (SGTI), fast-growth (FGTI), and discrete (DTI). Essentially, thermodynamic integration

involves the generalized force ($\delta H/\delta\lambda$), where λ is a coupling parameter that can be varied continuously (STGI and FGTI) or in discrete steps (DTI). This coupling parameter drives the system between two states, where at one state $\lambda = 1$ and a second state where $\lambda = 0$. Because of the continuous nature of λ and the speed with which λ evolves in fast-growth conditions, the system is never actually in equilibrium. Therefore, accuracy is entirely dependent on having small systems or long simulation times. While DTI avoids problems associated with the system being away from equilibrium, it runs into issues with sampling where if the free energy gradients are large for discrete λ values, numerical integration becomes computationally taxing. This issue is magnified at the states for $\lambda=1$ and $\lambda=0$ (which are incidentally the final and initial states of the system in question)[17]. The method used in this thesis is FGTI, and the analysis of results follows Jarzynski's work averaging, Bennett's Acceptance Ratio (BAR), and the Crooks Gaussian intersection method (CGI).

Jarzynski has proven that the difference in free energy ΔF is directly related to a series of nonequilibrium work computations[18] shown in Equation 1:

$$e^{-\beta\Delta F} = \langle e^{-\beta W_\tau} \rangle \quad (1)$$

where the brackets denote an average over an ensemble of n trajectories originating from a canonical ensemble. In this equation, β is the reciprocal thermal energy ($1/K_BT$) where K_B is Boltzmann's constant and T is absolute temperature, and W_τ is the work function over an arbitrary time length τ. Either the Helmholtz or the Gibbs free energy can be estimated using Equation 1. Helmholtz free energy is predicted in the NVT ensemble, while Gibbs free energy is predicted in the NPT ensemble. Through the relationship in

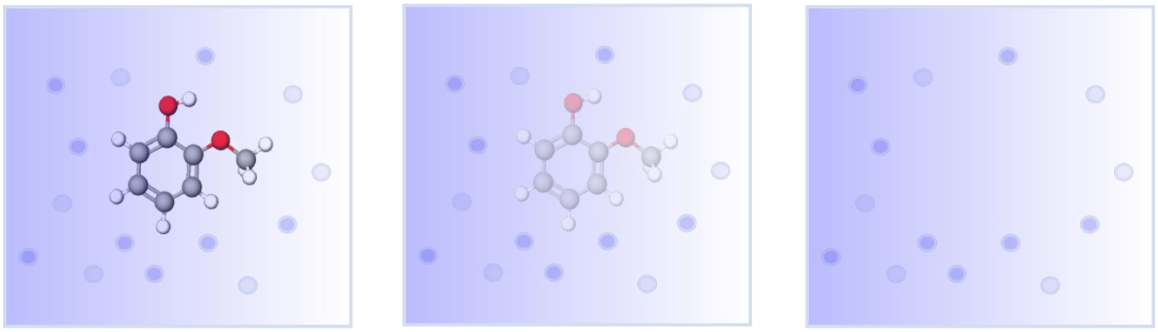
Equation 2, the Gibbs free energy is directly related to the chemical potential, because the number of moles, pressure, and temperature were held constant:

$$\mu_i = \left(\frac{\partial G}{\partial N_i}\right)_{T,P,N_{j\neq i}} \quad (2)$$

The work function W_τ is defined using the previously mentioned coupling parameter, λ, in Equation 3:

$$W_\tau = \int_0^1 \frac{\delta H_\lambda}{\delta_\lambda} d\lambda \quad (3)$$

The coupling parameter λ switches the system from a defined state A to a new state B over the simulation length τ, defined by the Hamiltonians H_A and H_B, where it is shown that $H_\lambda = (1-\lambda)H_A + \lambda H_B$. Through using a very long simulation length, and thus a large switching time from state A to state B, the system stays close enough to equilibrium conditions that it can be assumed that the dissipated work is negligible allowing the work function to represent the free energy difference, $\Delta F = W$ [17]. A visualization of the coupling parameter λ is shown in Figure 1, which depicts 2-methoxyphenol in the solvent system—specifically, the reverse ensemble where λ = 1 corresponds to the fully present VOC in the solvent system at the start of the simulation.



$\lambda = 1$ $\lambda = 0.5$ $\lambda = 0$

Figure 1. Visualization of the coupling parameter λ in the reverse ensemble of 2-methoxyphenol

In running many different simulations for each volatile compound, the probability distribution of values for the work is approximated by a Gaussian function $P(W)$ in Equation 4:

$$P_{f,r}(W) \approx \frac{1}{\sigma_{f,r}\sqrt{2\pi}} exp\left[-\frac{(W-W_{f,r})^2}{2\sigma_{f,r}^2}\right] \quad (4)$$

where $W_{f,r}$ are the means and $\sigma_{f,r}$ are the standard deviations of the work distributions. The "f" and "r" denote the forward and reverse ensembles. In the forward ensemble λ increases from 0 to 1. In the reverse ensemble λ decreases from 1 to 0 over time[17]. An alternate approach that circumvents relying on the Jarzynski equality is based on the Crooks Fluctuation Theorem where the forward and reverse ensemble distributions can be expressed as a ratio seen in Equation 5, where:

$$\frac{P_f(W)}{P_r(-W)} = e^{\beta(W-\Delta F)} \quad (5)$$

If it is assumed that the distributions for the forward and reverse work are smooth enough, a maximum likelihood on Bennett's Acceptance Ratio yields the following Equation 6 under the assumption that there is an equal number of forward and reverse ensemble distributions[17]:

$$\langle \frac{1}{1+exp[\beta(W-\Delta F)]} \rangle_f = \langle \frac{1}{1+exp[-\beta(W-\Delta F)]} \rangle_r \quad (6)$$

The difference in free energy can then be directly calculated using Equation 6.

The last method, that subsequently also employs the Crooks Fluctuation Theorem is known as the Crooks Gaussian Intersection. Each work value is calculated using Equation 3, from the individual trajectories for the switching process based on the coupling parameter. Equation 5 is the foundation upon which this method is built. Based on the equality shown, ΔF is the amount of work necessary for the left-hand side of the equation to be unity. Simply put, ΔF is the intersection point of the two work distributions, where $P_f(W) = P_r(-W)$ and is the point of interest for these simulations[17]. This intersection point is shown in Figure 2.

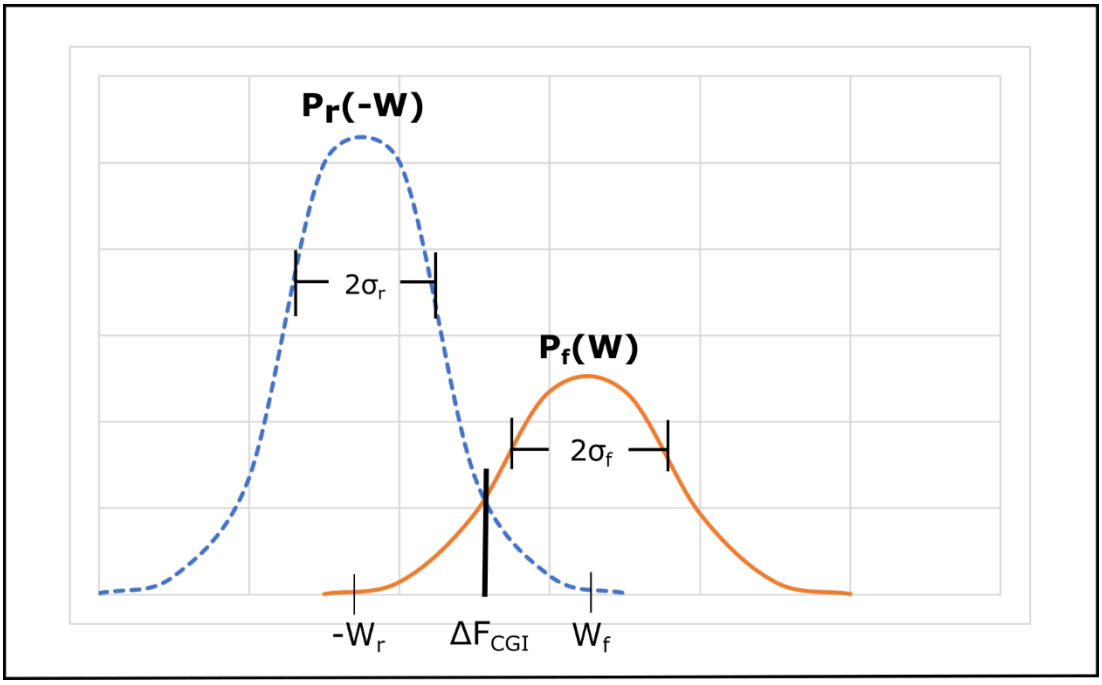


Figure 2. Illustration of the Gaussian work distributions and the intersection point. The solid line indicates the forward A to B distribution while the dashed line indicates the reverse B to A distribution. σ_f and σ_r are the standard deviations of the forward and reverse ensembles, respectively, while W_f and $-W_r$ are the means of the forward and reverse ensembles, respectively.

The intersection point, ΔF, can be directly calculated using Equation 7:

$$\Delta F = \frac{\frac{W_f}{\sigma_f^2} - \frac{-W_r}{\sigma_r^2} \pm \sqrt{\frac{1}{\sigma_f^2 \sigma_r^2}(W_f + W_r)^2 + 2\left(\frac{1}{\sigma_f^2} - \frac{1}{\sigma_r^2}\right) ln \frac{\sigma_r}{\sigma_f}}}{\frac{1}{\sigma_f^2} - \frac{1}{\sigma_r^2}} \quad (7)$$

In Equation 7, W_f and $-W_r$ are the means of the forward and reverse Gaussian functions, respectively. The standard deviations, σ_f and σ_r, are also for the forward and reverse Gaussian functions, respectively.

Goette and Grubmüller note that when $\sigma_f \neq \sigma_r$, these cases will generally have two intersection points. Because ΔF only has one unique solution, the intersection point that is located in the tail region of the distributions is generally neglected. The intersection point that lies closest to $(W_f + W_r)/2$ is accepted as the correct solution for the estimate of ΔF, the difference in free energy[17].

As was the case of all the simulations presented herein, sometimes W_f and $-W_r$ are too close to each other to calculate an intersection. Visually, this is when both distributions are on top of each other (almost appearing as one distribution) as opposed to overlapping as seen in Figure 1. In these cases, Goette and Grubmüller empirically chose to use the mean of W_f and $-W_r$ as the best estimate of the ΔF value[17]. This occurred for each of the 18 volatile organic compounds, and as such the mean was used to estimate ΔF.

It should also be noted that while directly determining the intersection point is possible using histograms, this is not advisable without a Gaussian distribution function. This is due to the large statistical error introduced when calculating the intersection in this way. The bin containing the intersection point is the only bin of work values used when calculating the intersection. Goette and Grubmüller note that this is particularly damaging when the forward and reverse distributions have a very small overlap which leads this intersection value to be miniscule or zero[17]. As such, Gaussian approximations were used in this thesis.

Goette and Grubmüller also applied a Kolmogorov-Smirnov-test to their simulated distributions to ensure that they were distributed as a Gaussian function. Testing the hypothesis that the 1000 simulated values were distributed in this manner, the

pair obtained significance levels α=0.10 and α=0.50 for the forward and reverse distributions, respectively. These significance levels implied that at the expected 5% or lower significance level, the hypothesis could not be rejected. Thus, it was assumed that the Gaussian approximation held for the tested values. Of note, with typical Gaussian distributions, there are decreasing sets of values as one moves away from the main body of the function and into the tail function. Because of the smaller data set, Kolmogorov-Smirnov-tests cannot be accurately applied for these data points. However, with most of the work distributions, the intersection of the forward and reverse ensembles generally lies in the overlapping main body areas of the two distributions, where the tail data points have no real weight. Therefore, any statistical uncertainty in these tail values has no actual weight on the final free energy estimates for the system[17].

Finally, Goette and Grubmüller tested the accuracy and convergence for each of the above methods. To test the convergence for varying numbers of trajectories, the researchers carried out test simulations on two systems with the slow-growth thermodynamic integration results as the basis for reference. For both test systems, using SGTI enabled the systems to converge after approximately 40 ns. In comparison, each of the new methods tested (except for one, not used in this thesis) converged beyond 7.5 ns. And when these new methods converged for the test systems, despite being significantly faster simulations, they agreed with the slow-growth reference result within an acceptable statistical accuracy threshold. When comparing the accuracy of the traditional methods with the newer methods proposed in their paper, the two found no significant differences in the resulting free energies that were calculated[17].

MD simulations require starting structures that contain the identities and positions of each of the atoms in the system. One of the most common formats for MD structure files is from the Protein Databank (PDB) and contains the file extension *.pdb*. While many online databases have *.pdb* files ready for download, many contain errors. Therefore, existing *.pdb* files were edited manually, or molecules were completely redrawn. Creating a *.pdb* file is straightforward using programs such as GaussView. Within the graphical user interface, one can create a ball and stick model of a molecule, with specifications of bond types, angles, dihedrals, etc. and save the molecule in the *.pdb* format, which will contain atom names, coordinates, and information about atom connectivity.

Once a *.pdb* file was created for each VOC, the next step involved generating a forcefield based on the *.pdb* file. A force field is simply the collection of equations used to describe the molecular interactions within and between each molecule in the system. During an MD simulation, the internal structure of molecules evolves over time through changing bond lengths, bond angles, and dihedrals angles. Van der Waals and electrostatic forces act between non-bonded atoms within and between molecules to additionally drive molecular rearrangement. The force field, therefore, describes the potential energies of bonds, angles, dihedrals of each molecule, as well as the electrostatics and Lenard-Jones potentials that fall into the category of non-bonded interactions. The bonded terms cover the covalent bonded potentials. Equation 8 is the harmonic energy equation generally used for force fields[19]:

$$\boldsymbol{E_{bonded}} + E_{nonbonded}$$

$$= \sum_{\boldsymbol{bonds}} \boldsymbol{K_b(b - b_0)^2} + \sum_{\boldsymbol{angles}} \boldsymbol{K_\theta(\theta - \theta_0)^2} + \sum_{\substack{\boldsymbol{improper} \\ \boldsymbol{dihedrals}}} \boldsymbol{K_\varphi(\varphi - \varphi_0)^2}$$

$$+ \sum_{\boldsymbol{dihedrals}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{6}} \boldsymbol{K_{\emptyset,n}(1 + \cos(n\emptyset - \delta_n))} + \sum_{\substack{nonbonded \\ pairs\ ij}} \frac{q_i q_j}{4\pi D r_{ij}}$$

$$+ \sum_{\substack{nonbonded \\ pairs\ ij}} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] \quad (8)$$

Where, K_b is the bond force constant, b_0 is the reference bond length, K_θ is the angle force constant, θ_0 the reference valence angle, K_φ the improper dihedral force constant, φ_0 the improper dihedral angle reference (usually 0), *n* is the dihedral multiplicity, δ_n the dihedral phase, $K_{\emptyset,n}$ the dihedral amplitude, $q_i q_j$ the partial charges, ε_{ij} the Lennard-Jones well depth , $R_{min,ij}$ the Lennard-Jones radius, and r_{ij} the distance between two particles.

For bonds, the force is spring-like in nature. For angle potentials, molecules have preferred bond angles and any deviation in this serves to change the potential energy. Lastly, there are dihedral bonded potentials which, while like bond angle, involve the orientation of the molecule in 3D space. The van der Waals forces of the molecules are approximated using the Lenard-Jones potential model, which describes the attractive and repulsive forces between atoms that arise from temporary dipoles. The repulsive and attractive forces between molecules is primarily a function of the distance between pairs as in Figure 3.

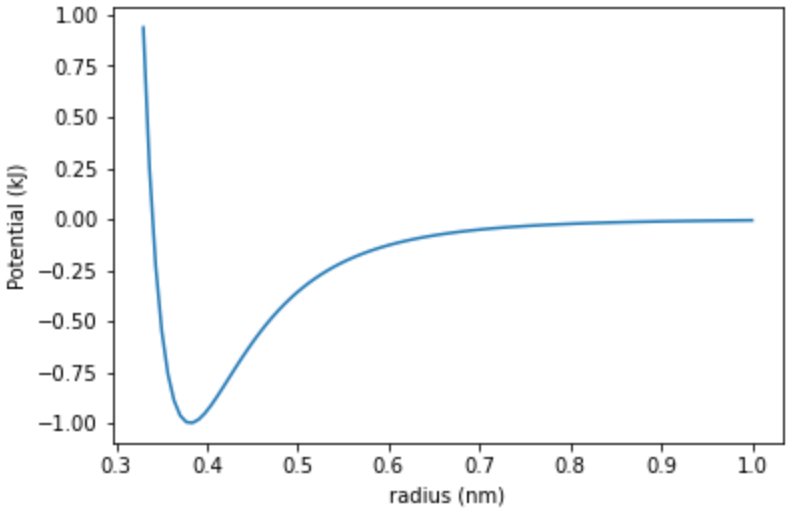


Figure 3. Plot showing the Lenard-Jones potential as a function of the distance between two particles

When atoms are too close, they have an unfavorable positive interatomic potential. When atoms are at a moderate distance, they have a favorable negative interatomic potential. When atoms are far apart, they have no interatomic potential. The second term of non-bonded potentials comes from the contribution of electrostatics between molecules. These forces are approximated by assigning a point charge to every atom in the system, and while not perfect, is a suitable approximation in most cases. In general, the non-bonded terms contribute heavily to the overall energy in the force field.

To create the forcefield, a python script was used around the ANTECHAMBER software, within the AMBER MD package, to simplify the generation of topologies and parameters for use with GROMACS. ACPYPE (AnteChamber Python Parser interfacE) is a simple script used to generate the force field of the system based on the *.pdb* files for each of the VOCs. This free, open source application readily calculates partial charges

and generates topology and parameters in different formats for use with molecular dynamics simulations[20]. The default method, and the one used in this thesis, is the BCC method, which is a semi-empirical method parameterized to reproduce specific charges. This method is slower than other methods but is more accurate. The output of this script is a usable file containing force field parameters for ethanol and each VOC taking most parameters from the General AMBER Force Field (GAFF). The force field used was the AM1-BCC force field. The water solvent model used was the TIP3P model with explicit solvent dynamics. GAFF was designed to be compatible with existing AMBER force fields for proteins and nucleic acids, but also has parameters for most organic and pharmaceutical molecules[21]. Mobley and his coworkers found that the solvation free energy of small molecules in TIP3P water is accurately predicted by MD simulations when using GAFF and AM1-BCC, as used herein. When comparing simulated values and experimental values, they found an average error of 0.47 ± 0.06 kcal/mol[10, 11]. The topology, which is a description of the connectivity of the atoms in the system, is also generated by the script. The combined force field and topology files for GROMACS have the *.itp* extension, which allows the quick addition of new, non-standard molecules into MD simulations.

To fill the simulation box with molecules and generate an initial structure for the MD simulation, Packmol was used. This application packs a specified number of molecules within a region of space, using the *.pdb* files of each of the constituent molecules as inputs. To calculate the number of water and ethanol molecules to use when filling the simulation box, basic stoichiometry was used along with an assumption of 120-130 proof bourbon—equating to a baseline of 60% alcohol by volume. Changing the

percent of alcohol within the solvent would drastically affect results, because the VOCs have different solubility values in water and ethanol. Assuming a volume of 125 nm^3, and incorporating the densities of water and ethanol, it was determined that the box would hold 1666 molecules of water and 773 molecules of ethanol (as well as one VOC molecule). This volume was chosen based on the size of particles in the system being on a 1-2 nm scale. Systems smaller than this would not provide a realistic picture of the behavior occurring, while systems larger than this, while more representative, would take greater computational power. A GROMACS topology file (.top) must be manually created and edited to match the number of solvent molecules with the VOC. Despite filling the box with the correct number of molecules that will fit in it, the box size must still be specified. Using the editconf command within GROMACS, the box size was set to 5x5x5 nm^3. GROMACS version 2020 was used for all simulations presented here.

After a structure and topology has been generated for the complete system, the next step involves minimizing the energy of the system. Due to imperfections in the system setup, it is possible to have starting configurations that cause forces to be too large due to overlapping van der Waals radii, causing spikes in velocity early in the simulation. Therefore, driving the system down potential energy gradients toward a minimum energy for the system allows for an optimal and stable starting point for the simulation.

Using the grompp function in GROMACS outputs a binary *.tpr* file containing the assembled structure starting point, the topology, and simulation parameters. Energy minimization simulations are generally quick. For the energy minimization, the simulation was set to stop once the maximum force in the system reached less than 1000 kJ/mol/nm. This is a generally recommended force value for ensuring an optimal starting

position for the actual MD simulations to run. The coordinates of the atoms in the system were shifted to drive the system's potential energy downhill using the steepest descent method for up to 50,000 steps, with an energy step size of 0.01 kJ/mol/nm. To minimize the number of potential and force calculations in the system, Lennard-Jones potentials and forces are truncated after a certain distance. This is called the cutoff distance, and it was set to 1.0 nm for all simulations presented here. The cutoff distance is selected to coincide with the asymptotic portion of the Lennard-Jones potential curve as shown in Figure 3. Short range electrostatic potentials and forces are likewise calculated directly below the cutoff distance. However, electrostatic interactions act over much longer distances than Lennard-Jones interactions, and therefore long-range electrostatic interactions must also be considered. The Particle-mesh Ewald (PME) method for calculating long-range electrostatics was employed. PME assigns charges to the grid using interpolation as opposed to direct summation of vectors. Transforming the grid into a 3D object using Fourier transformation, the reciprocal energy is calculated through one summation of the grid. The parameters of grid size are automatically tuned by GROMACS to maintain fast simulations with minimal numerical errors.

Because Lennard-Jones truncation causes a discontinuity in the potential, the potential is always shifted by a constant value such that it always equals 0 at the cutoff distance. Periodic boundary conditions are used for all simulations to prevent edge effects that would arise from surrounding the solvent box with a vacuum. Periodic boundary conditions essentially create an infinite working space in a "Pac-man"-like manner—where a particle traveling in the positive X direction will hit the edge of the cube and enter back into the cube from the opposite face. For the short-range electrostatics, the

cutoff distance becomes important and is chosen to be 1.0 nm. This cutoff distance is also the same for short-range Van der Waals interactions. As seen in Figure 2, as the distance from two particles increases, the potential slowly goes to 0. By choosing a cut-off of 1.0 nm, the assumption is made that there are no interactions occurring between particles past this distance. With all these parameters, and a successful energy minimization completed, the next step is equilibrating the system.

The position of the solvent molecules must be shifted from their random initial positions generated by Packmol toward a more realistic structure around the VOC. Without equilibration, certain MD simulations will become unstable and crash. Other simulations will produce unreliable results, because the sampling is of states that are far from equilibrium. Equilibration takes a longer amount of time than running energy minimization and can be run in the NPT or NVT ensemble (or both). The systems in this study were equilibrated under an NPT ensemble, run for 500,000 steps and a time step of 0.002 ps for a total of 1.0 ns of simulation time, which is enough time for solvent molecules to rearrange with respect to each other and with respect to the VOC.

Equilibration simulations are true MD simulations in that they use numerical integration of the Newtonian equations of motion to propagate the system coordinates through time, in contrast to the energy minimization simulations previously described, in which atom coordinates are simply driven down potential energy gradients. MD simulations in the NPT ensemble require the addition of a thermostat and barostat to the system. For the thermostat, the coupling scheme was set to V-rescale. With a thermostat, the temperature is controlled by adding or removing the kinetic energy in the system through changing atomic velocities. The setpoint was 300K with a τ value of 0.1 ps.

Using 2-methoxyphenol as an example, Figure 4 demonstrates the function of the thermostat in keeping the temperature maintained at an average of 300K. Despite fluctuations from ~292K up to ~308K, the overall moving average was maintained at 300K.

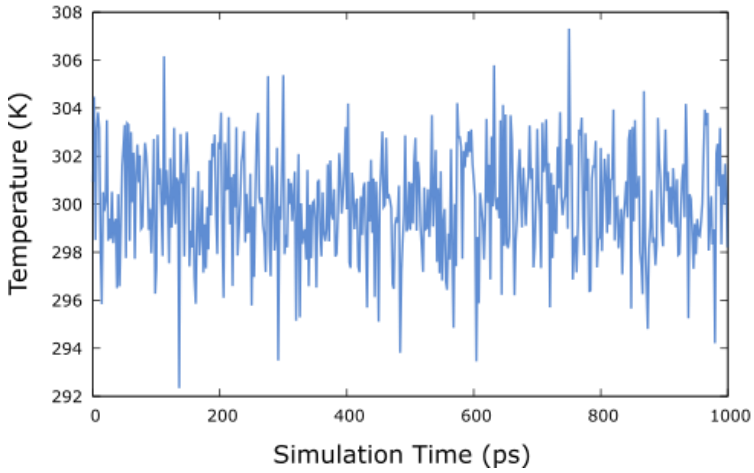


Figure 4. Temperature profile of 2-methoxyphenol during the NPT equilibration simulation

For the barostat, the pressure was set to a reference pressure of 1.0 bar with a τ value of 2.0 ps and compressibility of 4.5×10^{-5} bar^{-1}, which is the isothermal compressibility of water (since the solvent model is based largely on water). This is also a standard for use with GROMACS. The pressure coupling algorithm used was the Berendsen barostat, which is somewhat analogous to the temperature coupling scheme. The Berendsen barostat is an algorithm that holds the pressure constant through changing the box size of the simulation. It is important when using this algorithm that the coupling type is also set to be isotropic, so that every dimension of the box is changed equally as opposed to increasing or decreasing certain faces exclusively. The pressure profile of 2-methoxyphenol is shown in Figure 5.

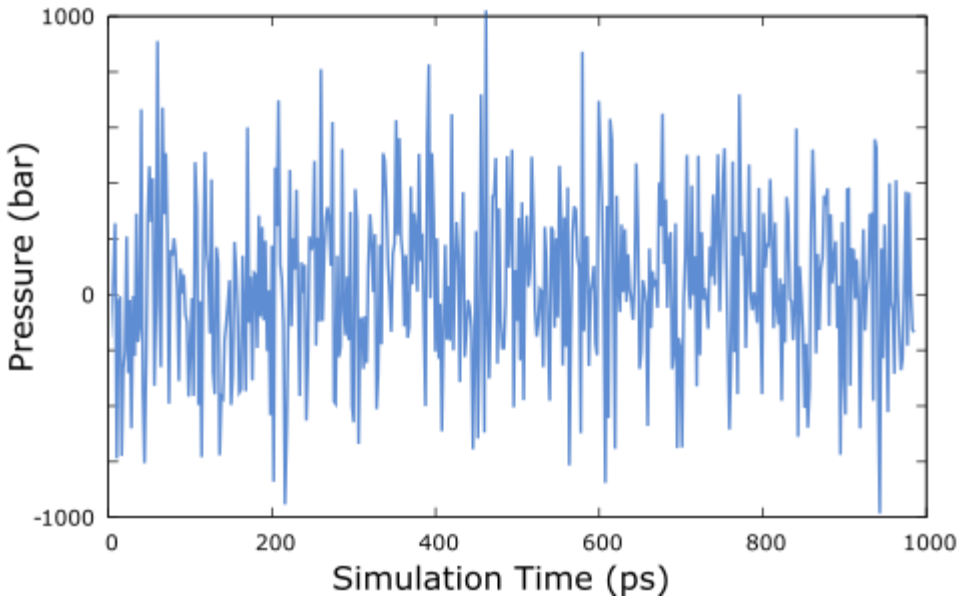


Figure 5. Pressure profile of 2-methoxyphenol during the NPT equilibration simulation.

At first glance, Figure 5 looks similar to the temperature profile presented in Figure 4, however, with further inspection of the axes the difference becomes apparent. The pressure fluctuated for 2-methoxyphenol over a range of 2000 bar which could seem alarming after effort was put in to minimizing the system. Fortunately, pressure tends to fluctuate heavily over an MD simulation of incompressible fluids, and this type of profile is very normal. While the setpoint was held at 1.0 bar, the actual average of the system was 6.04 bar for the equilibration of 2-methoxyphenol. However, the root-mean-square difference of the fluctuations was 369.16 bar, which is so large that, statistically, the value of 6.04 bar would be identical to 1.0 bar. Over longer timescales, an extended simulation's pressure would equilibrate to 1.0 bar eventually. So, while, at times, the system pressure dips down to values such as -678 bar at timestep 128 ps, these low points are balanced out by pressure highs of equal magnitudes resulting in a 'constant' pressure that is at the setpoint. Lastly, the density of the equilibrated, solvated system of

water/ethanol and 2-methoxyphenol was 855.71 kg/m^3 which is reasonably close to the density of 120-proof bourbon at 891.1 kg/m^3.

The equilibration step is also the first simulation run where velocities of the particles in the system are introduced. GROMACS generates atomic velocities at the beginning of the simulation based on the starting temperature setpoint of 300K and the Maxwell-Boltzmann distribution.

To check that the system is appropriately equilibrated, analysis was run prior to the production MD simulations. For all systems, the box size equilibrated to 127 nm^3, which was adequately close to the initial box size of 125 nm^3. The volume of the system was set to the initial 125 nm^3, then as the system is minimized and velocities are introduced, the system expands until reaching a maximum. After this, the system converges rapidly to a resting, equilibrated volume of 127 nm^3 for the rest of the simulation from ~100 ps onwards. This behavior is shown in Figure 6.

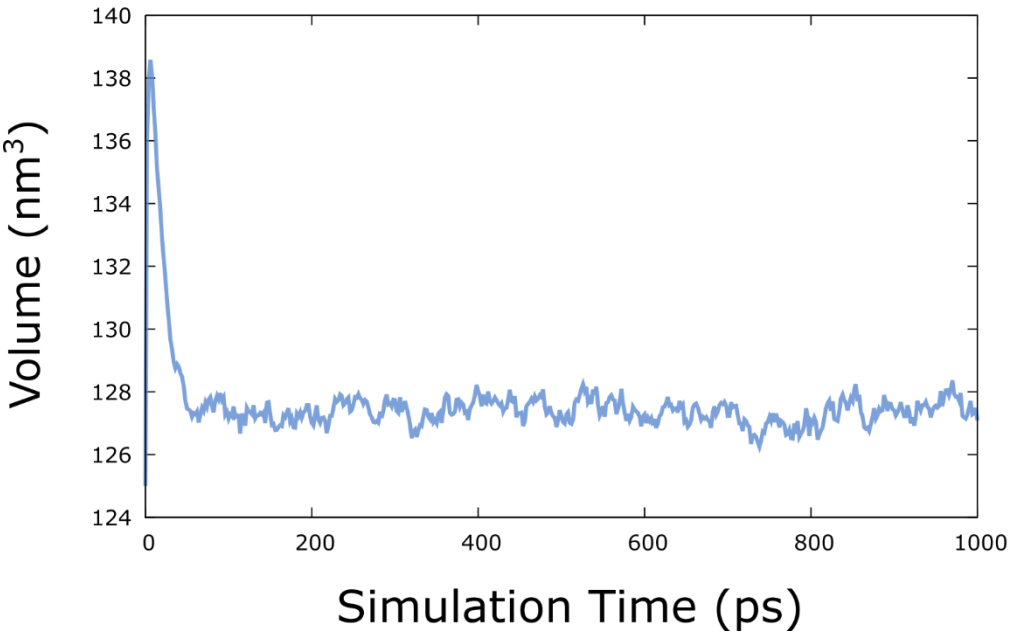


Figure 6. Volume profile of 2-methoxyphenol during the NPT equilibration simulation.

With a well-equilibrated system, the next step was running a production MD simulation for data collection. Before running the actual simulations for data collection, 3 test simulations were run at increasing simulation lengths (500,000 steps, 5,000,000 steps, and 50,000,000 steps corresponding to 1.0, 10, and 100 ns, respectively) to check for convergence over the period of the simulation. All 3 of these simulation lengths were returning similar values for the change in enthalpy of the system. Because of this, the shortest simulation length was chosen at 500,000 steps so that many simulations could be run for each of the 18 VOCs.

The production MD simulation parameters are the same for the equilibration simulations, aside from two exceptions. The first is the addition of the coupling parameter λ as described in the introduction section and visualized in Figure 1, specifically, the reverse ensemble where $\lambda = 1$ corresponds to state A where the VOC is fully present in the solvent system and slowly disappears to state B.

The second change was that instead of using the Berendsen barostat, the barostat was changed to use the Parrinello-Rahman algorithm. The Berendsen barostat is ideal for the equilibration step because it is a fast, first-order algorithm that will rapidly equilibrate a system. However, this method is not as reliable for a full-length production run simulation. The Parrinello-Rahman algorithm allows for maintaining the correct canonical or isothermal-isobaric ensemble and is more reliable for simulating thermodynamic properties. The downside to using this algorithm is that it has much slower approach to the setpoint pressure than using the Berendsen barostat due to its second-order nature.

Fifty transition simulations were run for the forward ensemble and fifty simulations were run for the reverse ensemble for each of the 18 VOCs. This took considerable computational effort and would not have been possible without a system for queueing simulations on the research cluster, as the production simulations, which remove major restraints on the system, take much longer to perform.

Once these production runs were completed, the change in Gibbs energy of solvation was calculated using a Python script pmx, created by Gapsys and de Groot, that analyzes the forward and reverse trajectories for each run, combining them into histograms[22]. This script also calculated the Gibbs free energy using the CGI, BAR, and Jarzynski methods. As previously described, the intercept of the two histograms is used to find the final value of Gibbs energy for the system. The Crook's Gaussian Intercept plots for 2-methoxyphenol can be seen in Figure 7. However, as was the case for most of the VOCs tested, the overlap of the two histograms was so close that instead of using the actual intercept of the two Gaussian fits, the Python script calculated the mean of the two data sets.

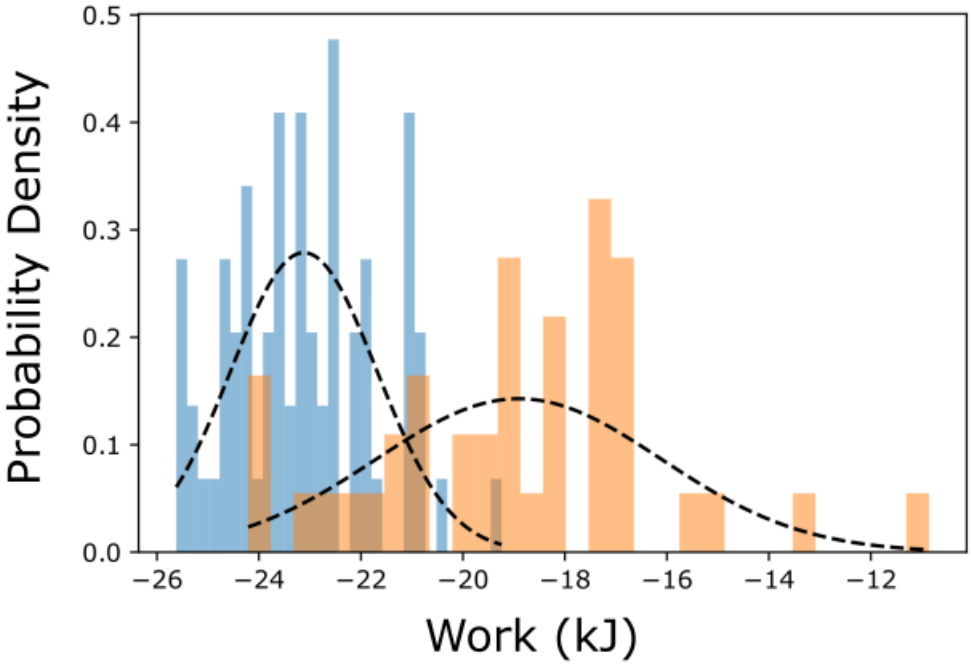


Figure 7. Crooks Gaussian Intercept plot for 2-methoxyphenol, demonstrating the intersection of the two Gaussian fits.

Following the production runs for each VOC in both the forward and reverse ensembles, the entire process had to be repeated for a VOC in a pure vacuum to complete the thermodynamic cycle of solvation. This is because the total change in Gibbs energy of solvation is the difference in solution energy and gas-phase energy. The thermodynamic cycle is seen in Figure 8.

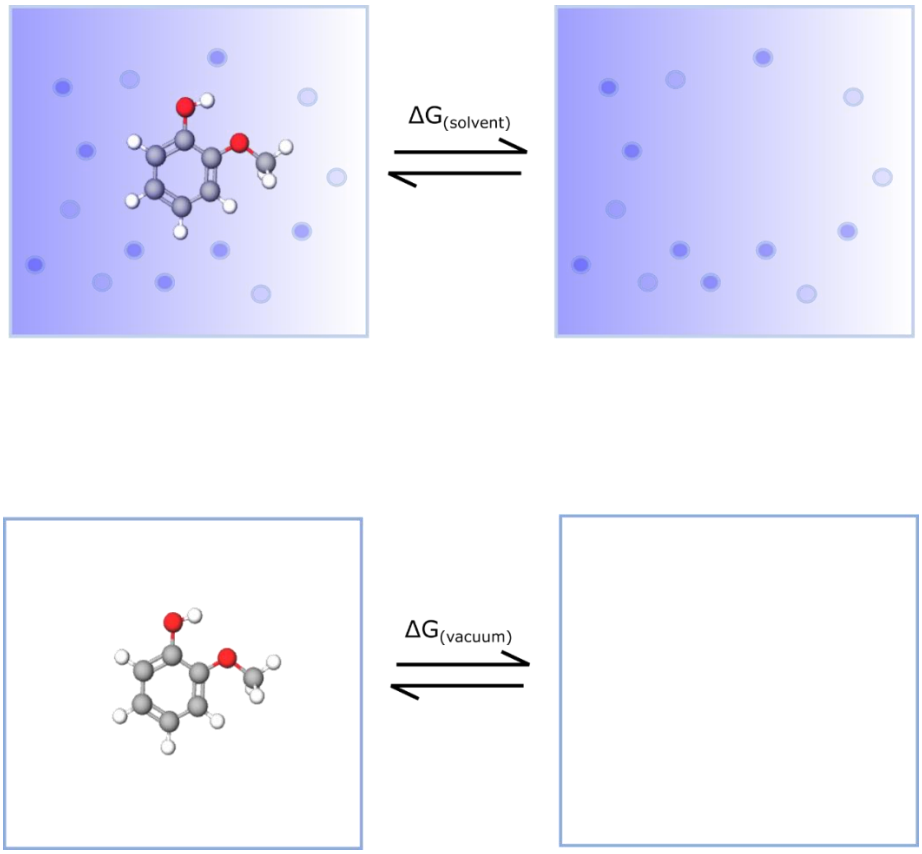


Figure 8. Thermodynamic cycle as it relates to the solvation of 2-methoxyphenol in a solvent system and in a vacuum state, where going from left to right is $\Delta G_{solvation}$.

The resultant change in Gibbs free energy of solvation is then used for the purposes of calculating the final Henry's law constant. While the chemical potential is usually defined by a system's internal energy, using a Legendre transformation on the definition of internal energy for Gibbs free energy yields the expression for chemical potential based on Gibbs free energy, where the chemical potential is simply the partial derivative of Gibbs free energy with respect to the number of moles of species i. Because of this relationship, the calculated Gibbs energy is used in place of μ_i.

Following this, the calculation of the Henry's Law constant is very straightforward using Equation 9:

$$\Delta G^{\circ}_{solv} = -RT\ ln\left(H_{cp}\frac{P^{\circ}}{m^{\circ}}\right) \quad (9)$$

Where, P° and m° are the reference pressure and molar standard. P° is chosen to be 1 bar, and m° is chosen to be 1 mol/kg.

Rearranging, and solving for H_{cp} yields Equation 10:

$$H_{cp} = \left(\frac{m^{\circ}}{P^{\circ}}\right)\exp\left(\frac{-\Delta G^{\circ}_{solv}}{RT}\right) \quad (10)$$

Where, after unit conversion, the standard units for H_{cp} are mol/m^{3}Pa.

Results and Discussion

With the culmination of data collection, data analysis could begin. Simulations were first checked for the convergence of enthalpy. This is due to the simple relation where the Gibbs free energy change is directly related to the enthalpy and entropy change: $\Delta G = \Delta H - T\Delta S$.

Using 2-methoxyphenol as a visual example to demonstrate the convergence, Figure 9 shows the change in enthalpy per change in λ vs. the simulation time. It is important that the enthalpy converges while the coupling parameter changes, so that the Gibbs free energy of solvation also converges. Integrating the dH/dλ results in the work value, and any inflections would indicate adverse system behavior.

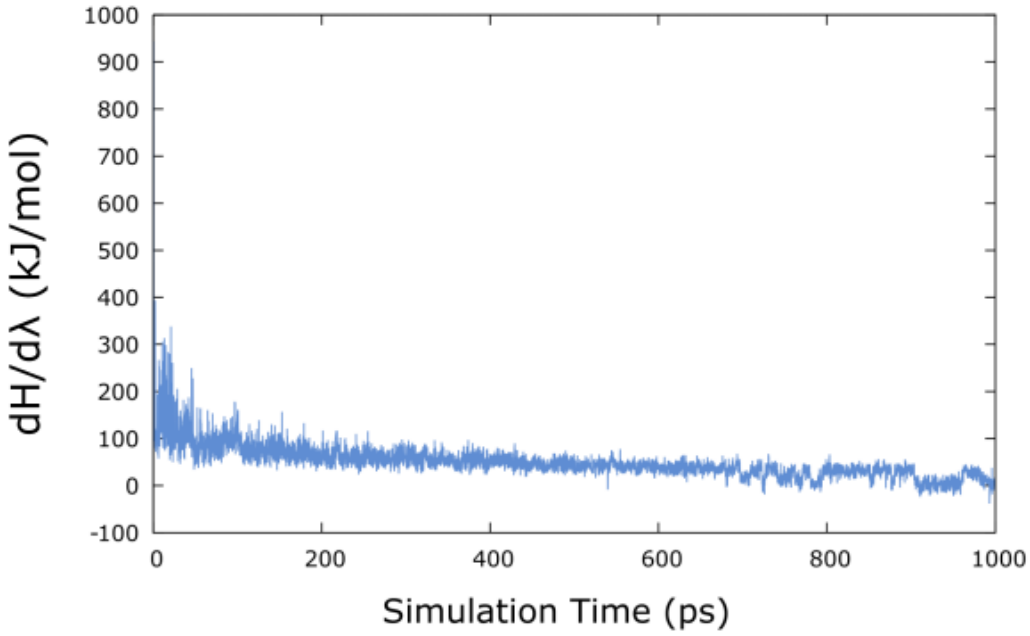


Figure. 9 2-methoxyphenol profile of change in Enthalpy over changing λ vs. time.

The model of a production simulation can be seen for 2-methoxyphenol in Figure 10. The box itself is the 125 nm^3 box, with periodic boundary conditions (which can be seen as some of the water and ethanol molecules are outside of the box lines). The water molecules are the red lines. The ethanol molecules are in light blue lines. In total there are 1666 water molecules and 773 ethanol molecules as previously mentioned. The VOC, 2-methoxyphenol, is represented using the van der Waals representation, so it is scaled up in size compared to the solvent molecules. In reality, this is a very tightly packed box filled completely with molecules. This can be seen when using the VDW representation for all molecules shown in Figure 11.

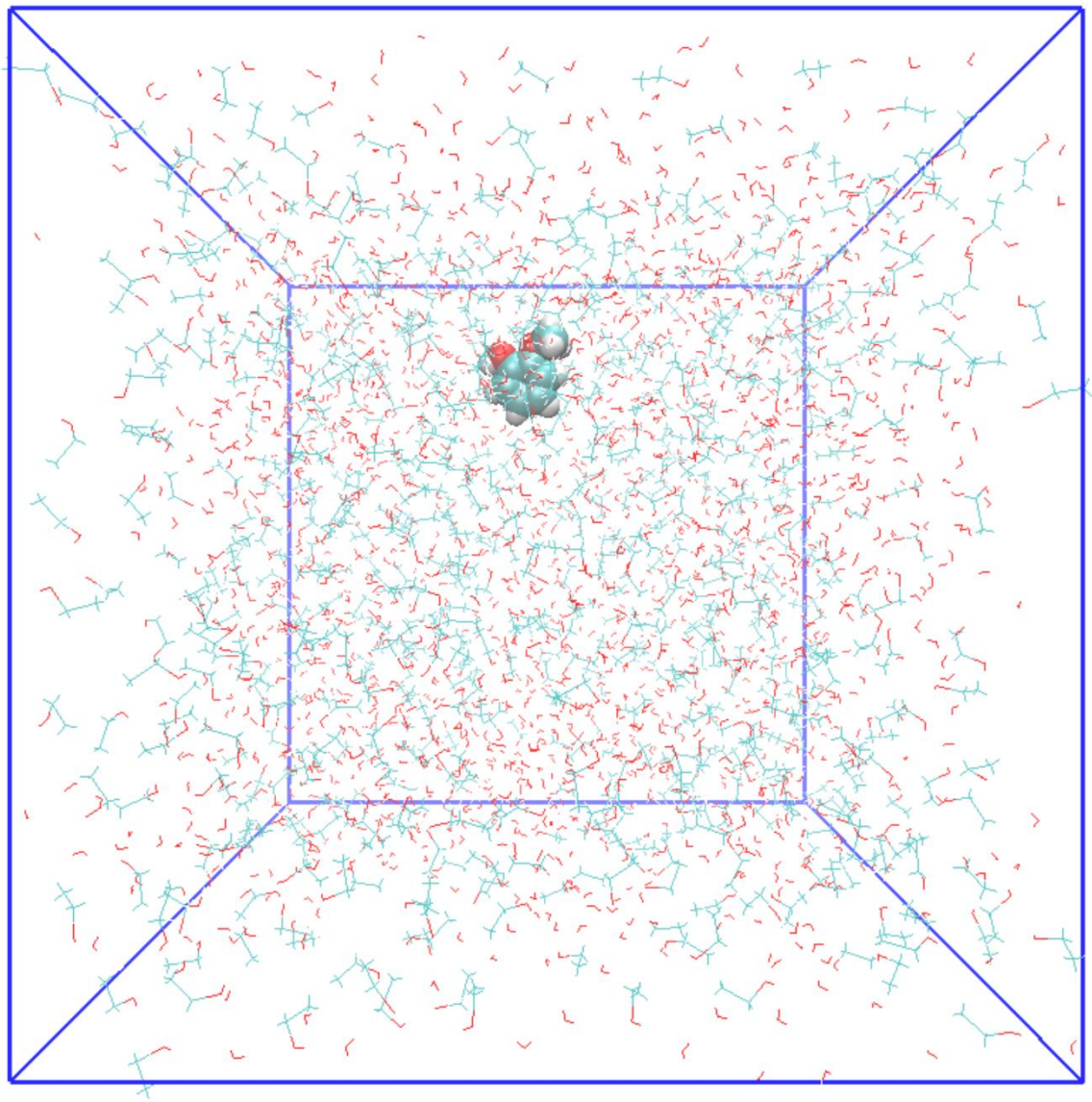


Figure 10. Snapshot of 2-methoxyphenol production run where the system is contained within the 125 nm^3 box, while periodic boundary conditions are implemented. Ethanol (light blue) and water (red) surround 2-methoxyphenol (VDW visualization).

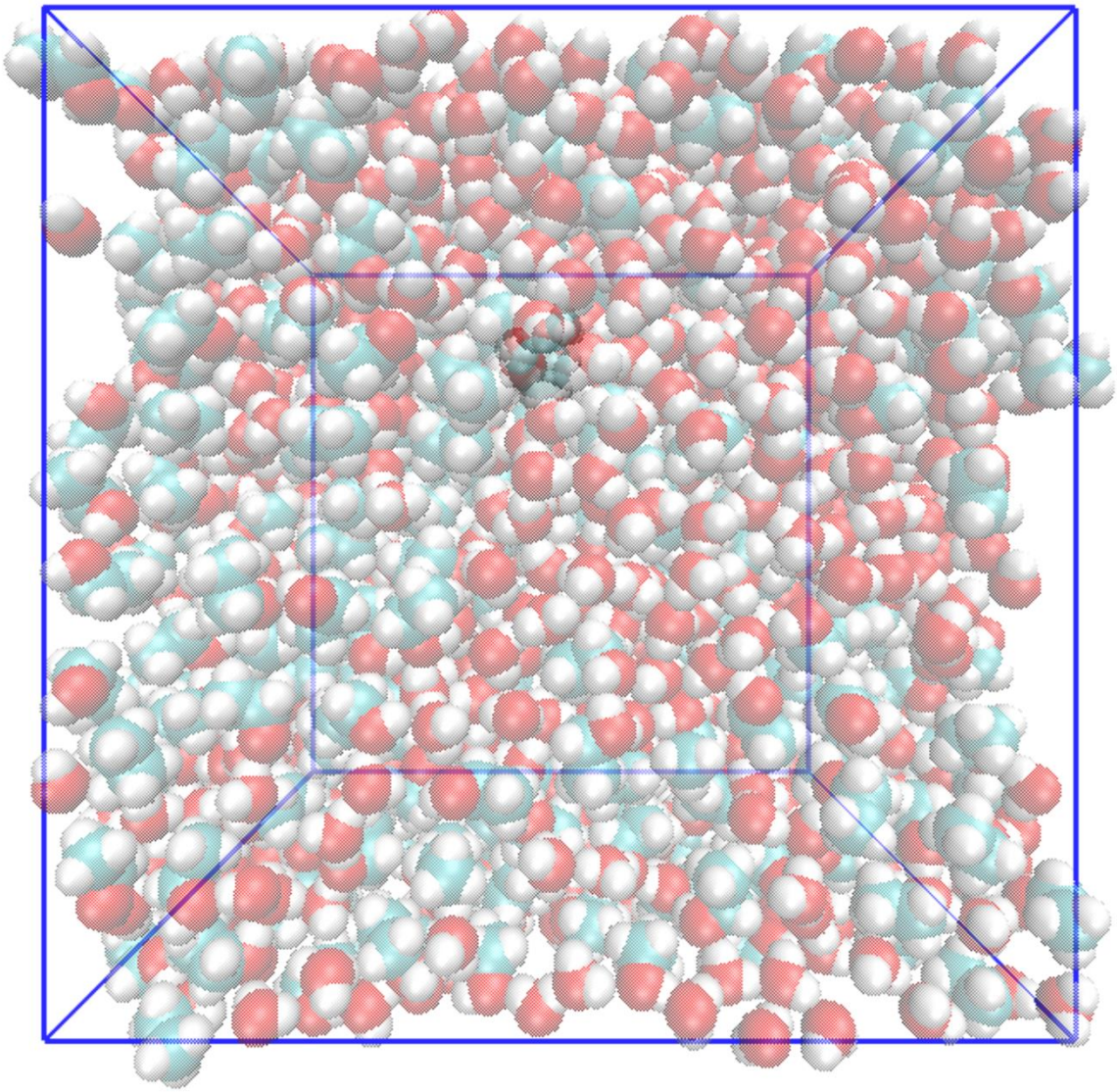


Figure 11. Snapshot of the same system in Figure 11 but using van der Waals visualizations for each molecule. Water and ethanol are transparent, while 2-methoxyphenol is centered and towards the top of the box.

The 18 volatile compounds chosen for this study are depicted using ball-and-stick 3D models in Figure 12.

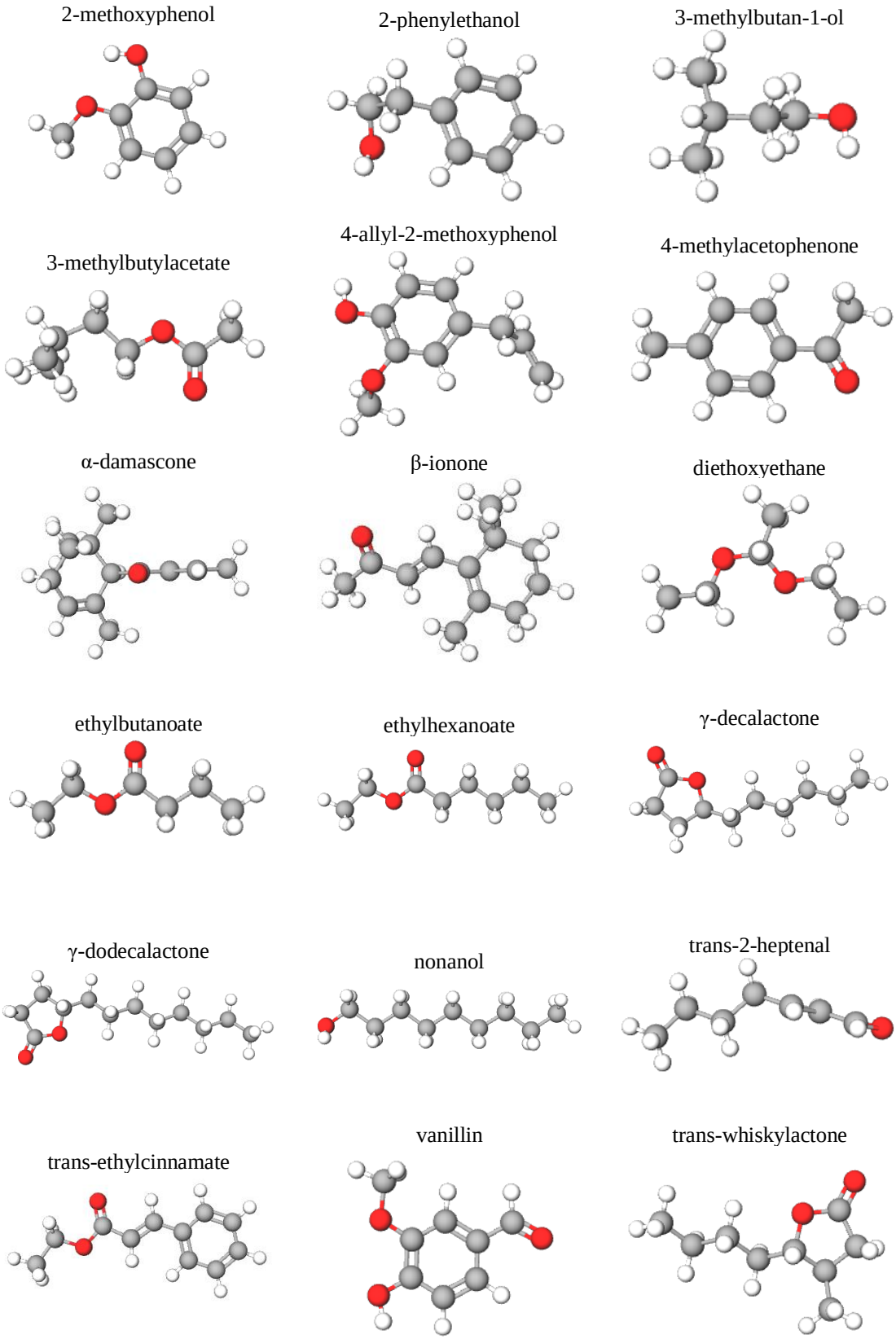


Figure 12. VOC Models

Using the previously mentioned Python script, pmx, change in Gibbs free energy of solvation was calculated using the forward and reverse trajectories for every solvated VOC using 3 different methods: Crooks-Gaussian Intercept, Bennett Acceptance Ratio, and Jarzynski's work averaging. These are shown in Table 1. The associated error for each method was the error from analytical integration, as opposed to bootstrap error.

Table 1

ΔG Values in Solvated System for Each Method.

VOC	CGI ΔG (kJ/mol)	BAR ΔG (kJ/mol)	Jarz ΔG (kJ/mol)	CGI Error (±)	BAR Error (±)	Jarz Error (±)
trans-2-heptenal	-19.41	-19.42	-19.44	0.12	0.09	0.11
nonanol	-20.4	-20.4	-20.42	0.13	0.09	0.1
2-phenylethanol	-22.67	-22.69	-22.64	0.14	0.12	0.22
3-methylbutan-1-ol	-26.28	-26.17	-26.16	0.12	0.08	0.13
4-allyl-2-methoxyphenol	-45.62	-45.66	-45.63	0.25	0.17	0.22
α-damascone	-46.81	-46.81	-46.8	0.13	0.12	0.17
2-methoxyphenol	-48.39	-48.28	-48.23	0.17	0.15	0.21
4-methylacetophenone	-57.27	-57.41	-57.43	0.17	0.11	0.14
γ-dodecalactone	-82.83	-82.92	-82.93	0.14	0.08	0.12
γ-decalactone	-83.25	-83.33	-83.34	0.15	0.12	0.14
t-whiskylactone	-86.9	-86.9	-86.85	0.16	0.11	0.17
ethylbutanoate	-94.73	-94.74	-94.78	0.11	0.1	0.15
ethylhexanoate	-95.11	-95.08	-95.08	0.18	0.11	0.15
t-ethylcinnamate	-106.93	-106.91	-106.94	0.18	0.12	0.15
diethoxyethane	-122.4	-122.63	-122.67	0.2	0.13	0.14
β-ionone	-131.9	-132.02	-132.05	0.14	0.11	0.11
3-methylbutylacetate	-139.12	-139.27	-139.32	0.14	0.11	0.14
vanillin	-33.47	-33.69	-33.82	0.23	0.25	0.27

Following the tabulation of Gibbs free energy change data for the solvated systems, as mentioned in methods, simulations were repeated using vacuum conditions with results shown in Table 2 for each method.

Table 2

ΔG Values Under Vacuum for Each Method

VOC	CGI ΔG (kJ/mol)	BAR ΔG (kJ/mol)	Jarz ΔG (kJ/mol)	CGI Error (±)	BAR Error (±)	Jarz Error (±)
trans-2-heptenal	10	11.74	11.48	0.19	0.39	0.47
nonanol	6.79	7.43	7.52	0.17	0.3	0.22
2-phenylethanol	-9.7	-9.59	-9.69	0.09	0.13	0.14
3-methylbutan-1-ol	-2.74	-3.66	-3.67	0.15	0.26	0.09
4-allyl-2-methoxyphenol	-19.46	-20.61	-21.19	0.78	0.51	0.83
α-damascone	-10.41	-10.57	-10.29	0.24	0.2	0.27
2-methoxyphenol	-21.11	-21.38	-21.51	0.18	0.27	0.45
4-methylacetophenone	-24.87	-23.08	-23.23	0.27	0.51	0.11
γ-dodecalactone	-45.07	-43.44	-43.26	0.3	0.61	0.36
γ-decalactone	-44.58	-42.88	-42.91	0.3	0.59	0.33
t-whiskylactone	-49.25	-46.89	-46.95	0.31	0.65	0.31
ethylbutanoate	-62.92	-62.19	-62.36	0.18	0.29	0.15
ethylhexanoate	-63.66	-61.86	-61.97	0.26	0.51	0.32
t-ethylcinnamate	-77.36	-75.59	-75.58	0.3	0.54	0.19
diethoxyethane	-99.04	-97.76	-97.9	0.18	0.31	0.17
β-ionone	-113.68	-113.68	-113.68	0	0	0
3-methylbutylacetate	-105.53	-105.76	-105.81	0.32	0.51	0.11
vanillin	13.94	15.07	15.12	0.47	0.81	0.28

Following the collection of both solvated system data and vacuum data, the vacuum Gibbs free energy data was subtracted from the solvated data. From this, the associated error for the Crook's Gaussian (CGI) method was added and subtracted to yield an upper and lower estimate for the change in Gibb's free energy of solvation. Using the formula for calculating Henry's law constants, the upper and lower CGI ΔG values were used to find the final resulting constants, seen in Table 3. The CGI method was used due to its associated Kolmogorov-Smirnov tests that check for Gaussian distribution quality.

Table 3

Henry's Law Constants Using Crook's Gaussian ΔG Data

VOC	CGI ΔG (J/mol)	CGI ΔG upper bound (J/mol)	CGI ΔG lower bound (J/mol)	H^{cp} (mol/m^3Pa)	H^{cp} lower bound (mol/m^3Pa)	H^{cp} upper bound (mol/m^3Pa)
trans-2-heptenal	-29410	-29290	-29530	1.304	1.243	1.368
nonanol	-27190	-27060	-27320	0.535	0.508	0.564
2-phenylethanol	-12970	-12830	-13110	0.002	0.002	0.002
3-methylbutan-1-ol	-23540	-23420	-23660	0.124	0.118	0.130
4-allyl-2-methoxyphenol	-26160	-25910	-26410	0.354	0.320	0.392
α-damascone	-36400	-36270	-36530	21.494	20.402	22.644
2-methoxyphenol	-27280	-27110	-27450	0.555	0.518	0.594
4-methylacetophenone	-32400	-32230	-32570	4.323	4.039	4.628
γ-dodecalactone	-37760	-37620	-37900	37.078	35.055	39.219
γ-decalactone	-38670	-38520	-38820	53.404	50.287	56.714
t-whiskylactone	-37650	-37490	-37810	35.479	33.274	37.829
ethylbutanoate	-31810	-31700	-31920	3.413	3.265	3.567
ethylhexanoate	-31450	-31270	-31630	2.954	2.748	3.175
t-ethylcinnamate	-29570	-29390	-29750	1.390	1.293	1.494
diethoxyethane	-23360	-23160	-23560	0.115	0.106	0.125
β-ionone	-18220	-18080	-18360	0.015	0.014	0.016
3-methylbutylacetate	-33590	-33450	-33730	6.967	6.587	7.369
vanillin	-47410	-47180	-47640	1775.807	1619.376	1947.349

Of the 14 calculated constants with associated literature data in aqueous solution, five compounds were on the same order of magnitude as the experimental constants. Three compounds were off by one order of magnitude compared to literature. The remaining 5 compounds deviated by two or more orders of magnitude from experimental values, with 2-phenylethanol deviating the most at a difference of five orders of magnitude. Research was conducted into the solubility of these compounds in both water and ethanol to try to better understand the data and the formation of any apparent trends

as VOC solubility in either water or ethanol relates to the Henry's law constant value. Each VOC, its experimental Henry's law constant, calculated Henry's law constant, solubility in water and ethanol, and aroma are tabulated in Table 4.

Table 4

Henry's Law Constants for Each VOC with Associated Upper and Lower Error Bounds

VOC	H^{cp} upper bound (mol/m^3Pa)	H^{cp} lower bound (mol/m^3Pa)	Experimental H^{cp} in water (mol/m^3Pa)	Experimental Reference	Solubility in H_2O	Solubility in Eth.	Aroma[5]
trans-2-heptenal	1.368	1.243	0.05	23	Insoluble	Soluble	Fatty, green
nonanol	0.564	0.508	0.11	24	140 mg/L	Miscible	soapy
2-phenylethanol	0.002	0.002	>37	25	16,000 mg/L	1 mL/2 mL in 50%	flowery
3-methylbutan-1-ol	0.130	0.118	0.46	23	26,700 mg/L	Miscible	malty
4-allyl-2-methoxyphenol	0.392	0.320	5.1	26	2400 mg/L	Soluble	Clove-like
α-damascone	22.644	20.402	None	none	Insoluble	1 mL/10 mL 95%	Cooked apple
2-methoxyphenol	0.594	0.518	7.7	27	187,000 mg/L	Very soluble	phenolic
4-methylacetophenone	4.628	4.039	1.1	28	Insoluble	Very soluble	Almond-like
γ-dodecalactone	39.219	35.055	none	none	Insoluble	1 mL/1 mL 95%	Peach-like
γ-decalactone	56.714	50.287	None	none	Conflicting	1 mL/ 1 mL	Peach-like
t-whiskylactone	37.829	33.274	None	none	Soluble	None	Coconut-like
ethylbutanoate	3.567	3.265	0.029	29	4900 mg/L	Miscible	fruity
ethylhexanoate	3.175	2.748	0.014	30	Insoluble	1 mL/2 mL 70%	fruity
t-ethylcinnamate	1.494	1.293	0.162	31	Insoluble	Miscible	fruity
diethoxyethane	0.125	0.106	0.1	26	44,000 mg/L	Miscible	fruity
β-ionone	0.016	0.014	1.2	32	169 mg/L	Miscible	Violet-like
3-methylbutylacetate	7.369	6.587	0.026	33	2000 mg/L	Miscible	fruity
vanillin	1947.349	1619.376	4700	26	11,020 mg/L	Very soluble	Vanilla-like

At first, these results seem somewhat inconclusive. While some VOCs are similar in value and magnitude to reported experimental values, most deviate by 1-2 orders of magnitude in terms of accuracy. However, standard error was calculated for each method of calculating the Gibbs free energy (3 measurements per compound) and this error ranged from 0.007 to 0.1 kJ/mol. This standard error was calculated using the standard deviation between the methods divided by the square root of the number of measurements. To investigate any correlation in the calculated data and the literature data, Spearman correlations were created to try to rank the VOCs based on the size and molecular weight of the molecules and how this possibly affects the trend of Henry's law constant value. The Spearman's rank correlation coefficient, ρ, is a metric for determining the rank correlation and relationship between two variables. The correlation investigated first was the relationship between Henry's law constants and the molecular weight of the VOCs. The values of each were ranked from smallest to largest for the constant vs. molecular weight, then using Equation 10, the rank correlation coefficient was calculated.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{10}$$

Where $\sum d_i^2$ is the summation of the difference between ranks for each variable squared. Because there were 18 VOCs, n = 18. The Spearman coefficient ranges from -1 to 1, where -1 indicates a perfectly negative linear correlation, 1 indicates a perfectly positive linear correlation, and 0 indicates no correlation at all. Ideally, the coefficient should be close to, but not 1.

When determining the correlation between the Henry's law constant and molecular weight, the calculated ρ was 0.385—which indicates a broad, but clustered

correlation between the two. Using a similar process, the correlation for the Henry's law constant and topological polar surface area was investigated. The calculated ρ for this case was 0.212, indicating a slight correlation between the two, but not as significant as the previous case.

Comparing these values to the coefficients obtained when instead using the literature Henry's law constant, first the relationship between molecular weight and constant was investigated. The calculated ρ for the experimental data for the first case was 0.292. The calculated ρ for the case between the constants and the topological polar surface area was 0.281. Using the experimental data provided similar results as with using the data determined from MD simulations, where the Spearman coefficients did not differ in magnitude. As previously discussed, calculating Henry's law constants for compounds in bourbon using molecular dynamics simulations has not been done previously. However, due to the relatively small error associated with the values of Gibbs free energy for each VOC, these results are very precise over many iterations.

To investigate accuracy and validate the proposed novel method of calculating Henry's Law constants using MD, one compound (2-phenylethanol) was used to repeat the process presented in the methods section under two new conditions: (1) using a pure aqueous solvent, and (2) using a pure ethanol solvent. For 2-phenylethanol in water, the Henry's law constant was calculated to be 0.12 mol/m^3-Pa, compared to a literature value of >37 mol/m^3-Pa. 2-phenylethanol in ethanol yielded a Henry's Law constant of 0.0003 mol/m^3-Pa.

Conclusion

This study provides valuable insight into the feasibility of predicting thermodynamic data for volatile compounds in solvent systems more complex than water and could prove to be a valuable addition for the spirits industry. The calculated Henry's law constants for VOCs in bourbon were very precise, with standard error between methods of calculating ΔG ranging from 0.007 to 0.1 kJ/mol for calculated energies ranging from -139 kJ/mol up to -18 kJ/mol. Continuing, five calculated Henry's Law constants were on the same order of magnitude as their literature values, three compounds were within one order of magnitude, and the remaining compounds deviated by two or more orders of magnitude, up to a maximum of five.

Using the same process laid out in the methods section, 2-phenylethanol was chosen to be solvated in two new systems—one with an ethanol solvent and the other water. In doing this, the compound's calculated data could be more closely compared with the data presented in literature. While the Henry's Law constant for aqueous solvation of 2-phenylethanol is presented in literature as >37 mol/m^3-Pa, the calculated value herein was 0.120 mol/m^3-Pa. When compared with the value obtained using the bourbon solvent simulation, 0.002 mol/m^3-Pa, the result from aqueous solvation improves in accuracy by two orders of magnitude. Furthermore, the calculated Henry's Law constant for 2-phenylethanol in the ethanol solvent system was 0.0003 mol/m^3-Pa. 2-phenylethanol's solubility in water is 16,000 mg/L, while its solubility in ethanol is 1 mL/2 mL in 50% ethanol. These results coupled with the literature solubility data build confidence in the proposed novel method, where a compound's solubility in water relative to ethanol is reflected in the simulation of solvation in bourbon.

The accuracy and feasibility of running these simulations for this kind of work is generally well-established. Even though MD requires learning new technical knowledge, programming skills, etc., I would argue that the benefits of using simulations far outweighs any potential cost—whether this is time investment, cost investment, burden of knowledge, etc.

Subsequently, with the growing popularity of artificial intelligence and machine learning for predicting and forecasting data trends, developing a machine learning model to do this kind of computational work could also be beneficial. This would require a similar amount of data points as the ones collected in this study, but with the potential to expand to a much greater extent than using multiple simulation iterations. Admittedly, this would also require a new set of skills and knowledge that is not often found in chemical engineering curriculum.

Lastly, I recommend that the relationship between water and ethanol solubility is further investigated as it relates specifically to the magnitude of the Henry's law constants for bourbon VOCs. Would it be possible to predict the relative magnitude of one constant based on its quantitative solubility in water and ethanol, or is it dependent on many more parameters? Could we also interpolate between calculated constants for a water solvent system and an ethanol solvent system to get a new constant of the combined solvent? The results of this study lean toward agreement with this line of thinking.

While the presented calculations are more precise than accurate, they can be valuable in predicting the relative magnitudes of thermodynamic data, as evidenced by the validation cases of 2-phenylethanol. Future work in this area should focus on further

validation of methods and force fields used for calculating Henry's Law constants of VOCs in bourbon.

REFERENCES

1. Sander, R., Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos. Chem. Phys* **2015,** *15* (8), 4399-4981.
2. Kornstein, B.; Luckett, J., The economic and fiscal impacts of the distilling industry in Kentucky. *Kentucky Agricultural Development Fund, October* **2014**.
3. Fang, C.; Du, H.; Jia, W.; Xu, Y., Compositional differences and similarities between typical Chinese baijiu and western liquor as revealed by mass spectrometry-based metabolomics. *Metabolites* **2019,** *9* (1), 2.
4. Capobiango, M.; Oliveira, E.; Cardeal, Z., Evaluation of Methods Used for the Analysis of Volatile Organic Compounds of Sugarcane (Cachaça) and Fruit Spirits. *Food Analytical Methods* **2012,** *6*.
5. Poisson, L.; Schieberle, P., Characterization of the most odor-active compounds in an American Bourbon whisky by application of the aroma extract dilution analysis. *Journal of agricultural and food chemistry* **2008,** *56* (14), 5813-9.
6. Dearden, J. C.; Schüürmann, G., Quantitative structure-property relationships for predicting henry's law constant from molecular structure. *Environmental Toxicology and Chemistry: An International Journal* **2003,** *22* (8), 1755-1770.
7. Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics* **1980,** *72* (4), 2384-2393.
8. Geng, H.; Chen, F.; Ye, J.; Jiang, F., Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins. *Computational and structural biotechnology journal* **2019,** *17*, 1162-1170.
9. Lau, D.; Jian, W.; Yu, Z.; Hui, D., Nano-engineering of construction materials using molecular dynamics simulations: Prospects and challenges. *Composites Part B: Engineering* **2018,** *143*, 282-291.
10. Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A., Predictions of hydration free energies from all-atom molecular dynamics simulations. *J Phys Chem B* **2009,** *113* (14), 4533-4537.
11. Mobley, D. L.; Guthrie, J. P., FreeSolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design* **2014,** *28* (7), 711-720.
12. Poisson, L.; Schieberle, P., Characterization of the Key Aroma Compounds in an American Bourbon Whisky by Quantitative Measurements, Aroma Recombination, and Omission Studies. *Journal of Agricultural and Food Chemistry* **2008,** *56* (14), 5820-5826.
13. SALO, P.; NYKÄNEN, L.; SUOMALAINEN, H., ODOR THRESHOLDS AND RELATIVE INTENSITIES OF VOLATILE AROMA COMPONENTS IN AN ARTIFICIAL BEVERAGE IMITATING WHISKY. *Journal of Food Science* **1972,** *37* (3), 394-398.
14. Liger-Belair, G.; Villaume, S., Losses of dissolved CO2 through the cork stopper during Champagne aging: toward a multiparameter modeling. *J Agric Food Chem* **2011,** *59* (8), 4051-6.
15. Bonhommeau, D. A.; Perret, A.; Nuzillard, J.-M.; Cilindre, C.; Cours, T.; Alijah, A.; Liger-Belair, G., Unveiling the Interplay Between Diffusing CO2 and Ethanol Molecules in Champagne Wines by Classical Molecular Dynamics and 13C NMR Spectroscopy. *The Journal of Physical Chemistry Letters* **2014,** *5* (24), 4232-4237.
16. Zwanzig, R. W., High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.

17. Goette, M.; Grubmüller, H., Accuracy and convergence of free energy differences calculated from nonequilibrium switching processes. *Journal of computational chemistry* **2009,** *30* (3), 447-456.
18. Jarzynski, C., Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters* **1997,** *78* (14), 2690-2693.
19. Vanommeslaeghe, K.; Guvench, O.; MacKerell, A. D., Jr., Molecular mechanics. *Curr Pharm Des* **2014,** *20* (20), 3281-3292.
20. Sousa da Silva, A. W.; Vranken, W. F., ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Research Notes* **2012,** *5* (1), 367.
21. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J Comput Chem* **2004,** *25* (9), 1157-74.
22. Gapsys, V.; de Groot, B. L., pmx Webserver: a user friendly interface for alchemistry. *Journal of chemical information and modeling* **2017,** *57* (2), 109-114.
23. Hilal, S. H.; Ayyampalayam, S. N.; Carreira, L. A., Air-liquid partition coefficient for a diverse set of organic compounds: Henry's Law Constant in water and hexadecane. *Environ Sci Technol* **2008,** *42* (24), 9231-6.
24. Shunthirasingham, C.; Cao, X.; Lei, Y. D.; Wania, F., Large Bubbles Reduce the Surface Sorption Artifact of the Inert Gas Stripping Method. *Journal of Chemical & Engineering Data* **2013,** *58* (3), 792-797.
25. Altschuh, J.; Brüggemann, R.; Santl, H.; Eichinger, G.; Piringer, O. G., Henry's law constants for a diverse set of organic chemicals: Experimental determination and comparison of estimation methods. *Chemosphere* **1999,** *39* (11), 1871-1887.
26. TOXicology data NET-work (TOXNET). In *HSDB: Hazardous Substances Data Bank*, National Library of Medicine (US), 2015.
27. Sagebiel, J. C.; Seiber, J. N.; Woodrow, J. E., Comparison of headspace and gas-stripping methods for determining the Henry's law constant (H) for organic compounds of low to intermediate H. *Chemosphere* **1992,** *25* (12), 1763-1768.
28. Abraham, M. H.; Andonian-Haftvan, J.; Whiting, G. S.; Leo, A.; Taft, R. S., Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. *Journal of the Chemical Society, Perkin Transactions 2* **1994,** (8), 1777-1791.
29. Fenclová, D.; Blahut, A.; Vrbka, P.; Dohnal, V.; Böhme, A., Temperature dependence of limiting activity coefficients, Henry's law constants, and related infinite dilution properties of C4–C6 isomeric n-alkyl ethanoates/ethyl n-alkanoates in water. Measurement, critical compilation, correlation, and recommended data. *Fluid Phase Equilibria* **2014,** *375*, 347-359.
30. Aprea, E.; Biasioli, F.; Märk, T. D.; Gasperi, F., PTR-MS study of esters in water and water/ethanol solutions: Fragmentation patterns and partition coefficients. *International Journal of Mass Spectrometry* **2007,** *262* (1), 114-121.
31. (ECHA), E. C. A.
32. Fichan, I.; Larroche, C.; Gros, J. B., Water Solubility, Vapor Pressure, and Activity Coefficients of Terpenes and Terpenoids. *Journal of Chemical & Engineering Data* **1999,** *44* (1), 56-62.
33. Mackay, D.; Shiu, W.-Y.; Lee, S. C., *Handbook of physical-chemical properties and environmental fate for organic chemicals*. CRC press: 2006.

CURRICULUM VITA

NAME: Chris Abney

ADDRESS: Department of Chemical Engineering
216 Eastern Pwky.
University of Louisville
Louisville, KY 40208

DOB: Louisville, Kentucky – July 28, 1995

EDUCATION
& TRAINING: B.S., Chemical Engineering
University of Louisville
2015-2020

M.Eng., Chemical Engineering
University of Louisville
2020-21

AWARDS: D.A. Richards/GE Memorial Scholarship
2020-21

PROFESSIONAL SOCIETIES: AiCHE