Journal of Visualized Experiments

Computation of Atmospheric Concentrations of Molecular Clusters from ab initio Thermochemistry

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE60964R1
Full Title:	Computation of Atmospheric Concentrations of Molecular Clusters from ab initio Thermochemistry
Section/Category:	JoVE Chemistry
Keywords:	Aerosols, atmospheric chemistry, amino acids, thermochemistry, ab initio, free energy, conformational analysis, potential energy surface, quantum chemistry, computational chemistry, cluster distribution, configurational sampling
Corresponding Author:	George Shields UNITED STATES
Corresponding Author's Institution:	
Corresponding Author E-Mail:	george.shields@furman.edu
Order of Authors:	Tuguldur Odbadrakh
	Ariel Gale
	Benjamin Ball
	Berhane Temelso
	George Shields
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US$4,200)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Greenville, SC, USA

Manuscript

Click here to access/download;Manuscript;60964_R1.docx

JOURNAL OF VISUALIZED EXPERIMENTS
1 Alewife Center #200
Cambridge, MA 02140
tel. 617.945.9051
www.jove.com

Standard Manuscript Template
Please Remove all Gray Text before Submitting

TITLE:

Computation of Atmospheric Concentrations of Molecular Clusters from ab initio Thermochemistry

AUTHORS AND AFFILIATIONS:

Tuguldur T. Odbadrakh[1], Ariel G. Gale[1], Benjamin T. Ball[1], Berhane Temelso[2], George C. Shields[1]

[1]Department of Chemistry, Furman University, Greenville, SC, USA
[2]College of Charleston, Charleston, SC, USA

Corresponding Authors:
George C. Shields
george.shields@furman.edu

Berhane Temelso
temelsob@cofc.edu

Email Addresses of Co-authors:

Tuguldur T. Odbadrakh	(togo.odbadrakh@furman.edu)
Ariel G. Gale	(ariel.gale@furman.edu)
Benjamin T. Ball	(tyler.ball@furman.edu)
Berhane Temelso	(temelsob@cofc.edu)

KEYWORDS:

Quantum chemistry, ab initio, thermochemistry, atmospheric chemistry, computational chemistry, aerosols, cluster distribution, configurational sampling

SUMMARY:

The atmospheric concentrations of weakly bound molecular clusters can be computed from the thermochemical properties of low energy structures found through a multi-step configurational sampling methodology utilizing a genetic algorithm and semi-empirical and ab initio quantum chemistry.

ABSTRACT:

The computational study of the formation and growth of atmospheric aerosols requires an accurate Gibbs free energy surface, which can be obtained from gas phase electronic structures and vibrational frequencies. These quantities are valid for those atmospheric clusters whose geometries correspond to a minimum on their potential energy surfaces. The Gibbs free energy of the minimum energy structure can be used to predict atmospheric concentrations of the cluster under a variety of conditions such as temperature and pressure. We present a computationally inexpensive procedure built on a genetic algorithm-based configurational sampling followed by a series of increasingly accurate screening calculations. The procedure starts by generating and evolving the geometries of a large set of configurations using semi-empirical models then refines the resulting unique structures at a series of high-level ab initio

levels of theory. Finally, thermodynamic corrections are computed for the resulting set of minimum-energy structures and used to compute the Gibbs free energies of formation, equilibrium constants, and atmospheric concentrations. We present the application of this procedure to the study of hydrated glycine clusters under ambient conditions.

INTRODUCTION:

The most uncertain parameter in atmospheric studies of climate change is the exact extent to which cloud particles reflect incoming solar radiation. Aerosols, which are particulate matter suspended in a gas, form cloud particles called cloud condensation nuclei (CCN) that scatter incoming radiation, thus preventing its absorption and the subsequent heating of the atmosphere[1]. A detailed understanding of this net cooling effect requires an understanding of the growth of aerosols into CCNs, which in turn requires an understanding of the growth of small molecular clusters into aerosol particles. Recent work has suggested that aerosol formation is initiated by molecular clusters of 3 nm in diameter or less[2]; however, this size regime is difficult to access using experimental techniques[3,4]. Therefore, a computational modeling approach is desired in order to overcome this experimental limitation.

Using our modeling approach described below, we can analyze the growth of any hydrated cluster. Because we are interested in the role of water in the formation of large biological molecules from smaller constituents in pre-biotic environments, we illustrate our approach with glycine. The challenges encountered and tools needed to address those research questions are very similar to those involved in the study of atmospheric aerosols and prenucleation clusters[5-15]. Here, we examine hydrated glycine clusters starting from an isolated glycine molecule followed by a series of stepwise additions of up to five water molecules. The final goal is to calculate the equilibrium concentrations of $Gly(H_2O)_{n=0-5}$ clusters in the atmosphere at room temperature at sea-level and a relative humidity (RH) of 100 %.

A small number of these subnanometer molecular clusters grow into a metastable critical cluster (1-3 nm in diameter) either by adding other vapor molecules or coagulating on existing clusters. These critical clusters have a favorable growth profile leading to the formation of much larger (up to 50-100 nm) cloud condensation nuclei (CCN), which directly affect the precipitation efficiency of clouds as well their ability to reflect incident light. Therefore, having a good understanding of the thermodynamics of molecular clusters and their equilibrium distributions should lead to more accurate predictions of the impact of aerosols on the global climate.

A descriptive model of aerosol formation requires accurate thermodynamics of molecular cluster formation. The computation of accurate thermodynamics of molecular cluster formation requires the identification of the most stable configurations, which involves finding the global and local minima on the cluster's potential energy surface (PES)[16]. This process is called configurational sampling and can be achieved through a variety of techniques, including those based on molecular dynamics (MD)[17-20], Monte Carlo (MC)[21,22], and genetic algorithms (GA)[23-25].

Different protocols have been developed over the years to obtain the structure and thermodynamics of atmospheric hydrates at a high level of theory. These protocols differed in the choice of (i) configurational sampling method, (ii) nature of low-level method used in the configurational sampling, and (iii) the hierarchy of higher level methods used to refine the results in the subsequent steps.

The configurational sampling methods included chemical intuition[26], random sampling[27,28], molecular dynamics (MD)[29,30], basin hopping (BH)[31], and genetic algorithm (GA)[24,25,32]. The most common low-level methods employed with these sampling methods are force fields or semi-empirical models such as PM6, PM7 and SCC-DFTB. These are often followed by DFT calculations with increasingly larger basis sets and more reliable functionals from the higher rungs of Jacob's ladder[33]. In some cases, these are followed by higher level wavefunction methods such as MP2, CCSD(T), and the cost efficient DLPNO-CCSD(T)[34,35].

Kildgaard et al.[36] developed a systematic method where water molecules are added at points on the Fibonacci spheres[37] around smaller hydrated or unhydrated clusters to generate candidates for larger clusters. Unphysical and redundant candidates are removed based on close contact thresholds and root-mean-square distance between different conformers. Subsequent optimizations using the PM6 semi-empirical method and a hierarchy of DFT and wavefunction methods are used to get a set of low energy conformers at a high level of theory. The artificial bee colony (ABC) algorithm[38] is a new configurational sampling approach that has recently been implemented by Zhang et al. to study molecular clusters in a program called ABCluster[39]. Kubecka et al.[40] used ABCluster for configurational sampling followed by low-level reoptimizations using the tight-binding GFN-xTB semi-empirical method[41]. They further refined the structures and energies using DFT methods followed by final energies using DLPNO-CCSD(T).

Regardless of the method, configurational sampling starts with a randomly- or nonrandomly-generated distribution of points on the PES. Each point corresponds to a specific geometry of the molecular cluster in question and is generated by the sampling method. Then the closest local minimum is found for each point by following the "downhill" direction on the PES. The set of minima thus found correspond to those geometries of the molecular cluster that are stable, at least for some time. Here, the shape of the PES and the evaluation of the energy at each point on the surface will be sensitive to the physical description of the system where a more accurate physical description results in a more computationally expensive energy calculation. We will specifically use the GA method implemented in the OGOLEM[25] program, which has been successfully applied to a variety of global optimization and configurational sampling problems[42-45], to generate the initial set of sampling points. The PES will be described by the PM7 model[46] implemented in the MOPAC2016 program[47]. This combination is employed because it generates a larger variety of points compared to the MD and MC methods and finds the local minima faster than more-detailed descriptions of the PES.

The set of GA-optimized local minima are taken as the starting geometries for a series of screening steps, which lead to a set of low lying minimum energy. This part of the protocol begins by optimizing the set of unique GA-optimized structures using density-functional theory (DFT) with a small basis set. This set of optimizations will generally give a smaller set of unique local minimum structures which are modeled in more detail compared to the GA-optimized semi-empirical structures. Then another round of DFT optimizations are performed on this smaller set of structures using a larger basis set. Again, this step will generally give a smaller set of unique structures which are modeled in more detail compared to the small basis DFT step. The final set of unique structures are then optimized to a tighter convergence and the harmonic vibrational frequencies are calculated. After this step we have everything we need to compute the equilibrium concentrations of the clusters in the atmosphere. The overall approach is summarized diagrammatically in **Figure 1**. We will use the PW91[48] generalized-gradient approximation (GGA) exchange-correlation functional in the Gaussian09[49] implementation of DFT along with two variations of the Pople[50] basis set (6-31+G* for the small basis step and 6-311++G** for the large basis step). This particular combination of exchange-correlation functional and basis set was chosen due to its previous success in computing accurate Gibbs free energies of formation for atmospheric clusters[51,52].

This protocol assumes that the user has access to a high-performance computing cluster with the PBS portable batch system[53], MOPAC2016 (http://openmopac.net/MOPAC2016.html)[47], OGOLEM (https://www.ogolem.org)[25], Gaussian 09 (https://gaussian.com)[49], and OpenBabel[54] (http://openbabel.org/wiki/Main_Page) software installed following their specific installation instructions. Each step in this protocol also uses a set of in-house shell and Python 2.7 scripts which must be saved to a directory that is included in the user's $PATH environmental variable. All necessary environmental modules and execution permissions to run all of the above programs must also be loaded into the user's session.

PROTOCOL:

1. Finding the minimum energy structure of isolated glycine and water

NOTE: The goal here is two-fold: (i) to obtain minimum energy structures of isolated water and glycine molecules for use in the genetic algorithm configurational sampling, (ii) and to compute the thermodynamic corrections to the gas phase energies of these molecules for use in the calculation of atmospheric concentrations.

1.1. Open a new session in Avogadro. Click **Build > Insert > Peptide > Gly > Insert Peptide** to generate a glycine monomer in the visualization window. Click **Extensions > Gaussian** and edit the first line in the text box to read '# pw91pw91/6-311++G** int(Acc2E=12,UltraFine) scf(conver=12) opt(tight,maxcyc=300) freq'.

1.1.1. Click **Generate** and save the command file as **glycine.com**. Please note that if the molecule has significant conformational flexibility, as glycine does[55], it is critical to perform conformational analysis to identify the global minimum structure and other low-lying conformers. OpenBabel[54] provides robust conformational search tools utilizing different algorithms and quick force fields. While conformers are allowed to relax and interconvert during GA and subsequent calculations, it is sometimes necessary to run multiple GA calculations, each starting with a different conformer.

1.2. Open a new session in Avogadro. Click the first button in the toolbar and choose **Element: Oxygen** in the Drawing Settings box while making sure **Adjust Hydrogens** is checked in. Click **Extensions > Gaussian** and edit the first line in the text box to read '# pw91pw91/6-311++G** int(Acc2E=12,UltraFine) scf(conver=12) opt(tight,maxcyc=300) freq'. Click **Generate** and save the command file as **water.com**.

1.3. Transfer the two **.com** files to the computing cluster and call Gaussian 09 in a submit script to start the calculation. Once the calculations finish, generate **.xyz** files of the minimum energy structures by calling OpenBabel. For glycine, the command-line command is '**obabel -ig09 glycine.log -oxyz > glycine.xyz**'. These two **.xyz** files will be used by the GA configurational sampling in the next step.

2. Genetic-algorithm-based configurational sampling of $Gly(H_2O)_{n=1-5}$ clusters

NOTE: The goal here is to obtain a set of low-energy structures for $Gly(H_2O)_{n=1-5}$ at the inexpensive semi-empirical level of theory, using the PM7[46] model implemented in MOPAC[47]. It is imperative that the working directory has the exact organization and structure as shown in **Figure 2**. This is to ensure that the custom shell and Python scripts work without failures.

2.1. Create a directory called **gly-h2o-n** where **n** is the number of water molecules. Create a subdirectory called **GA** under the **gly-h2o-n** directory to run generic algorithm calculations. Copy the OGOLEM input files (*Eg.* pm7.ogo), monomers Cartesian coordinates (***Eg.*** glycine.xyz, water.xyz) and PBS batch submission script (**Eg.** run.pbs) into the **GA** directory. When submitting the calculation, OGOLEM will create a new directory named as the prefix of the OGOLEM input file (Eg. pm7) in the GA directory and store newly generated coordinates there.

2.2. Once the calculation is complete, change directory to **gly-h2o-n/GA/pm7** and compute the rotational constants of the GA-optimized clusters with the command '**getRotConsts-GA.csh N 0 99**' where **N** is the number of atoms in the molecular cluster. This will generate a file called **rotConstsData_C** which contains a sorted list of all the GA-optimized cluster configurations, their energies, and their rotational constants.

2.3. Find and save the unique GA-optimized clusters with the command '**similarityAnalysis.py pm7 rotConstsData_C**' where **pm7** will be used as a file-naming label. This will generate a file called **uniqueStructures-pm7.data** which contains a sorted list of the unique GA-optimized configurations. This is a list of unique local minimum structures for the $Gly(H_2O)_n$ cluster optimized at the PM7 level of theory, and these structures are now ready to be refined using DFT.

2.4. Go up to the **gly-h2o-n/GA** directory and combine the results from multiple comparable GA runs using the **combineGA.csh** script. The syntax is **'combineGA.csh <label> <list of**

directories with GA runs>'. In this particular case, the command **'combineGA.csh pm7 pm7**' will generate a new unique structures list named '**uniqueStructures-pm7.data'** in the **gly-h2o-n/GA** directory.

3. Configurational sampling at the small-basis ab initio level of theory

NOTE: The goal here is to refine the configurational sampling of the Gly(H_2O)$_{n=1-5}$ clusters using a better quantum-mechanical description to obtain a smaller but more accurate set of Gly(H_2O)$_{n=1-5}$ cluster structures. The starting structures for this step are the outputs of Step 2.

3.1. Create a subdirectory called **QM** under the **gly-h2o-n** directory. Under the QM directory, create another subdirectory named **pw91-sb**. Copy the unique structures list (**uniqueStructures-pm7.data**) from the **gly-h2o-n/GA** directory to the **QM/pw91-sb** directory and change directory to that **gly-h2o-n/QM/pw91-sb**.

3.1.1. Run the small-basis DFT configurational sampling script with the command '**run-pw91-sb.csh uniqueStructures-pm7.data sb QUEUE 10**' where **QUEUE** is the preferred queue on the computing cluster. This script will automatically generate the inputs for Gaussian 09 and submit all the calculations. Enter '**test'** for the '**QUEUE'** to do a dry run.

3.2. Once the submitted calculations are complete, extract the energies and compute the rotational constants of the small-basis-optimized clusters with the command '**getRotConsts-dft-sb.csh pw91 N**' where **N** is the number of atoms in the cluster. That will create a file named **rotConstsData_C**. Now identify the unique structures with the command '**similarityAnalysis.py sb rotConstsData_C**'. There will now be a list of unique configurations optimized at the PW91/6-31+G* level of theory saved in the file **uniqueStructures-sb.data**.

3.3. Go up to the **gly-h2o-n/QM** directory and combine the results from multiple comparable QM runs using the **combineQM.csh** script. The syntax is **'combineQM.csh <label> <list of directories with QM calculations>**'. In this particular case, the command **'combineQM.csh sb pw91-sb**' will generate a new unique structures list named '**uniqueStructures-sb.data'** in the **gly-h2o-n/QM** directory.

4. Configurational sampling at the large-basis ab initio level of theory

NOTE: The goal here is to further refine our configurational sampling of the Gly(H_2O)$_{n=1-5}$ clusters using a better quantum-mechanical description. The starting structures for this step are the outputs of Step 3.

4.1. Create a subdirectory called **pw91-lb** under the **QM** directory. Copy the unique structures list (**uniqueStructures-sb.data**) from the **QM/pw91-sb** directory to the **QM/pw91-lb** directory and change directory to **QM/pw91-lb**. Run the large-basis DFT configurational sampling script with the command '**run-pw91-lb.csh uniqueStructures-sb.data lb QUEUE 10**' where **QUEUE** is the preferred queue on the computing cluster. This script will automatically generate the inputs for Gaussian 09 and submit all the calculations. Enter ‘**test’** for the ‘**QUEUE’** to do a dry run testing.

4.2. Once the submitted calculations are complete, compute the rotational constants of the large-basis-optimized clusters with the command ‘**getRotConsts-dft-lb.csh pw91 N**’ where **N** is the number of atoms in the cluster. Now identify the unique structures with the command ‘**similarityAnalysis.py lb rotConstsData_C**’. We now have a list of unique configurations optimized at the PW91/6-311++G** level of theory saved in the file **uniqueStructures-lb.data**.

5. Vibrational analysis

NOTE: The goal here is to obtain the vibrational structure and energies of the Gly(H_2O)$_{n=1-5}$ clusters in order to compute the desired thermochemical corrections.

5.1. Copy the unique structures list (**uniqueStructures-lb.data**) from the **QM/pw91-lb** directory to the **QM/pw91-lb/ultrafine** directory and change directory to **QM/pw91-lb/ultrafine**. Run the ultrafine large-basis DFT script with the command ‘**run-pw91-lb-ultrafine.csh uniqueStructures-lb.data uf QUEUE 10**’ where **QUEUE** is the preferred queue on the computing cluster. This script will automatically generate the inputs for Gaussian 09 and submit all the calculations. Enter ‘**test’** for the ‘**QUEUE’** to do a dry run testing.

5.2. Once the submitted calculations are complete, compute the rotational constants of the large-basis-optimized clusters with the command ‘**getRotConsts-dft-lb-ultrafine.csh pw91 N**’ where **N** is the number of atoms in the cluster. Now identify the unique structures with the command ‘**similarityAnalysis.py uf rotConstsData_C**’. We now have a list of unique configurations optimized at the PW91/6-311++G** level of theory saved in the file **uniqueStructures-uf.data**.

5.3. Compute the thermodynamic corrections with the command ‘**run-thermo-pw91.csh uniqueStructures-uf.data**’ and copy/paste the command-line output to a spreadsheet with a name like **gly-h2o-n.xls**.

6. Computing atmospheric concentrations of Gly(H_2O)$_{n=0-5}$ clusters at room temperature at sea-level

NOTE: This is accomplished by setting up a system of chemical equilibria for the sequential hydration of glycine as shown below.

$$\mathrm{Gly} + \mathrm{H_2O} \leftrightarrow \mathrm{Gly}\cdot(\mathrm{H_2O})$$
$$\mathrm{Gly}\cdot(\mathrm{H_2O}) + \mathrm{H_2O} \leftrightarrow \mathrm{Gly}\cdot(\mathrm{H_2O})_2$$
$$\mathrm{Gly}\cdot(\mathrm{H_2O})_2 + \mathrm{H_2O} \leftrightarrow \mathrm{Gly}\cdot(\mathrm{H_2O})_3$$
$$\mathrm{Gly}\cdot(\mathrm{H_2O})_3 + \mathrm{H_2O} \leftrightarrow \mathrm{Gly}\cdot(\mathrm{H_2O})_4$$
$$\mathrm{Gly}\cdot(\mathrm{H_2O})_4 + \mathrm{H_2O} \leftrightarrow \mathrm{Gly}\cdot(\mathrm{H_2O})_5$$

6.1. Compute the equilibrium constants K_n using $K_n = e^{-\Delta G_n/(k_B T)}$, where n is the level of hydration, ΔG_n is the Gibbs free energy change of the nth hydration reaction, k_B is Boltzmann's constant, and T is temperature.

$$K_1 = \frac{[\mathrm{Gly}\cdot(\mathrm{H_2O})]}{[\mathrm{Gly}][\mathrm{H_2}O]}$$
$$K_2 = \frac{[\mathrm{Gly}\cdot(\mathrm{H_2O})_2]}{[\mathrm{Gly}\cdot(\mathrm{H_2O})][\mathrm{H_2}O]}$$
$$K_3 = \frac{[\mathrm{Gly}\cdot(\mathrm{H_2O})_3]}{[\mathrm{Gly}\cdot(\mathrm{H_2O})_2][\mathrm{H_2}O]}$$
$$K_4 = \frac{[\mathrm{Gly}\cdot(\mathrm{H_2O})_4]}{[\mathrm{Gly}\cdot(\mathrm{H_2O})_3][\mathrm{H_2}O]}$$
$$K_5 = \frac{[\mathrm{Gly}\cdot(\mathrm{H_2O})_5]}{[\mathrm{Gly}\cdot(\mathrm{H_2O})_4][\mathrm{H_2}O]}$$

6.2. Set up the equation for the conservation of mass, using the assumption that the sum of the equilibrium concentrations of the hydrated and un-hydrated glycine clusters equals the initial concentration of isolated glycine $[\mathrm{Gly}]_0$. Rewrite this system of six simultaneous equations, using some algebraic rearrangement of the equilibrium constant expressions, as

$$[\mathrm{Gly}]_0 = [\mathrm{Gly}] + [\mathrm{Gly}\cdot(\mathrm{H_2O})] + [\mathrm{Gly}\cdot(\mathrm{H_2O})_2] + [\mathrm{Gly}\cdot(\mathrm{H_2O})_3] + [\mathrm{Gly}\cdot(\mathrm{H_2O})_4] + [\mathrm{Gly}\cdot(\mathrm{H_2O})_5]$$
$$K_1[\mathrm{Gly}][\mathrm{H_2O}] = [\mathrm{Gly}\cdot(\mathrm{H_2O})]$$
$$K_2[\mathrm{Gly}\cdot(\mathrm{H_2O})][\mathrm{H_2O}] = [\mathrm{Gly}\cdot(\mathrm{H_2O})_2]$$
$$K_3[\mathrm{Gly}\cdot(\mathrm{H_2O})_2][\mathrm{H_2O}] = [\mathrm{Gly}\cdot(\mathrm{H_2O})_3]$$
$$K_4[\mathrm{Gly}\cdot(\mathrm{H_2O})_3][\mathrm{H_2O}] = [\mathrm{Gly}\cdot(\mathrm{H_2O})_4]$$
$$K_5[\mathrm{Gly}\cdot(\mathrm{H_2O})_4][\mathrm{H_2O}] = [\mathrm{Gly}\cdot(\mathrm{H_2O})_5].$$

6.3. Solve the system of equations shown above to obtain the equilibrium concentrations of Gly(H_2O)$_{n=0-5}$ using an experimental value[56-58] for the concentration of glycine in the atmosphere, $[\mathrm{Gly}]_0 = 2.9 \times 10^6\ cm^{-3}$, and the concentration of water in the atmosphere at 100% relative humidity and a temperature of 298.15 K[59], $[H_2\mathrm{O}] = 7.7 \times 10^{17} cm^{-3}$.

REPRESENTATIVE RESULTS:

The first set of results from this protocol should be a set of low-energy structures of Gly(H_2O)$_{n=1-5}$ found through the configurational sampling procedure. These structures have been optimized at the PW91/6-311++G** level of theory and are assumed to be accurate for the purpose of this paper. There is no evidence to suggest that PW91/6-311++G** consistently underestimates or overestimates the binding energy of these clusters. Its ability to predict binding energies relative to MP2/CBS[32] and [DLPNO-]CCSD(T)/CBS[60,61] estimates and experiment[52] shows a lot of fluctuations. The same is true of most other density functionals.

The disk and memory usage by the GA code (OGOLEM) and semi-empirical codes (MOPAC) are very small by modern computer resource standards. The overall memory and disk usage for OGOLEM/MOPAC depends on how many threads one wants to use, and even then the resource usage will be small compared to the capabilities of most HPC systems. The resource needs of the QM methods depend on the size of the clusters and the level of theory used. The advantage of using this protocol is that one can vary the level of theory to be able to calculate the final set of low energy structures, keeping in mind that usually faster calculations lead to more uncertainty in accuracy of the results.

Generally, each value of n = 1 – 5 should yield a handful of low-energy structures within around 5 kcal mol^{-1} of the lowest-energy structure. Here, we focus on the first structure produced by the **run-thermo-pw91.csh** script for brevity. **Figure 3** shows the lowest electronic energy isomers of $Gly(H_2O)_{n=0-5}$ clusters. One can see that the hydrogen bond network grows in complexity as the number of water molecules increases, and even goes from a mostly planar network to a three-dimensional cage-like structure at n = 5. The rest of this text uses the energies and thermodynamic quantities corresponding to these five specific clusters.

Table 1 contains the thermodynamic quantities necessary to carry out the protocol. **Table 2** shows an example of the output of the **run-thermo-pw91.csh** script where the electronic energies, vibrational zero-point corrections, and the thermodynamic corrections at three different temperatures are printed. For each cluster (row), **E[PW91/6-311++G**]** corresponds to the gas phase electronic energies at the PW91/6-311++G** level of theory calculated on ultrafine integration grids in units of Hartree, as well as the zero-point vibrational energy (**ZPVE**) in units of kcal mol^{-1}. At each temperature, 216.65 K, 273.15 K, and 298.15 K, the thermodynamic corrections are listed, **ΔH** the enthalpy of formation, **S** the entropy of formation, and **ΔG** the Gibbs free energy of formation in units of kcal mol^{-1}. **Table 3** shows an example computation of the total Gibbs free energy change of hydration, as well as for sequential hydration. An example computation of the total Gibbs free energy change of hydration for the reaction

$$\mathrm{Gly}\ + \mathrm{H_2O}\ \leftrightarrow\ \mathrm{Gly}\cdot(\mathrm{H_2O})$$

starts with the computation of the electronic energy E_{PW91} as

$$\Delta E_{PW91} = \mathrm{E_{PW91}}[\mathrm{Gly}\cdot(\mathrm{H_2O})]\ -\ \mathrm{E_{PW91}}[\mathrm{Gly}]\ -\ \mathrm{E_{PW91}}[H_2\mathrm{O}]$$

where $\mathrm{E_{PW91}}[\mathrm{Gly}\cdot(\mathrm{H_2O})]$ is taken from **Table 2** column C, and $\mathrm{E_{PW91}}[\mathrm{Gly}]$ and $\mathrm{E_{PW91}}[H_2\mathrm{O}]$ are taken from **Table 1** column B. Next we calculate the total gas phase energy change $\Delta E(0)$ by including the change in the zero-point vibrational energy of the reaction as

$$\Delta E(0) = \Delta \mathrm{E_{PW91/6-311++G**}} + (\mathrm{E_{ZPVE}}[\mathrm{Gly}\cdot(\mathrm{H_2O})]\ -\ \mathrm{E_{ZPVE}}[\mathrm{Gly}]\ -\ \mathrm{E_{ZPVE}}[H_2\mathrm{O}])$$

to obtain column D. Here, $\Delta \mathrm{E_{PW91/6-311++G**}}$ is taken from **Table 3** column C, $\mathrm{E_{ZPVE}}[\mathrm{Gly}\cdot(\mathrm{H_2O})]$ from **Table 2** column D, and $\mathrm{E_{ZPVE}}[\mathrm{Gly}]$ and $\mathrm{E_{ZPVE}}[H_2\mathrm{O}]$ from **Table 1** column C. For the sake of brevity, we will move on to room temperature clusters, so we skip over the 216.65 K and 273.15 K data. At room temperature, we then calculate the enthalpy change of the reaction ΔH by correcting the gas phase energy change as

$$\Delta H = \Delta E(0) + (\Delta H[\mathrm{Gly} \cdot (\mathrm{H_2O})] \; - \; \Delta H[\mathrm{Gly}] \; - \; \Delta H[H_2\mathrm{O}])$$

where $\Delta E(0)$ is taken from **Table 3** column D, $\Delta H[\mathrm{Gly} \cdot (\mathrm{H_2O})]$ is taken from **Table 2** column K, and $\Delta H[\mathrm{Gly}]$ and $\Delta H[H_2\mathrm{O}]$ are taken from **Table 1** column J. Finally, we calculate the Gibbs free energy change of the reaction ΔG as

$$\Delta G = \Delta H \; - \; 298.15 \; K \; (S[\mathrm{Gly} \cdot (\mathrm{H_2O})] \; - \; S[\mathrm{Gly}] \; - \; S[H_2\mathrm{O}])$$

where ΔH is taken from **Table 3** column I, $S[\mathrm{Gly} \cdot (\mathrm{H_2O})]$ is taken from **Table 2** column L, and $S[\mathrm{Gly}]$ and $S[H_2\mathrm{O}]$ are taken from **Table 1** column K.

We now have the necessary quantities to compute the atmospheric concentrations of hydrated glycine as shown in **Step 6**. The results should resemble the data shown in **Table 4**, but small numerical differences are to be expected. **Table 4** shows the formulation of the system of six equations in **Step 6.2** into one matrix equation and its subsequent solution. We start by acknowledging the fact that the system of equations can be written as

$$\begin{pmatrix} K_1 w & -1 & 0 & 0 & 0 & 0 \\ 0 & K_2 w & -1 & 0 & 0 & 0 \\ 0 & 0 & K_3 w & -1 & 0 & 0 \\ 0 & 0 & 0 & K_4 w & -1 & 0 \\ 0 & 0 & 0 & 0 & K_5 w & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ g \end{pmatrix}$$

where K_n is the equilibrium constant for the nth sequential hydration of glycine, w is the concentration of water in the atmosphere, g is the initial concentration of isolated glycine in the atmosphere, and g_n is the equilibrium concentration of Gly(H_2O)$_n$. If we rewrite the above equation as $\boldsymbol{A}x = b$, we get $x = \boldsymbol{A}^{-1}b$ where $\boldsymbol{A}^{-1}$ is the inverse of matrix $\boldsymbol{A}$. This inverse can be easily computed using built-in spreadsheey functions as shown in **Table 4** to obtain the final results.

As presented, this protocol gives a qualitative understanding of the hydrated glycine populations in the atmosphere. Assuming a non-equilibrium concentration of isolated glycine of 2.9 million molecules per cubic centimeter, we see that the singly-hydrated glycine is the most abundant cluster in the atmosphere, followed by the triply- and quadruply-hydrated clusters. The larger clusters are predicted to be found in negligible amounts. Upon inspection of **Figure 3**, the abundance of the n = 1 – 4 clusters can be related to the stability and strain in the hydrogen bond network of the clusters. These clusters have the water molecules hydrogen bonded to the carboxylic acid moiety of glycine in a geometry closely resembling those of various hydrogen-bonded ring structures, making them especially stable.

FIGURE AND TABLE LEGENDS:

Figure 1. Schematic description of the current procedure. A large pool of guess structures generated by the genetic algorithm (GA) is refined by a series of PW91 geometry optimizations until a set of converged structures are obtained. The vibrational frequencies of these structures are computed and used to compute the Gibbs free energy of formation, which is in turn used to compute the equilibrium concentrations of the clusters under ambient conditions.

Figure 2. Representative directory structure for each cluster. The in-house scripts included in this protocol require the directory structure shown above, where n is the number of water molecules. For each n in **gly-h2o-n**, there are the following subdirectories: GA for genetic algorithm with a GA/pm7 directory, QM for quantum mechanics with QM/pw91-sb for PW91/6-31+G*, QM/pw91-lb for PW91/6-311++G**, and QM/pw91-lb/ultrafine for optimizations and final vibrational calculations on ultrafine integration grids.

Figure 3. Representative low energy structures of Gly(H_2O)$_{n=0-5}$. These clusters were the electronic energy global minima optimized at the PW91/6-311++G** level of theory.

Table 1. Monomer energies. Electronic energies are in units of Hartree while all other quantities are in units of kcal mol^{-1}. Water and glycine were optimized at the PW91/6-311++G** level of theory and vibrational frequencies were computed. The thermodynamic corrections for a pressure of 1 atm and temperature of 298.15 K were computed using the thermo.pl script.

Table 2. Cluster energies. The energies of the lowest-energy Gly(H_2O)$_{n=1-5}$ structures found using our procedure outlined in **Figure 1**. Electronic energies are in units of Hartree while all other quantities are in units of kcal mol^{-1}.

Table 3. Hydration energies. The total energy of hydration and energy of sequential hydration for Gly(H_2O)$_{n=1-5}$ in units of kcal mol^{-1}. Here, E[PW91/6-311++G**] is the change in the electronic energy, ΔE(0) is the zero-point vibrational energy (ZPVE) corrected change in energy, ΔH(T) is the enthalpy change at temperature T, and ΔG(T) is the Gibbs free energy change of hydration of each Gly(H_2O)$_{n=1-5}$ cluster.

Table 4. Atmospheric concentrations. Table containing the computation of the equilibrium concentrations of atmospheric Gly(H_2O)$_{n=1-5}$ clusters. Here, k_B is Boltzmann's constant, R is the ideal gas constant, h is the Planck constant, kCtoJ is a conversion factor from kcal mol^{-1} to J mol^{-1}, amutoKg is a conversion factor from atomic mass units to kilograms, NtoAtm is a conversion factor from number of molecules per cubic centimeter to partial pressure in atmospheres, and Na is Avogadro's number. We used the experimental values[56-58] of $[\mathrm{Gly}]_0 = 2.9 \times 10^6\ \mathrm{cm}^{-3}$ and $[H_2\mathrm{O}] = 7.7 \times 10^{17} cm^{-3}$ at 100% relative humidity of and $\mathrm{T} = 298.15\ \mathrm{K}$[59].

DISCUSSION:

The accuracy of the data generated by this protocol depends mainly on three things: (i) the variety of configurations sampled by Step 2, (ii) the accuracy of the electronic structure of the system, (iii) and the accuracy of the thermodynamic corrections. Each of these factors can be addressed by modifying the method by editing the included scripts. The first factor is easily overcome with the use of a larger initial pool of randomly generated structures, more numerous iterations of the GA, and a looser definition of the criteria involved in the GA. In addition, one may use a different semi-empirical method such as the self-consistent charge density-functional tight-binding (SCC-DFTB)[62] model and the effective fragment potential (EFP)[63] model in order to explore the effects of different physical descriptions. The main limitation here is the inability of the method to form or break covalent bonds, meaning that the monomers are frozen. The GA procedure only finds the most stable relative positions of these frozen monomers according to the semi-empirical description.

The accuracy of the electronic structure of the system can be improved in a variety of ways, each with its computational cost. One may choose a better density functional, such as M06-2X[64] and wB97X-V[65], or quantum chemical method such as the Møller-Plesset[66-68] (MPn) perturbation theories and coupled-cluster[69] (CC) methods in order to improve the physical description of the system. In the hierarchy of functionals, the performance generally improves upon going from generalized-gradient approximation (GGA) functionals like PW91 to range-separated hybrid functionals like wB97X-D and meta-GGA hybrid functionals like M06-2X. The disadvantage of DFT methods is that a systematic convergence towards an accurate value is not possible; however, DFT methods are computationally inexpensive and there is a wide variety of functionals for a wide variety of applications.

Energies calculated using wavefunction methods like MP2 and CCSD(T) in conjunction with correlation consistent basis sets of increasing cardinal number ([aug-]cc-pV[D,T,Q,...]Z) converge towards their complete basis set limit systematically, but the computational cost of each calculation becomes prohibitive as the system size grows. Further refinement of the electronic structure can be accomplished by using explicitly correlated basis sets[70] and by extrapolating to the complete basis set (CBS)[71] limit. Our recent work suggests that a density-fitted explicitly correlated second-order Møller-Plesset (DF-MP2-F12) perturbative approach yields energies approaching that of MP2/CBS computations[32]. Modification of the current protocol to use different electronic structure methods involves two steps: (i) prepare a template input file following the syntax given by the software, (ii) and edit the **run-pw91-sb.csh**, **run-pw91-lb.csh**, and **run-pw91-lb-ultrafine.csh** scripts to generate the correct input file syntax as well as the correct submit script for the software.

Lastly, the accuracy of the thermodynamic corrections depends on the electronic structure method as well as the description of the PES around the global minimum. An accurate description of the PES requires the computation of third- and higher-order derivatives of the PES with respect to displacements in the nuclear degrees of freedom, such as the quartic force field[72,73] (QFF), which is an exceptionally costly task. The current protocol uses the harmonic oscillator approximation to the vibrational frequencies, resulting in the need to compute only up to second derivatives of the PES. This approach becomes problematic in systems with high anharmonicity, such as very floppy molecules and symmetric double-well potentials due to the large difference in the true PES and the harmonic PES. Furthermore, the cost of having a high-quality PES from a computationally demanding electronic structure method only compounds the problem of cost for vibrational frequency calculations. One approach to overcome this is to use the electronic energies from a high-quality electronic structure calculation along with vibrational frequencies computed on a lower quality PES, resulting in a balance between cost and accuracy. The current protocol can be modified to use different PES descriptions as described in the previous paragraph; however, one may also edit the vibrational frequency keywords in the scripts and templates to compute anharmonic vibrational frequencies.

Two crucial issues for any configurational sampling protocol are the initial method for sampling the potential energy surface and the criteria used to identify each cluster. We have made extensive use of a variety of methods in our previous work. For the first issue, the initial method for sampling the potential energy surface, we have made the choice of using GA with semi-empirical methods based on these factors. Configurational sampling using chemical intuition[26], random sampling, and molecular dynamics (MD)[29,30], fail to find putative global minima regularly for clusters larger than 10 monomers, as we observed in our studies of water clusters[18]. We have successfully used basin hopping (BH) to study the complex PES of $(H_2O)_{11}$[74], but it required the manual inclusion of some potential low energy isomers the BH algorithm did not find. A comparison of the performance of BH and GA in finding the global minimum of water clusters, $(H_2O)_{n=10-20}$ demonstrated that GA consistently found the global minimum faster than BH[75]. GA as implemented in OGOLEM and CLUSTER is very versatile because it can be applied to any molecular cluster and it can interface with a vast number of packages with classical force field, semi-empirical, density functional, and ab initio capabilities. The choice of PM7 is driven by its speed and reasonable accuracy. Virtually any other semi-empirical method would have significantly higher computational cost.

As for the second issue, we have explored using different criteria to identify unique structures ranging from electronic energies, dipole moments, overlap RMSDs and rotational constants. Using dipole moments proved difficult because both the dipole moment components were dependent on the molecule’s orientation and the total dipole moment was very sensitive to geometry differences in such a way that it was difficult to set thresholds determining is structures are the same or unique. A combination of electronic energies and rotational constants proved to be most useful.

The current criteria for deeming two structures unique is based on an energy difference threshold of 0.10 kcal mol^{-1} and rotational constant difference of 1%. Therefore, two structures are considered different if their energies differ by more than 0.10 kcal mol^{-1} (~0.00015 a.u.) AND any of their three rotational constants (A, B, C) differ by more than 1%. Substantial internal benchmarks over the years found these thresholds to be reasonable choices. Our configurational sampling approach and screening methodology has been applied to very weakly bound clusters such as polyaromatic hydrocarbons complexed with water[76,77] as well as strongly bound ternary sulfate hydrates containing ammonia and amines[32]. For clusters where there are different protonation states to be considered, the best approach is to run various GA calculations, each starting with monomers in different protonation states. This ensures that structures with different protonation states are carefully considered. However, the low-level

DFT calculations often allow protonation states to change during the course of the geometry optimization, thereby yielding the most stable protonation state regardless of the starting geometry.

Our GA configurational sampling methods should work well even for floppy molecules as long as the GA codes are interfaced with general, non-parameterized methods that allow the monomers to adopt different configurations during the course of the GA run. For example, interfacing GA with PM7 would allow monomers' structures to change, but if their bonds break as would happen when protonation states change, the structures may get discarded as unacceptable candidates.

We have considered different ways of correcting the shortcomings of the harmonic approximation, especially those arising from low vibrational frequencies. Incorporating the quasi-harmonic approximation into the current methodology is not difficult. However, there are still questions about the quasi-harmonic method, especially when it comes to the cutoff frequency below which it will be applied. Also, there are no rigorous benchmarking works examining the reliability of the quasi-RRHO approximation even though conventional wisdom suggests it should be an improvement over RRHO approximation.

The protocol thus presented may be generalized to any system of noncovalently-bound gas phase molecular clusters. It may also be generalized to use any semi-empirical method, electronic structure method and software, and vibrational analysis method and software by editing the scripts and templates. This assumes that the user is comfortable with the Unix command-line interface, Python scripting, and high-performance computing. The unfamiliar syntax and look of the Unix operating system and lack of scripting experience is the largest pitfall in this protocol and is where new students struggle the most. This protocol has been used successfully in a variety of implementations for years in our group, mostly focusing on the effects of sulfuric acid and ammonia on aerosol formation. Further improvements to this protocol will involve a more robust interface to more electronic structure software, alternative implementations of the genetic algorithm, and possibly the use of newer methods for faster computations of electronic and vibrational energies. Our current applications of this protocol are exploring the importance of amino acids in the early stages of aerosol formation in the current atmosphere and in the formation of larger biological molecules in prebiotic environments.

ACKNOWLEDGMENTS:

This project was supported by grants CHE-1662030, CHE-1721511, and CHE-1903871 from the National Science Foundation, and used the high-performance computing resources of the MERCURY Consortium (http://www.mercuryconsortium.org).

DISCLOSURES:

None.

REFERENCES:

1. Foster, P., Ramaswamy, V., In *Climate Change 2007 The Scientific Basis*. Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., Miller, H. L., Eds. Cambridge University Press. Cambridge, U.K. (2007).

2. Kulmala, M. et al. Toward direct measurement of atmospheric nucleation. *Science*. **318** (5847), 89-92 (2007).

3. Sipila, M. et al. The role of sulfuric acid in atmospheric nucleation. *Science*. **327** (5970), 1243-1246 (2010).

4. Jiang, J. et al. First measurement of neutral atmospheric cluster and 1 – 2 nm particle number size distributions during nucleation events. *Aerosol Science and Technology.* **45** (4), ii-v (2011).

5. Dunn, M.E., Pokon, E.K., Shields, G.C. Thermodynamics of forming water clusters at various Temperatures and Pressures by Gaussian-2, Gaussian-3, Complete Basis Set-QB3, and Complete Basis Set-APNO model chemistries; implications for atmospheric chemistry. *Journal of the American Chemical Society*. **126** (8), 2647-2653 (2004).

6. Pickard, F.C., Pokon, E.K., Liptak, M.D., Shields, G.C. Comparison of CBSQB3, CBSAPNO, G2, and G3 thermochemical predictions with experiment for formation of ionic clusters of hydronium and hydroxide ions complexed with water. *Journal of Chemical Physics.* **122**, 024302 (2005).

7. Pickard, F.C., Dunn, M.E., Shields, G.C. Comparison of Model Chemistry and Density Functional Theory Thermochemical Predictions with Experiment for Formation of Ionic Clusters of the Ammonium Cation Complexed with Water and Ammonia; Atmospheric Implications. *Journal of Physical Chemistry A.* **109** (22), 4905-4910 (2005).

8. Alongi, K.S., Dibble, T.S., Shields, G.C., Kirschner, K.N. Exploration of the Potential Energy Surfaces, Prediction of Atmospheric Concentrations, and Vibrational Spectra of the HO_2•••$(H_2O)_n$ (n=1-2) Hydrogen Bonded Complexes. Journal of Physical Chemistry A. **110** (10), 3686-3691 (2006).

9. Allodi, M.A., Dunn, M.E., Livada, J., Kirschner, K.N. Do Hydroxyl Radical-Water Clusters, $OH(H_2O)_n$, n=1-5, Exist in the Atmosphere? *Journal of Physical Chemistry A.* **110** (49), 13283-13289 (2006).

10. Kirschner, K.N., Hartt, G.M., Evans, T.M., Shields, G.C. In Search of $CS_2(H_2O)_{n=1\text{-}4}$ Clusters. *Journal of Chemical Physics.* **126**, 154320 (2007).

11. Hartt, G.M., Kirschner, K.N., Shields, G.C. Hydration of OCS with One to Four Water Molecules in Atmospheric and Laboratory Conditions. *Journal of Physical Chemistry A.* **112** (19), 4490-4495 (2008).

12. Morrell, T.E., Shields, G.C. Atmospheric Implications for Formation of Clusters of Ammonium and 1–10 Water Molecules. *Journal of Physical Chemistry A.* **114** (12), 4266-4271 (2010).

13. Temelso, B. et al. Quantum Mechanical Study of Sulfuric Acid Hydration: Atmospheric Implications. *Journal of Physical Chemistry A.* **116** (9), 2209-2204 (2012).

14. Husar, D.E., Temelso, B., Ashworth, A.L., Shields, G.C. Hydration of the Bisulfate Ion: Atmospheric Implications. *Journal of Physical Chemistry A.* **116** (21), 5151-5163 (2012).

15. Bustos, D.J., Temelso, B., Shields, G.C. Hydration of the Sulfuric Acid – Methylamine Complex and Implications for Aerosol Formation. *Journal of Physical Chemistry A.* **118** (35), 7430-7441 (2014).

16. Wales, D. J., Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science*. **27** (5432), 1368-1372 (1999).

17. Day, M. B., Kirschner, K. N., Shields, G. C. Global search for minimum energy $(H_2O)_n$ clusters, n = 3 – 5. *The Journal of Physical Chemistry A*. **109** (30), 6773-6778 (2005).

18. Shields, R. M., Temelso, B., Archer, K. A., Morrell, T. E., Shields, G. C. Accurate predictions of water cluster formation, $(H_2O)_{n=2-10}$. *The Journal of Physical Chemistry A*. **114** (43), 11725-11737 (2010).

19. Temelso, B., Archer, K. A., Shields, G. C. Benchmark structures and binding energies of small water clusters with anharmonicity corrections. *The Journal of Physical Chemistry A*. **115** (43), 12034-12046 (2011).

20. Temelso, B., Shields, G. C. The role of anharmonicity in hydrogen-bonded systems: The case of water clusters. *The Journal of Chemical Theory and Computation*. **7** (9), 2804-2817 (2011).

21. Von Freyberg, B., Braun, W. Efficient search for all low energy conformations of polypeptides by Monte Carlo methods. *The Journal of Computational Chemistry*. **12** (9), 1065-1076 (1991).

22. Rakshit, A., Yamaguchi, T., Asada, T., Bandyopadhyay, P. Understanding the structure and hydrogen bonding network of $(H_2O)_{32}$ and $(H_2O)_{33}$: An improved Monte Carlo temperature basin paving (MCTBP) method of quantum theory of atoms in molecules (QTAIM) analysis. *RSC Advances*. **7** (30), 18401-18417 (2017).

23. Deaven, D. M., Ho, K. M., Molecular geometry optimization with a genetic algorithm. *Physical Review Letters*. **75**, 288-291 (1995).

24. Hartke, B. Application of evolutionary algorithms to global cluster geometry optimization. *Applications of Evolutionary Computation in Chemistry.* Springer. Berlin (2004).

25. Dieterich, J. M., Hartke, B. OGOLEM: Global cluster structure optimization for arbitrary mixtures of flexible molecules. A multiscaling, object-oriented approach. *Molecular Physics*. **108** (3-4), 279-291 (2010).

26. Herb, J. Nadykto, A. B., Yu, F. Large ternary hydrogen-bonded pre-nucleation clusters in the Earth's atmosphere. *Chemical Physics Letters.* **518**, 7– 14 (2011).

27. Ortega et al. From quantum chemical formation free energies to evaporation rates. *Atmospheric Chemistry and Physics.* **12** (1), 225– 235 (2012).

28. Elm, J., Bilde, M., Mikkelsen, K.V. Influence of Nucleation Precursors on the Reaction Kinetics of Methanol with the OH Radical. *Journal of Physical Chemistry A.* **117** (30), 6695– 6701 (2013).

29. Loukonen,V. et al. Enhancing effect of dimethylamine in sulfuric acid nucleation in the presence of water – a computational study. *Atmospheric Chemistry and Physics.* **10** (10), 4961−4974 (2010).

30. Temelso, B., Phan, T.N., Shields, G.C. Computational study of the hydration of sulfuric acid dimers: implications for acid dissociation and aerosol formation. *Journal of Physical Chemistry A.* **116** (39), 9745−9758 (2012).

31. Jiang, S. et al. Study of $Cl^-(H_2O)_n$ (n = 1–4) using basin-hopping method coupled with density functional theory**.** *Journal of Computational Chemistry.* **35** (2), 159– 165 (2014).

32. Temelso, B. et al. Effect of mixing ammonia and alkylamines on sulfate aerosol formation. *Journal of Physical Chemistry A.* **122** (6), 1612−1622 (2018).

33. Perdew, J.P., Ruzsinszky, A., Tao, J. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *Journal of Chemical Physics.* **123**, 062201 (2005).

34. Riplinger, C., Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *Journal of Chemical Physics.* **138**, 034106 (2013).

35. Riplinger, C., Pinski, P., Becker, U., Valeev, E.F., Neese, F. Sparse maps--A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *Journal of Chemical Physics.* **144** (2), 024109 (2016).

36. Kildgaard, J.V., Mikkelsen, K.V., Bilde, M., Elm, J. Hydration of atmospheric molecular clusters: a new method for systematic configurational sampling. *Journal of Physical Chemistry A.* **122** (22), 5026-5036 (2018).

37. González, Á. Measurement of areas on a sphere Using Fibonacci and latitude–longitude lattices *Mathematical Geosciences.* **42**, 49– 64 (2010).

38. Karaboga, D., Basturk, B. On the performance of artificial bee colony (ABC) algorithm. *Applied Soft Computing.* **8** (1), 687– 697 (2008).

39. Zhang, J., Doig, M. Global optimization of rigid molecules using the artificial bee colony algorithm. *Physical Chemistry Chemical Physics.* **18** (4), 3003– 3010 (2016).

40. Kubecka, J.,Besel, V., Kurten, T., Myllys, N., Vehkamaki, H. Configurational sampling of noncovalent (atmospheric) molecular clusters: sulfuric acid and guanidine. *Journal of Physical Chemistry A.* **123** (28), 6022−6033 (2019).

41. Grimme, S., Bannwarth, C., Shushkov, P. A Robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent Interactions of large molecular systems parametrized for all spd-block elements (*Z* = 1–86). *Journal of Chemical Theory and Computation.* **13** (5), 1989– 2009 (2017).

42. Buck, U., Pradzynski, C. C., Zeuch, T., Dieterich, J. M., Hartke, B. A size resolved investigation of large water clusters. *Physical Chemistry Chemical Physics*. **16** (15), 6859-4871 (2014).

43. Forck, R. M. et al. Structural diversity in sodium doped water trimers. *Physical Chemistry Chemical Physics*. **14** (25), 9054-9057.

44. Witt, C., Dieterich, J. M., Hartke, B. Cluster structures influenced by interaction with a surface. *Physical Chemistry Chemical Physics*. **20** (23), 15661-15670 (2018).

45. Freitbert, A., Dieterich, J. M., Hartke, B. Exploring self-organization of molecular tether molecules on a gold surface by global structure optimization. *The Journal of Computational Chemistry*. **40** (22), 1978-1989 (2019).

46. Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *The Journal of Molecular Modeling*. **19** (1), 1-32 (2013).

47. Stewart, J. J. P. MOPAC2012 Computational Chemistry. http://openmopac.net (2012).

48. Burke, K., Perdew, J. P., Wang, Yue. Derivation of a generalized gradient approximation: The PW91 density functional. In *Electronic Density Functional Theory*. Springer, Boston, MA. 81-111 (1998).

49. Frisch, M. J. et al. Gaussian 09, Revision A.02. Gaussian, Inc., Wallingford, CT (2016).

50. Ditchfield, R., Hehre, W. J., Pople, J. A. Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics*. **54** (2), 724 (1971).

51. Elm, J., Bilde, M., Mikkelsen, K. V. Assessment of density functional theory in predicting structures and free energies of reaction of atmospheric prenucleation clusters. *The Journal of Chemical Theory and Computation*. **8** (6), 2071-2077 (2012).

52. Elm, J., Mikkelsen, K. V. Computational approaches for efficiently modelling of small atmospheric clusters. *Chemical Physics Letters*. **615**, 26-29 (2014).

53. Bayucan, A. et al. PBS Portable Batch System. MRJ Technology Solutions. Mountain View, CA (1999).

54. O'Boyle, N. M. et al. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*. **3**, 33, (2011).

55. Csaszar, A.G. Conformers of gaseous glycine. *Journal of the American Chemical Society*. **114** (24), 9568-9575 (1992).

56. Zhang, Q., Anastasio, C. Free and combined amino compounds in atmospheric fine particles (PM2.5) and fog waters from Northern California. *Atmospheric Environment*. **37** (16), 2247-2258 (2003).

57. Matsumoto, K., Uematsu, M. Free amino acids in marine aerosols over the western North Pacific Ocean. *Atmospheric Environment*. **39** (11), 2163-2170 (2005).

58. Mandalakis, M., Apostolaki, M., Stephanou, E.G. Trace analysis of free and combined amino acids in atmospheric aerosols by gas chromatography-mass spectrometry. *Journal of Chromatography A*. **1217** (1), 143-150 (2010).

59. Seinfeld, J.H., Pandis, S.N. Atmospheric Chemistry and Physics, 3rd Ed., John Wiley & Sons. Hoboken, N.J. (2016).

60. Myllys, N., Elm, J., Halonen, R., Kurten, T., Vehkamaki, H. Coupled cluster evaluation of atmospheric acid-base clusters with up to 10 molecules. *The Journal of Physical Chemistry A.* **120** (4), 621–630 (2016).

61. Elm, J., Bilde, M., Mikkelsen, K.V. Assessment of binding energies of atmospherically relevant clusters. *Physical Chemistry Chemical Physics*. **15** (39), (2013).

62. Elstner, M. The SCC-DFTB method and its application to biological systems. *Theoretical Chemistry Accounts*. **116** (1-3), 316-325 (2006).

63. Kaliman, I. A., Slipchenko, L. V. LIBEFP: A new parallel implementation of the effective fragment potential method as a portable software library. *The Journal of Computational Chemistry*. **34** (26), 2284-2292 (2013).

64. Zhao, Y., Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and trasition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*. **120** (1-3), 215-241 (2008).

65. Mardirossian, N., Head-Gordon, M. wB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Physical Chemistry Chemical Physics*. **16** (21), 9904-9924 (2014).

66. Head-Gordon, M., Pople, J. A., Frisch, M. J. MP2 energy evaluation by direct methods. *Chemical Physics Letters*. **153** (6), 503-506 (1988).

67. Pople, J. A., Seeger, R., Krishnan, R. Variational configuration interaction methods and comparison with perturbation theory. *The International Journal of Quantum Chemistry*. **12** (S11), 149-163 (1977).

68. Pople, J. A., Binkley, J. S., Seeger, R. Theoretical models incorporating electron correlation. *The International Journal of Quantum Chemistry.* **10** (S10), 1-19 (1976).

69. Monkhorst, H. J. Calculation of properties with the coupled-cluster method. *The International Journal of Quantum Chemistry*. **12** (S11), 421-432 (1977).

70. Klopper, W., Manby, F. R., Ten-No, S., Valeev, E. F. R12 methods in explicitly correlated molecular electronic structure theory. *International Reviews in Physical Chemistry*. **25**, 427-468 (2006).

71. Hattig, C. Optimization of auxiliary basis sets for RI-MP2 and RI-CC2 calculations: Core-valence and quintuple-z basis sets for H to Ar and QZVPP basis sets for Li to Kr. *Physical Chemistry Chemical Physics*. **7** (1), 59-66 (2005).

72. Barone, V. Anharmonic vibrational properties by a fully automated second-order perturbative approach. *The Journal of Chemical Physics*. **122**, 014108 (2005).

73. Barone, V. Vibrational zero-point energies and thermodynamic functions beyond the harmonic approximation. *The Journal of Chemical Physics*. **120** (7), 3059-3065 (2004).

74. Temelso, B. et al. Exploring the Rich Potential Energy Surface of $(H_2O)_{11}$ and Its Physical Implications. *Journal of Chemical Theory and Computation.* **14** (2), 1141−1153 (2018).

75. Kabrede, H., Hentschke, R. Global minima of water clusters $(H_2O)_N$, $N \leq 25$, described by three empirical potentials. *Journal of Physical Chemistry B.* **107** (16) (2003).

76. Steber, A.L. et al. Capturing the Elusive Water Trimer from the Stepwise Growth of Water on the Surface of a Polycyclic Aromatic Hydrocarbon Acenaphthene. *Journal of Physical Chemistry Letters.* **8** (23), 5744-5750 (2017).

77. Perez, C. et al. Corrannulene and its complex with water: A tiny cup of water. *Physical Chemistry Chemical Physics*. **19** (22), 14214-14223 (2017).

Figure 1

Click here to access/download;Figure;figure-1.png

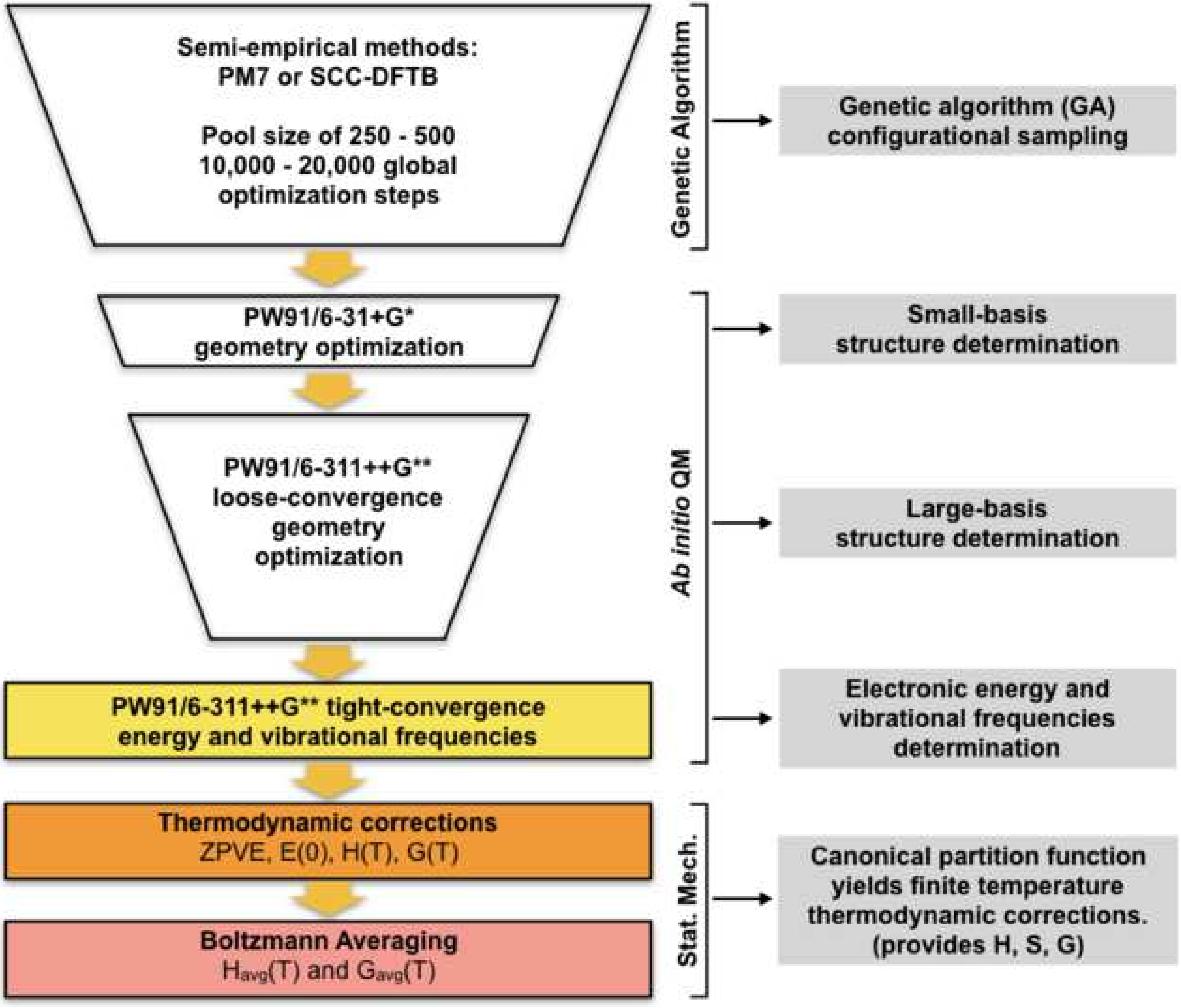
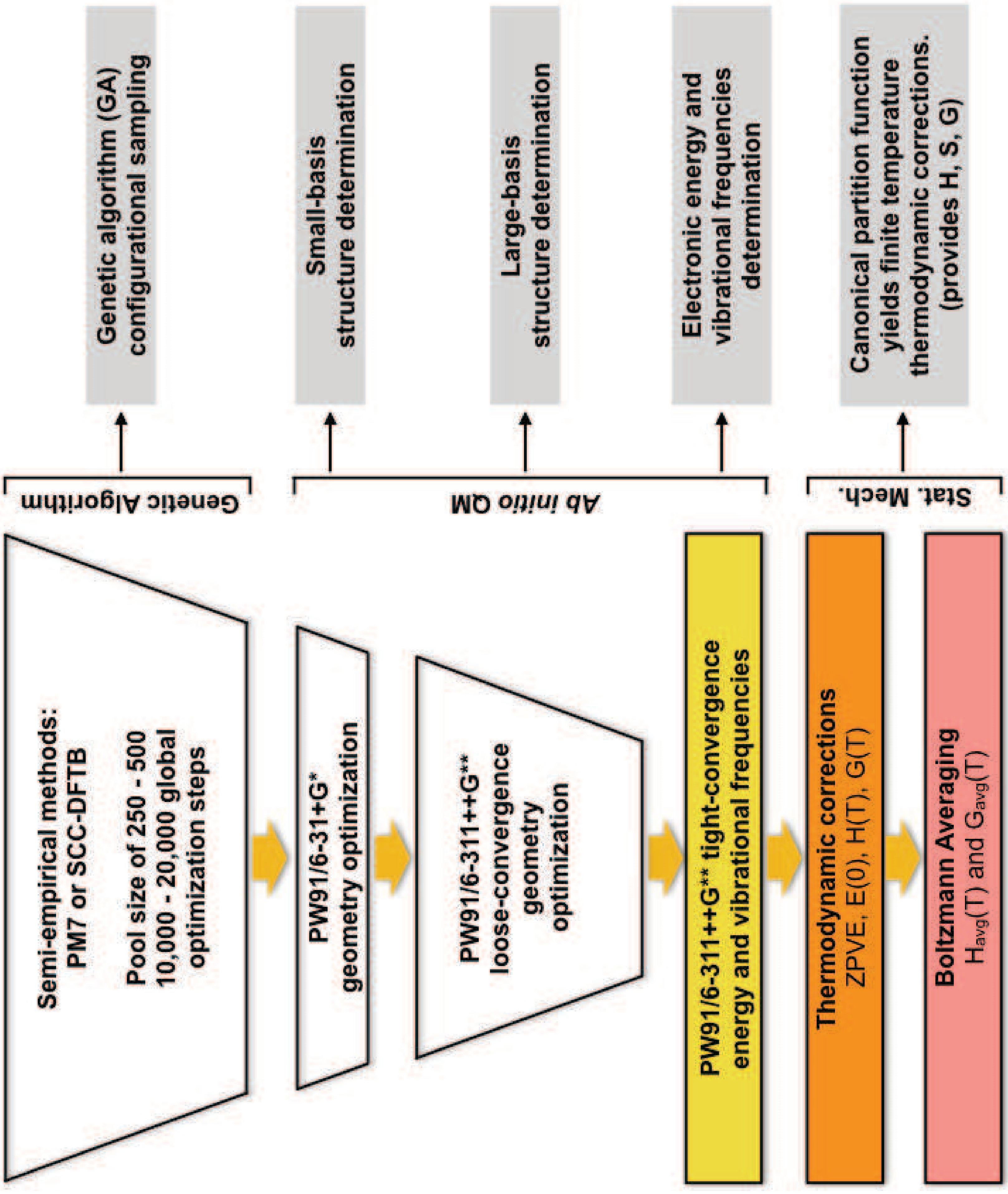


Figure 1

Click here to access/download;Figure;figure-1.eps



Click here to access/download;Figure;Figure-2.eps

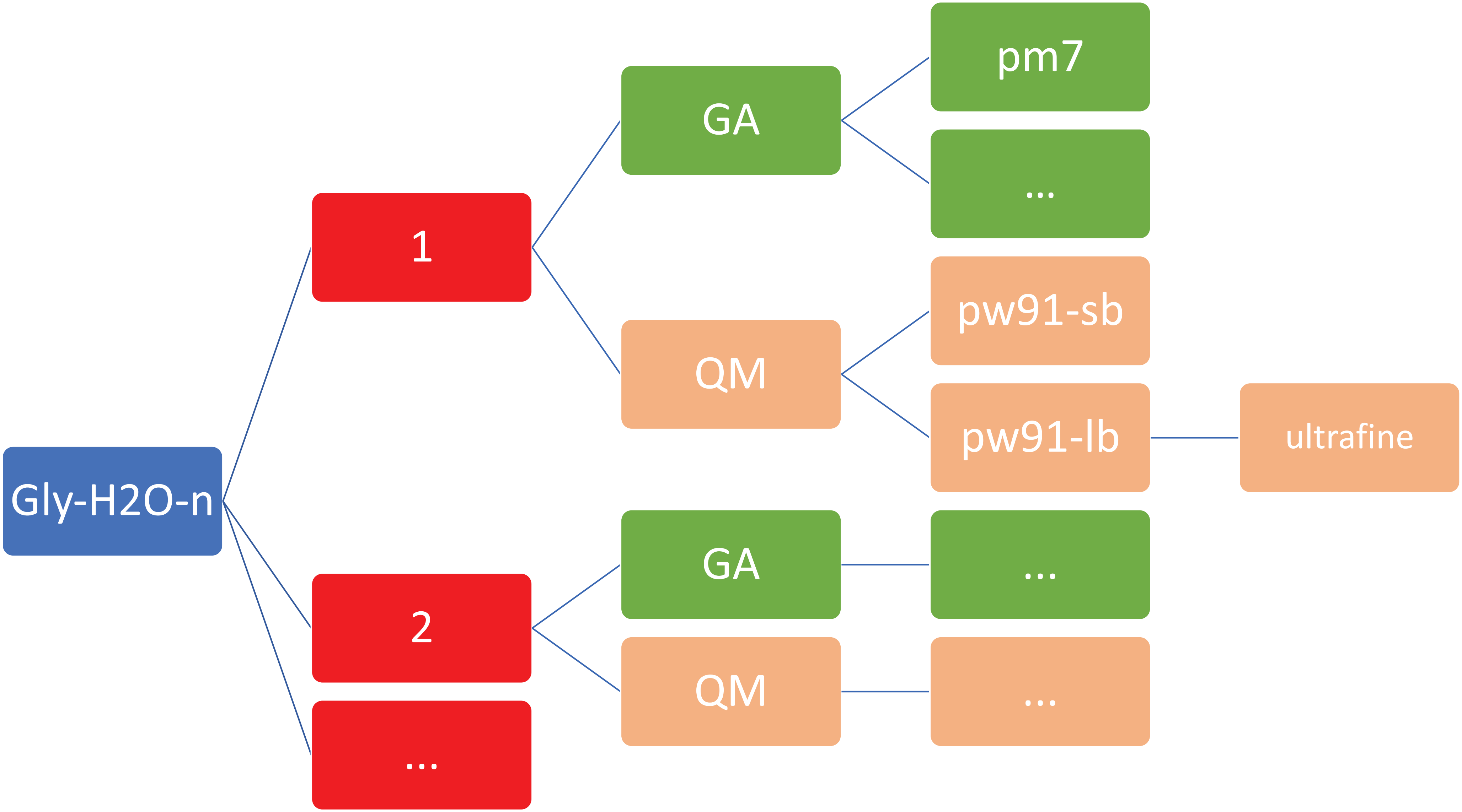


Figure 3

Click here to access/download;Figure;figure-3.eps

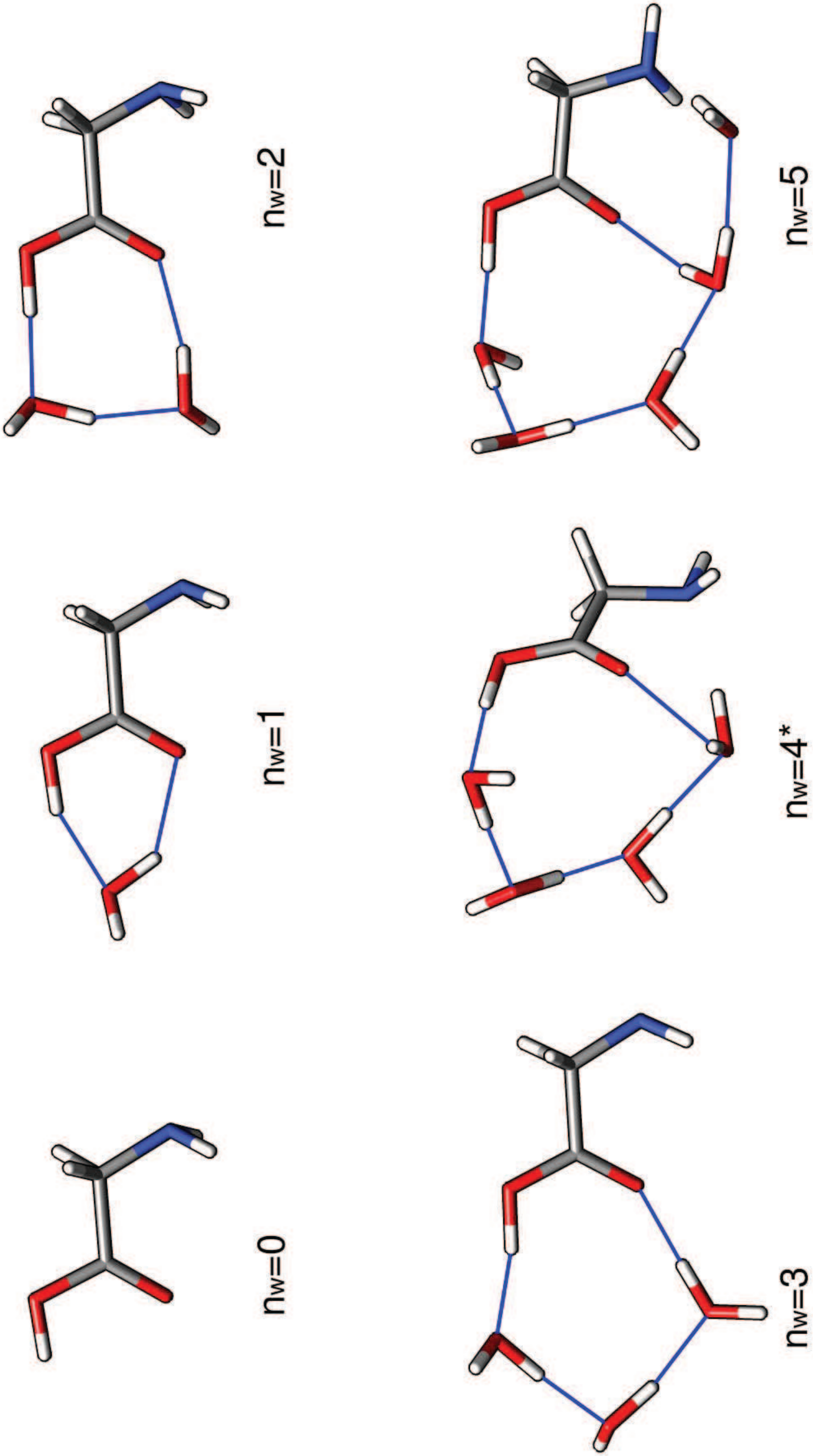


Table 1

Click here to access/download;Table;table-1.xlsx

	E[PW91/6-311++G**]		216.65 K			
	LB-UF	ZPVE	ΔH	S	ΔG	ΔH
water	-76.430500	13.04	1.72	42.59	5.54	2.17
glycine	#########	48.55	2.65	69.53	36.14	3.70

273.15 K		**298.15 K**		
S	**ΔG**	**ΔH**	**S**	**ΔG**
44.44	3.08	2.37	45.14	1.96
73.81	32.09	4.22	75.61	30.22

Table 2

Click here to access/download;Table;table-2.xlsx

		E[PW91/6-311++G**]	0 K	216.65 K	
n	name	LB-UF	ZPVE	ΔH	S
1	gly-h2o-1	-360.88481	63.96	3.61	80.12
2	gly-h2o-2	-437.33763	79.33	4.53	90.86
3	gly-h2o-3	-513.78620	94.52	5.67	105.08
4	gly-h2o-4	-590.23667	109.80	6.03	104.98
5	gly-h2o-5	-666.68845	125.80	7.26	121.70

	273.15 K			298.15 K		
ΔG	**ΔH**	**S**	**ΔG**	**ΔH**	**S**	**ΔG**
50.22	5.12	86.27	45.52	5.85	88.83	43.33
64.17	6.46	98.78	58.81	7.40	102.06	56.30
77.42	8.08	114.94	71.19	9.23	119.00	68.27
91.30	8.78	116.21	84.40	10.11	120.87	81.14
106.69	10.47	134.83	99.44	12.01	140.24	96.00

Table 3

Click here to access/download;Table;table-3.xlsx

		Total Hydration: Gly +			
		E[PW91/6-311++G**]		216.65	
n	system name	LB-UF	ΔE(0)	ΔH(T)	ΔG(T)
1	gly-h2o-1	-12.22	-9.85	-10.61	-3.68
2	gly-h2o-2	-26.22	-21.53	-23.10	-9.27
3	gly-h2o-3	-37.56	-30.72	-32.88	-12.90
4	gly-h2o-4	-50.10	-40.34	-43.48	-15.87
5	gly-h2o-5	-63.45	-51.41	-55.42	-20.58

- nH2O <-> Gly(H2O)n				Sequential Hyd		
273.15		298.15				216
ΔH(T)	ΔG(T)	ΔH(T)	ΔG(T)	LB-UF	ΔE(0)	ΔH(T)
-10.61	-1.87	-10.59	-1.07	-12.22	-9.85	-10.61
-23.11	-5.66	-23.09	-4.06	-14.00	-11.68	-12.49
-32.87	-7.69	-32.82	-5.38	-11.34	-9.19	-9.78
-43.54	-8.71	-43.51	-5.55	-12.54	-9.62	-10.60
-55.51	-11.48	-55.48	-7.45	-13.35	-11.07	-11.94

dration: Gly(H2O)n-1 + H2O <-> Gly(H2O)n				
.65	273.15		298.15	
ΔG(T)	H(T)	ΔG(T)	ΔH(T)	ΔG(T)
-3.68	-10.61	-1.87	-10.59	-1.07
-5.59	-12.50	-3.79	-12.50	-2.99
-3.63	-9.76	-2.03	-9.73	-1.32
-2.97	-10.67	-1.02	-10.69	-0.17
-4.71	-11.97	-2.77	-11.97	-1.90

Table 4

Click here to access/download;Table;table-4.xlsx

kB	1.99E-03
R	8.314472
h	6.63E-34
kCtoJ	4184
amutoKg	1.66E-27
NtoAtm	2.69E+19
Na	6.02E+23
Relative Hum	100
Water Concei	7.7E+17
Glycine Conc‹	2.90E+06

		S
	0	
n	**LB-UF**	**ΔE(0)**
1	-12.22	-9.85
2	-14.00	-11.68
3	-11.34	-9.19
4	-12.54	-9.62
5	-13.35	-11.07

A =	1.74E-01
	0
	0
	0
	0
	1

A^-1 =	3.09410606
	-0.4608995
	-2.0523679
	-0.5453306
	-0.020798
	-0.01471

equential Hydration: Gly(H2O)n-1 + H2O <-> Gly(H2O)n						
216.65		273.15		298.15		Equilb
ΔH(T)	ΔG(T)	ΔH(T)	ΔG(T)	ΔH(T)	ΔG(T)	216.65
-10.61	-3.68	-10.61	-1.87	-10.59	-1.07	5159.15
-12.49	-5.59	-12.50	-3.79	-12.50	-2.99	435999.98
-9.78	-3.63	-9.76	-2.03	-9.73	-1.32	4593.41
-10.60	-2.97	-10.67	-1.02	-10.69	-0.17	991.50
-11.94	-4.71	-11.97	-2.77	-11.97	-1.90	56454.18

For the matrix equation Ax = B at room temperature:

-1	0	0	0	0
4.45E+00	-1	0	0	0
0	2.66E-01	-1	0	0
0	0	3.81E-02	-1	0
0	0	0	7.07E-01	-1
1	1	1	1	1

B =

0.59133825	0.49091002	0.78688447	0.46089953	0.46089953
0.10303161	0.08553353	0.13710254	0.08030467	0.08030467
-0.5412042	0.38087754	0.61051234	0.3575936	0.3575936
-0.1438023	-0.8987978	0.16221801	0.09501548	0.09501548
-0.0054844	-0.0342787	-0.9938133	0.00362374	0.00362374
-0.003879	-0.0242446	-0.7029041	-0.997437	0.00256299

X =

orium constants kn	
273.15	**298.15**
31.36	6.09
1078.11	155.56
42.11	9.28
6.55	1.33
164.62	24.71

$$\begin{vmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1.08E-13 \end{vmatrix}$$

$$\begin{vmatrix} 4.9688E-14 \\ 8.6574E-15 \\ 3.8551E-14 \\ 1.0243E-14 \\ 3.9066E-16 \\ 2.7631E-16 \end{vmatrix} = \begin{vmatrix} \mathbf{1.34E+06} \\ \mathbf{2.33E+05} \\ \mathbf{1.04E+06} \\ \mathbf{2.76E+05} \\ \mathbf{1.05E+04} \\ \mathbf{7.43E+03} \end{vmatrix}$$

n in Gly-Wn	**216.65 K**	**273.15 K**	**298.15 K**	**RH (%)**
0	0.86	0.92	0.94	20
1	0.033	0.035	0.033	20
2	0.1	0.045	0.029	20
3	0.0035	0.0023	0.0015	20
4	0.000026	0.000018	0.000012	20
5	0.000011	0	0	20
0	0.52	0.69	0.76	50
1	0.049	0.066	0.067	50
2	0.4	0.21	0.15	50
4	0.033	0.027	0.02	50
5	0.00061	0.00053	0.00038	50
6	0.00063	0.00026	0.00013	50
0	0.21	0.36	0.46	100
1	0.04	0.069	0.08	100
2	0.63	0.45	0.36	100
3	0.11	0.11	0.095	100
4	0.0039	0.0045	0.0036	100
5	0.0081	0.0044	0.0026	100

# T =	**216.65**	**RH = 20**	**S = 2.9*10^6 molecules/cm3**	
# n[W]	**n[SA-W](/cm**	**n[SA-W]/nS**		
0		2.5*10^(6)		0.86
1		9.5*10^(4)		0.033
2		3.*10^(5)		0.1
3		1.*10^(4)		0.0035
4		7.5*10^(1)		0.000026
5		3.1*10^(1)		0.000011
#T =	**216.65**	**RH = 50**	**S = 2.9*10^6 molecules/cm3**	
#n[W]	**n[SA-W](/cm**	**n[SA-W]/nS**		
0		1.5*10^(6)		0.52
1		1.4*10^(5)		0.049
2		1.1*10^(6)		0.4
3		9.7*10^(4)		0.033
4		1.8*10^(3)		0.00061
5		1.8*10^(3)		0.00063
#T =	**216.65**	**RH = 100**	**S = 2.9*10^6 molecules/cm3**	

#n[W]	n[SA-W](/cm	n[SA-W]/nS
0	6.*10^(5)	0.21
1	1.1*10^(5)	0.04
2	1.8*10^(6)	0.63
3	3.1*10^(5)	0.11
4	1.1*10^(4)	0.0039
5	2.3*10^(4)	0.0081

#T =	273.15	RH = 20	S = 2.9*10^6 molecules/cm3
#n[W]	n[SA-W](/cm	n[SA-W]/nS	
0	2.7*10^(6)	0.92	
1	1.*10^(5)	0.035	
2	1.3*10^(5)	0.045	
3	6.6*10^(3)	0.0023	
4	5.2*10^(1)	0.000018	
5	1.*10^(1)	0	
#T =	273.15	RH = 50	S = 2.9*10^6 molecules/cm3
#n[W]	n[SA-W](/cm	n[SA-W]/nS	
0	2.*10^(6)	0.69	
1	1.9*10^(5)	0.066	
2	6.2*10^(5)	0.21	
3	7.8*10^(4)	0.027	
4	1.5*10^(3)	0.00053	
5	7.6*10^(2)	0.00026	
#T =	273.15	RH = 100	S = 2.9*10^6 molecules/cm3
#n[W]	n[SA-W](/cm	n[SA-W]/nS	
0	1.1*10^(6)	0.36	
1	2.*10^(5)	0.069	
2	1.3*10^(6)	0.45	
3	3.3*10^(5)	0.11	
4	1.3*10^(4)	0.0045	
5	1.3*10^(4)	0.0044	

#T =	298.15	RH = 20	S = 2.9*10^6 molecules/cm3
#n[W]	n[SA-W](/cm	n[SA-W]/nS	
0	2.7*10^(6)	0.94	
1	9.5*10^(4)	0.033	
2	8.4*10^(4)	0.029	

	3		4.5*10^(3)		0.0015
	4		3.4*10^(1)		0.000012
	5		4.8	0	
#T =		**298.15**	**RH = 50**	**S = 2.9*10^6 molecules/cm3**	
#n[W]		**n[SA-W](/cm**	**n[SA-W]/nS**		
	0		2.2*10^(6)		0.76
	1		1.9*10^(5)		0.067
	2		4.3*10^(5)		0.15
	3		5.7*10^(4)		0.02
	4		1.1*10^(3)		0.00038
	5		3.9*10^(2)		0.00013
#T =		**298.15**	**RH = 100**	**S = 2.9*10^6 molecules/cm3**	
#n[W]		**n[SA-W](/cm**	**n[SA-W]/nS**		
	0		1.3*10^(6)		0.46
	1		2.3*10^(5)		0.08
	2		1.*10^(6)		0.36
	3		2.8*10^(5)		0.095
	4		1.1*10^(4)		0.0036
	5		7.4*10^(3)		0.0026

Table of Materials Click here to access/download;Table of Materials;Shields2019MaterialsList.xlsx

Name of Material/ Equipment
Avogadro
Gaussian [09/16] Software
MOPAC 2016
OGOLEM Software
OGOLEM Software
OpenBabel
calcRotConsts.py
calcRotConsts.py
calcSymmetry.csh
combine-GA.csh
combine-QM.csh
gaussianE.csh
gaussianFreqs.csh
getrotconsts
getRotConsts-dft-lb.csh
getRotConsts-dft-lb-ultrafine.csh
getRotConsts-dft-sb.csh
getRotConsts-GA.csh
global-minimum-coords.xyz
make-thermo-gaussian.csh
README.docx
runogolem.csh
run-pw91-lb.csh
run-pw91-lb-ultrafine.csh
run-pw91-sb.csh
run-thermo-pw91.csh
similarityAnalysis.py
symmetry
symmetry.c
template-marcy.pbs
template-pw91.com

template-pw91-HL.com

thermo.pl

Company	Catalog Number
https://avogadro.cc	
http://www.gaussian.com/	
http://openmopac.net/MOPAC2016.html	
https://www.ogolem.org	
https://www.ogolem.org/	
http://openbabel.org/wiki/Main_Page	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	
(C) 1996, 2003 S. Patchkovskii, Serguei.Patchkovskii@sympatico.ca	
Shields Group, Department of Chemistry, Furman University	
Shields Group, Department of Chemistry, Furman University	

Shields Group, Department of Chemistry, Furman University
https://www.nist.gov/mml/csd/chemical-informatics-research-group/products-and-services/program-computing-ideal-gas

Comments/Description
Open-source molecular visualization program
Commercial ab initio electronic structure program
Open-source semi-empirical program
Genetic algorithm-based global optimization program
Genetic algorithm code for global optimization of chemical problems
Open-source cheminformatics library
Python script to compute rotational constants
Python script to compute rotational constants
Shell script to calculate symmetry number of a molecule given Cartesian coordinates
Shell script to combine energy and rotational constants from different GA directories
Shell script to combine energy and rotational constants from different QM directories
Shell script to extract Gaussian 09 energies
Shell script to extract Gaussian 09 vibrational frequencies
Executable to calculate rotational constants given a molecule's Cartesian coordinates
Shell script to compute rotational constants for a batch of large basis DFT optimized structures
Shell script to compute rotational constants for a batch of ultrafine DFT optimized structures
Shell script to compute rotational constants for a batch of small basis DFT optimized structures
Shell script to compute rotational constants for a batch of genetic algorithm optimized structures
Cartesian coordinates of global minimum structures of gly-(h2o)n, where n=0-5
Shell script to extract data from Gaussian output files and make input files for the thermo.pl script
Clarifications to help readers use the scripts effectively
Shell script to run OGOLEM
Shell script to run a batch of large basis DFT optimization calculations
Shell script to run a batch of ultrafine DFT optimization calculations
Shell script to run a batch of small basis DFT optimization calculations
Shell script to compute the thermodynamic corrections for a batch of DFT optimized structures
Python script to determine unique structures based on rotational constants and energies
Executable to calculate molecular symmetry given Cartesian coordinates
C code to determine the molecular symmstry of a molecule given Cartesian coordinates
Template for a PBS submit script which uses OGOLEM
Template Gaussian 09 input

Template Gaussian 09 input for ultrafine DFT optimization

Perl open-source script to compute ideal gas thermodynamic corrections

Click here to access/download;Rebuttal Letter;Responses to Editorial and Reviewers.docx

Responses to Editorial and Reviewers' comments:

Dear Editor:

We submit our revised paper, which we think is much improved, thanks to the peer review process. I have made all the major changes in red font in the paper, 60961_R0.Revised.docx, so it is clear where changes occur. Font color can easily be changed once peer review is finished. Please see comments below.

Sincerely,

George

Editorial comments:

The manuscript has been modified and the updated manuscript, **60961_R0.docx**, is attached and located in your Editorial Manager account. **Please use the updated version to make your revisions.**

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues. Check.
2. Please obtain explicit copyright permission to reuse any figures from a previous publication. Explicit permission can be expressed in the form of a letter from the editor or a link to the editorial policy that allows re-prints. Please upload this information as a .doc or .docx file to your Editorial Manager account. The Figure must be cited appropriately in the Figure Legend, i.e. "This figure has been modified from [citation]." No reused figures.
3. Please provide at least 6 keywords or phrases. Done
4. Please add a one-line space between each of your protocol steps. Done
5. Please revise the text in Protocol to avoid the use of any personal pronouns (e.g., "we", "you", "our" etc.). Done
6. Step 6.1: Please ensure that all text is written in the imperative tense. Done
7. Step 6.2: Please ensure that all text is written in the imperative tense. Done

Reviewers' comments:

Reviewer #1:

Manuscript Summary:

The manuscript, "Computation of Atmospheric Concentrations of Molecular Clusters from ab initio Thermochemistry," by Odbadrakh et al, describes a procedure for generating glycine/water clusters with varying numbers of waters, and then finding the relative concentrations of each cluster. This procedure may be applied to other types of systems to generate novel clusters for a range of applications. The work is very thoroughly described, with rationales for choices made at each step. The procedure is easy to follow and done in such a way that the reader can understand the underlying principles and make changes/substitutions as desired. The work should be of broad interest and will be an excellent contribution to the field.

Major Concerns:
None

Minor Concerns:
On line 250, I question the use of "outputted" as the verb. I would use simply "output" or perhaps "produced." Changed to “produced” I would appreciate the addition of information on the active and disk memory used for each step of the procedure. We added these sentences in the first paragraph of the Representative Results section: The disk and memory usage by the GA code (OGOLEM) and semi-empirical codes (MOPAC) are very small by modern computer resource standards. The overall memory and disk usage for OGOLEM/MOPAC depends on how many threads one wants to use, and even then the resource usage will be small compared to the capabilities of most HPC systems. The resource needs of the QM methods depend on the size of the clusters and the level of theory used. The advantage of using this protocol is that one can vary the level of theory to be able to calculate the final set of low energy structures, keeping in mind that usually faster calculations lead to more uncertainty in accuracy of the results.

Reviewer #2:

Odbadrakh and co-workers report a computational protocol for studying the formation of atmospheric molecular clusters. The hydration of glycine, with up to five water molecules is used as an example. The protocol addresses some of the fundamental issues in studying atmospheric cluster formation such as the difficulty of locating the global minimum and the choice of electronic structure method. The paper very well presents the protocol and I believe it could be of great use in teaching.

The manuscript is well written, however, there are some exaggerated claims about the direct usefulness of the protocol in relation to atmospheric chemistry and some errors that need to be corrected before I can recommend publication in Journal of Visualized Experiments.

Comments:

As a criteria for the Journal of Visualized Experiments is that "Previously published protocols must be cited properly" the author should comment on and cite the following papers that report similar protocols:

Elm et al, J. Phys. Chem. A 2015, 119, 8414-8421.
Kubecka et al, J. Phys. Chem. A 2019, 123, 6022-6033.

We have added the following paragraphs with appropriate references:

Different protocols have been developed over the years to get the structure and thermodynamics of atmospheric hydrates at a high level of theory. These protocols differed in the choice of (i) configurational sampling method (ii) nature of low-level method used in the configurational sampling (iii) the hierarchy of higher level methods used to refine the results in the subsequent steps.

The configurational sampling methods included chemical intuition, random sampling, to molecular dynamics (MD), basin hopping (BH), and genetic algorithm(GA). The most common low-level methods employed with these sampling methods are force fields or semiempircal models such as PM6, PM7 and SCC-DFTB. These are often followed by DFT calculations with increasingly larger basis sets and more reliable functionals from the higher rungs of Jacob's ladder. In some cases, these are followed by higher level wavefunction methods such as MP2, CCSD(T), and the cost efficient DLPNO-CCSD(T)

Kildgaard et al. developed a systematic method where water molecules are added at points on the Fibonacci spheres around smaller hydrated or unhydrated clusters to generate candidates for larger clusters. Unphysical and redundant candidates are removed based on close contact thresholds and root-mean-square distance between different conformers. Subsequent optimizations using the PM6 semiempircal method and a hierarchy of DFT and wavefunction methods are used to get a set of low energy conformers at a high level of theory.

The artificial bee colony (ABC) algorithm is a new configurational sampling approach that has recently been implemented by Zhang et al. to study molecular clusters in a program called ABCluster. Kubecka et al. used ABCluster for configurational sampling followed by low-level reoptimizations using the tight-binding GFN-xTB semi-empirical method. They further refined the structures and energies using DFT methods followed by final energies using DLPNO-CCSD(T).

Line 35-36: "The Gibbs free energy of the minimum energy structure can be used to predict atmospheric concentrations of the cluster under a variety of conditions such as altitude and humidity level."

It does not appear that the authors explore the effects of altitude or humidity level in the manuscript. It should at least be mentioned in the text how this is performed if you state it in the abstract.

We changed the wording to "temperature and pressure" instead of "altitude and humidity level". Temperature decreases with altitude, and humidity decreases. We make it clear we are talking about a specific temperature and pressure later in the text.

Line 56-57: "... starting from an isolated glycine molecule ..."

Why was glycine chosen as the system of choice for the protocol? From a pedagogical point of view it seems like a poor choice, as to the best of my knowledge there are no ambient measurements of glycine concentrations and hence no indications that glycine should be involved in nucleation. A more logical choice would be sulfuric acid, as there is an abundance of literature to compare with and it is actually believed to be involved in new particle formation. The authors need to motivate, from an atmospheric perspective, why glycine was chosen.

Besides molecules of importance to atmospheric chemistry, we are interested in the role of water in the formation of large biological molecules from smaller constituents in pre-biotic environments. The challenges encountered and tools needed to address those research questions are very similar to those involved in the study of atmospheric aerosols. Therefore, the choice of glycine hydration as an example should be acceptable. We have added a few sentences describing this in the introduction, along with appropriate references.

Furthermore, the number of potential vapor molecules that could be involved in secondary aerosol formation has been increasing steadily over the years from just sulfuric acid and water to include the likes of ammonia, amines, diamines, volatile organics, highly oxidized molecules, guanidine, …, etc. And it is not totally inconceivable that small amino acids like glycine could play a part.

Line 60-61: "This gives us insight into the distribution of different molecular clusters in the atmosphere, leading to more accurate parameters in global climate models."

It is a too strong statement that cluster distributions directly yield more accurate parameters in climate models. Please rephrase.

We replaced that sentence with this statement: A small number of these subnanometer molecular clusters grow into a metastable critical cluster (1-3 nm in diameter) either by adding other vapor molecules or coagulating on existing clusters. These critical clusters have a favorable growth profile leading to the formation of much larger (up to 50-100 nm) cloud condensation nuclei (CCN) which directly affect the precipitation efficiency of clouds as well their ability to reflect incident light. Therefore, having a good understanding of the thermodynamics of molecular clusters and their equilibrium distributions should lead to more accurate predictions of the impact of aerosols on the global climate.

Line 84-85: "The set of GA-optimized global minima are taken as the starting geometries 1 for a series of screening steps which ultimately gives us the global minimum energy structure."
You are never certain that the global minimum structure is obtained. With the presented protocol you might have a good guess for (and be close in energy of) the global minimum.

We changed this to, "The set of GA-optimized global minima are taken as the starting geometries for a series of screening steps which leads to a set of low lying minimum energy structures."

Line 248-250: "One may take a Boltzmann average of this set of structures at each value of n to obtain a more accurate representation of the Gibbs free energy change of hydration"

Taking the Boltzmann average is an incorrect approach to account for multiple isomers. This will lead to an increase in the Gibbs free energy corresponding to an effective decrease in the number of available microstates. The existence of multiple isomers will always result in an increase in the number of available microstates, which leads to a decrease in Gibbs free energy. I suggest the authors simply remove this sentence. Otherwise I suggest the authors look into the following paper (L. Partanen, J. Phys. Chem. A, 2016, 120, 43, 8613-8624) for a more rigorous method to account for multiple isomers.

Agreed. We removed this sentence.

Line 251, Figure 3: It would be beneficial to further comment on the validity of the located "global" minima structures in Figure 3. One of the most powerful tools in modelling atmospheric molecular clusters is visual inspection and rationalizing whether what you cal- culated makes sense. By inspecting Figure 3, I would question that you have located the correct global minima structures for the (glycine)(H2O) and (glycine)(H2O)4 clusters. For in- stance how do your (glycine)(H2O) cluster compare in free energy to the one where the water molecule is residing at the carboxylic acid moiety? Same question with the (glycine)(H2O)4 cluster: Why is the water molecules interacting with the amino group and not the carboxylic acid group? Inspection of the calculated free energies in Table 3 gives a further indication that the (glycine)(H2O)4 cluster might not be the global minimum one as it is higher in free energy compared to the trihydrate.

Originally we weren't planning to present the global minima, because the point of this paper is HOW we generate structures and determine their energies, not to produce an exhaustive search of glycine-water clusters—but we realized this was confusing so we've changed it to the global minima which are in line with the reviewers insights.

Also what structure was used for the glycine monomer? If it was simply inserted from the "build peptide" function in Avogadro, it might be stuck in an unfavourable higher energy local minimum.

We checked and the reviewer was right, we weren't using the global minima. We added this sentence in 1.1 of the protocol: Please note that if the molecule has significant conformational flexibility, as glycine does[55], it is critical to perform conformational analysis to identify the global minimum structure and other low-lying conformers.

Line 246-247: "... and are assumed to be accurate for the purpose of this paper ..."

While I agree that the level is sufficient for the purpose of presenting the protocol, it should be specified that this level of theory might overestimate the binding energies of the clusters.

There is no evidence to suggest that PW91/6-311++G** consistently underestimates or overestimates the binding energy of these clusters. Its ability to predict binding energies relative to MP2/CBS and [DLPNO-]CCSD(T)/CBS estimates and experiment shows a lot of fluctuations. The same is true of most other density functionals.

We have added the above two sentences and four references to back this statement up.

Line 258, Tables: Related to the comment above: The free energy numbers reported in the

Tables seem quite too negative at least for the mono- and dihydrates. How do these values compare to other literature? One important aspect when applying a quantum chemical protocol is also to be critical about the output. So it would benefit if the authors conveyed such critical thinking in their manuscript.

This was fixed by using the low energy glycine monomer. Thank you Dr. Referee!

Also it is stated in the figure captions that electronic energies are in hartrees and all other are in kcal/mol. The entropy contribution S should be in cal/mol·K.

We agree that makes the most sense.

Line 274-301, page 6: There seems to be some errors in the calculation of the thermochemistry and presentation in the Tables. For instance, in Table 3, the columns for SB, LB, LB-UF and CBS are identical. Please check all the numbers and the calculated thermochemistry. This is left over from a default we have for using MP2 and doing CBS calculations, so we have removed it.

Furthermore, it is not very easy to follow what is going on in Table 1-3. For instance, the energies are in hartrees, but the dH and dG contributions are in kcal/mol. It is easier to convey your message if you use the same units. Also why do you include the dHvib and dSvib terms? They are not used for anything.

Line 280-281: "Although not included in this protocol, we can also calculate the complete basis set (CBS) limit electronic energy, resulting in the CBS column." This is only done for MP2 and CCSD(T) calculations, not for DFT, so we removed it.

How is the extrapolation to the basis set limit performed? Complete basis set extrapolation using the 6-31+G* and 6-311++G** basis sets seems like extremely bad practice as the Pople style basis sets show no clear basis set convergence behaviour. I highly recommend that the inclusion of ΔECBS is removed from the protocol as it promotes bad practice that might not be apparent for non-specialists/students using the protocol. See above.

Line 292: Please carefully check that you are not applying the ZPE twice in this equation.

We aren't.

Line 384-385: "One may choose a better density functional, such as M06-2X and wB97X-V, ..."

Please justify the claim and elaborate why these functionals should be better than PW91.

We've added this sentence:

In the hierarchy of functionals, the performance generally improves upon going from generalized-gradient approximation (GGA) functionals like PW91 to range-separated hybrid functionals like wB97X-D and meta-GGA hybrid functionals like M06-2X.

Line 389-390: "Quantum chemical methods such as the MPn and CC approaches do converge systematically towards an accurate value, ..."

It is incorrect that MPn systematically converge towards an accurate value. The MPn series have been shown to be divergent (Olsen et al, J. Chem. Phys.105, 5082 (1996) and Leininger et al, J. Chem. Phys, 112, 9213 (2000)) and there is no guarantee that higher order MPn methods are more accurate than MP2.

We added this sentence to be clearer:

Energies calculated using wavefunction methods like MP2 and CCSD(T) in conjunction with correlation consistent basis sets of increasing cardinal number ([aug-]ccpV[D,T,Q,...]Z) converge towards their complete basis set limit systematically.

Line 327-329: "From this small set of Gly(H2O)n=1-5 data, we may conclude that given a glycine concentration of several million molecules per cubic centimeter, glycine cannot act as a cloud condensation nucleus since hydration stops at n > 3 water molecules."

For non-specialist this is a slightly misleading sentence, as it gives the impression that one can study the hydration of a single chemical specie and directly make claims about the compounds involvement in forming CCN. This is certainly not the case as these are two very different size scales. For instance, for aerosol particles to "activate" and act as CCN they need to be around 50 nm, so it is no surprise that a single glycine molecule cannot grow into a CCN. Essentially, no single molecule can act as CCN.

This is a good point. We agree and have removed this sentence to keep the focus on what we know from the calculations.

Line 366-367: "We used the default values of [Gly]0 = 2.9 × 10^6 cm−3 and [H2O] = 7.7 × 10^17 cm−3 "

Please specify where these values are coming from.

We added the three references where the [GLY]0 concentration comes from. The water value is for 100% relative humidity at 298 K, and we added some words to make that clearer along with the reference.

Line 429-430: "Our current applications of this protocol are exploring the importance of amino acids in the atmosphere on both the prebiotic and current Earth.

Again this is a too strong statement as the thermochemistry at the PW91/6-311++G** can hardly be trustworthy. It gives the impression that you can apply this level of theory and directly infer effects on the global scale.

> I think the main purpose of this paper and its publication in JoVE is to demonstrate a clear protocol others could use to study other molecular clusters. We changed the sentence to this:
>
> "Our current applications of this protocol are exploring the importance of amino acids in the early stages of aerosol formation in the current atmosphere and in the formation of larger biological molecules in prebiotic environments."

Line 428-429: "possibly the use of machine-learning methods for faster computations of electronic and vibrational energies"

While I certainly agree that machine learning methods will evolve such calculations in the future it seems too vaguely described to just briefly mention here. If you want to mention machine learning as a perspective please further elaborate on which approach you want to take. It seems more like buzzword dropping here as it is extremely vaguely formulated how it should be useful.

Changed "machine-learning" to "newer".

Minor comments:
Line 139-140: where n is be the number of molecules → where n is the number of water molecules
Line 84: global → local
Line 385 and 394: Moller-Plesset → Møller-Plesset

Changes made.

Reviewer #3:

T. Odbadrakh and co-workers have developed a protocol for configurational sampling of hydrogen-bonded molecular clusters, and presented its application to the glycine - water system. This study nicely complements recently published work on much more strongly bound clusters, and the presented toolbox, together with the very detailed and user-friendly instructions, will be very useful to researchers in the field. I thus warmly recommend publication subject to a few minor revisions.

-First, I'd like to draw the authors' attention to a very recent study on a very similar subject: "Configurational Sampling of Noncovalent (Atmospheric) Molecular Clusters: Sulfuric Acid and Guanidine" by J. Kubecka and co-workers, in J. Phys. Chem. A 2019, 123, 6022–6033. (https://pubs.acs.org/doi/pdf/10.1021/acs.jpca.9b03853). The two approaches have much in common, for example the stepping from a semi-empirically generated set of input structures via "low-level" DFT (with smaller basis sets and/or looser criteria), to "high-level" DFT (with larger basis set and/or tighter criteria). It is quite encouraging to see different groups converging completely independently on what seems to be a reasonable approach for obtaining "educated guesses" for the global minimum structures of atmospheric clusters. I would even argue that all the differences in the actual quantum chemical protocols, for example the exact choice of density functional and basis set, the inclusion (or not) of higher-level energy corrections, and so on, are mostly a matter of taste (as well as of the desired level of accuracy). The two crucial issues, and indeed the defining and critical features of any configurational sampling protocol, are in my mind

We make reference to Kubecka's paper earlier in the paper. In these sentences:

Kubecka et al.'s the ABCluster method for configurational sampling followed by low-level reoptimizations using the tight-binding GFN-xTB semi-empirical method. They further refined the structures and energies using DFT methods followed by final energies using DLPNO-CCSD(T).

1)the initial method for sampling the potential energy surface (step 2 in the authors' protocol), and

2)the criteria used to identify unique clusters (i.e. the criteria used to determine when two cluster structures are NOT unique, and eliminate one of them).

To address these two issues, we added the following to the discussion, with appropriate references:

Two crucial issues for any configurational sampling protocol are the initial method for sampling the potential energy surface and the criteria used to identify each cluster. We have made extensive use of a variety of methods in our previous work. For the first issue, the initial method for sampling the potential energy surface, we have made the choice of using GA with semiempirical methods is based on these factors:

1. Configurational sampling using chemical intuition, random sampling, and molecular dynamics (MD), fail to find putative global minima regularly for clusters larger than 10 monomers, as we observed in our studies of water clusters. We have successfully used basin hopping (BH) to study the complex PES of $(H_2O)_{11}$, but it required the manual inclusion of some potential low energy isomers the BH algorithm did not find.
2. A comparison of the performance of BH and GA in finding the global minimum of water clusters, $(H_2O)_{n=10-20}$ demonstrated that GA consistently found the global minimum faster than BH.
3. GA as implemented in OGOLEM and CLUSTER is very versatile because it can be applied to any molecular cluster and it can interface with a vast number of packages with classical force field, semiempirical, density functional, and ab initio capabilities.

4. The choice of PM7 is driven by its speed and reasonable accuracy. Virtually any other semi-empirical method would have significantly higher computational cost.

As for the second issue, we have explored using different criteria to identify unique structures ranging from electronic energies, dipole moments, overlap RMSDs and rotational constants. Using dipole moments proved difficult because both the dipole moment components were dependent on the molecule's orientation and the total dipole moment was very sensitive to geometry differences in such a way that it was difficult to set thresholds determining is structures are the same or unique. A combination of electronic energies and rotational constants proved to be most useful.

The current criteria for deeming two structures unique is based on an energy difference threshold of 0.10 kcal/mol and rotational constant difference of 1%. Therefore, two structures are considered different if

1. their energies differ by more than 0.10 kcal/mol (~0.00015 a.u.) AND
2. any of their three rotational constants (A, B, C) differ by more than 1%.

Substantial internal benchmarks over the years found these thresholds to be reasonable choices.

I suggest that the authors spend a bit more time discussing the choices they have made for these two points, and compare them to other approaches proposed in the literature. (Note: in addition to the Kubecka et al study, the ABCcluster approach has recently become popular e.g. with groups in China working with atmospheric clusters, see for example prof. Xiuhui Zhang's work, e.g. at https://www.researchgate.net/profile/Xiuhui_Zhang. So probably it makes a good comparison for the present approach). What are the advantages and disadvantages for example of OGOLEM vs ABCcluster (for point 1), or rotational constants versus dipole moment, or radius of gyration, etc (for point 2)?

These two points are addressed above.

Could we expect the same method to work equally well for fairly weakly bound clusters (such as the glycine-H2O clusters in this study) as for very strongly bound clusters (such as the H2SO4-guanidine clusters in Kubecka et al)? Or would one approach work better for more weakly bound clusters, and the other for stronger clusters? How would the approach presented by the authors work for systems containing strong acids and bases (and thus potentially a mixture of different proton transfer states)? It seems that like ABCcluster, the proposed approach in its current form only works for rigid molecules - how could the sampling protocol be amended to account for floppy molecules with very many conformations already in the "monomers" (such as low-volatility organics contributing to CCN formation)?

Our configurational sampling approach and screening methodology has been applied to very weakly bound clusters such as polyaromatic hydrocarbons complexed with water as well as strongly bound ternary sulfate hydrates containing ammonia and amines.

For clusters where there are different protonation states to be considered, the best approach is to run various GA calculations, each starting with monomers in different protonation states. This ensures that structures with different protonation states are carefully considered. However, the low-level DFT calculations often allow protonation states to change during the course of the geometry optimization, thereby yielding the most stable protonation state regardless of the starting geometry.

Our GA configurational sampling methods should work well even for floppy molecules as long as the GA codes are interfaced with general, non-parameterized methods that allow the monomers to adopt different configurations during the course of the GA run. For example, interfacing GA with PM7 would allow monomers' structures to change, but if their bonds break as would happen when protonation states change, the structures may get discarded as unacceptable candidates.

In the future, it would be extremely interesting to do a "cross-comparison" of sampling methods on the cluster datasets already published in the literature, including a variety of different system types (e.g. weakly and strongly bound), and perhaps also testing some "notoriously difficult" structures (such as floppy polyfunctional organics). However that would be a good subject for a (long!) paper on its own, and far beyond the scope of the present manuscript - I'm just raising the point as a something for the authors to think about.

Such a study would certainly be very interesting and informative to a large community. There may comparisons of the performance of different global optimization methods for noble gas clusters, metal clusters and water clusters in the Cambridge Cluster Database (http://www-wales.ch.cam.ac.uk/CCD.html).

-As I implied above, the particular choice of quantum chemical methods in this study is quite acceptable (one could always quibble about functionals etc, but this has been done to death in the literature already), and the authors do a good job of discussing the various strengths and limitations of the approaches. One minor issue that is, however, missing from their discussion are the recently proposed computationally cheap options for treating the errors arising from low-frequency modes, such as the quasi-harmonic approximation (see e.g. references 55-57 in the Kubecka et al study reference above). This would seem to be an attractive option for correcting some of the worst deficiencies of the harmonic PES, without the need for expensive calculations of third and fourth derivatives. (Note: I'm not suggesting the authors actually perform these calculations for their glycine-H2O system, just that they mention this option in their discussion).

We have added this paragraph:

We have considered different ways of correcting the shortcomings of the harmonic approximation, especially those arising from low vibrational frequencies. Incorporating the quasi-harmonic approximation into the current methodology is not difficult. However, there are still questions about the quasi-harmonic method, especially when it comes to the cutoff frequency below which it will be applied. Also, there are not rigorous benchmarking works examining the reliability of the quasi-RRHO approximation even though conventional wisdom suggests it should be an improvement over RRHO approximation.

-The authors seem to have slightly misunderstood the physics (and/or the definitions) related to cloud condensation nuclei (CCN) formation. By definition, CCN are particles on which water vapour can condense. This process is usually called CCN activation, and it requires (again by definition) the saturation ratio of water to be above 1. (Note: CCN activation is essentially heterogeneous nucleation of PURE water on top of a surface.) Due to the Kelvin effect (i.e. the saturation vapor pressure above a curved surface is higher than that above a flat surface), larger particles are easier to activate (or in other words, they require a lower degree of supersaturation). For the supersaturations encountered in the atmosphere (rarely much more than a percent or so, i.e relative humidity of 101%), the threshold diameter for CCN activation is typically on the order of 100 nanometers (with some variation due to the chemical composition of the particle). The number of molecules even in the smallest CCNs is thus always at least in the tens of millions, usually much more. Single molecules, or even small clusters, NEVER EVER act as CCN. Or in other words, the supersaturations required for them to do so are ludicrously high, and never encountered in the atmosphere. Hygroscopic molecules, such as H2SO4, or glycine in this example, may well hydrate in the atmosphere, i.e. at atmospherically representative relative humidities (even below 100%), they may bind a few water molecules to themselves - for example the authors find that at some relative humidity (which could, by the way, be more clearly stated!), glycine binds on average two water molecules. This is in itself a reasonable result. However this "micro-hydration", where one molecule binds a handful (typically 1-5) of water molecules, is NOT analogous to CCN activation, which is condensation of water (as in: PURE water) on top of a pre-existing surface. (The final product of CCN activation is a cloud droplet, which contains many orders of magnitude more water than anything else. Or in yet other words, the water molecules involved in CCN activation are almost all bonded just to other water molecules, only a small minority actually bind to the molecules at the particle surface.) Thus the statement that "glycine cannot act as a cloud condensation nuclei" is certainly true, but about as relevant and novel as noting for example that "one glycine molecule is not massive enough to collapse into a black hole". (However, had the calculations suggested otherwise, it would certainly have proved that something is seriously wrong with the methodoIogy, so I guess this could be seen as a useful sanity check - but certainly not more than that.) I would suggest that the authors somewhat revise their discussion of CCN (both the sentence on glycine, and the more general sentence in the introduction) to avoid giving readers misleading impressions about CCN activation.

Agreed with the referee on all points. We removed any references to CCNs and focused instead on (1) the early stages of aerosol formation and (2) what the hydrate distribution says about glycine. It can be something like:

-The first two sentences in the abstract (especially the second sentence) is a bit hard to parse. I THINK I understand what they mean, but many readers surely will not. Please reformulate. Changed.

Bio of Tuguldur T. Odbadrakh

Click here to access/download;Supplemental File (Figures, Permissions, etc.);bio-odbadrakh.docx

Tuguldur T. Odbadrakh
Department of Chemistry
Furman University

Tuguldur Odbadrakh attended the West Virginia University for his B.S. in Chemistry, where he worked in the organic synthesis research group of Professor Xiaodong Shi. He contributed to two publications on one-pot synthesis methods. He then joined the computational chemistry research group Professor Kenneth Jordan at the University of Pittsburgh. Here, he worked on the Drude oscillator model of dispersion interactions between electrically neutral atoms and molecules. He also worked on the study of charged gas phase molecular clusters through vibrational spectroscopy, specifically the effects of hydration on an excess proton. Upon completion of his Ph.D. studies, Dr. Odbadrakh joined the research group of Professor George Shields at Furman University where he continues his study of gas phase molecular clusters while also contributing to the MERCURY Consortium for undergraduate computational chemistry research.

Bio of Ariel G. Gale

Click here to access/download;Supplemental File (Figures, Permissions, etc.);bio-gale.docx

Ariel G. Gale
Department of Chemistry
Furman University

Ariel G. Gale is a senior undergraduate student at Furman University's Department of Chemistry. She is majoring in Chemistry and minoring in women's and gender studies. She went to high school in Knoxville, TN before joining the research group of Professor George C. Shields. Since then she has focused her research efforts on understanding the rise of oligopeptides before life began on Earth and has conducted two Summers of full-time research and two years of part-time research. Her academic efforts earned her the Barry M. Goldwater scholarship as well as a Beckman scholarship. Ariel plans to attend graduate school to pursue her Ph.D. after her studies at Furman University.

Bio of Benjamin T. Ball
Click here to access/download;Supplemental File (Figures, Permissions, etc.);bio-ball.docx

Benjamin T. Ball
Department of Chemistry
Furman University

Benjamin T. Ball attended high school at Mountain Heritage High School in Burnsville, North Carolina. He is currently a senior at Furman University majoring in Environmental Chemistry at the Department of Chemistry. He joined the research group of George C. Shields in the summer of 2018 and has been a researcher since then. He conducted full-time research for two Summer terms and part-time research during the Fall and Spring terms. His work focuses on the thermodynamics of the hydration of sulfuric acid and various amino acids in the atmosphere. He plans to attend graduate school to pursue a Ph.D. after his time at Furman University.

Click here to access/download;Supplemental File (Figures, Permissions, etc.);bio-temelso.docx

Berhane Temelso

College of Charleston

Berhane Temelso is a senior HPC system administrator and research computing consultant at College of Charleston (CofC). He is a computational chemist by training and he has worked as a research scientist and HPC system administrator with George C. Shields and MERCURY consortium at Furman University, Bucknell University and Armstrong Atlantic University for many years prior to moving to CofC. He received his Ph.D. in computational chemistry from Georgia Institute of Technology in Atlanta, GA and B.A. in physics from Berea College in Berea, KY. His Ph.D. work explored the ability of the most rigorous first-principles computational methods to reproduce molecular properties derived from experiment.

Prior to moving to CofC, Dr. Temelso managed MERCURY consortium's HPC clusters named Skylight and Marcy as a system administrator. He provided technical research support to MERCURY users and promoted the use of HPC in chemistry and other fields. In March 2017, he was also named an inaugural Foresight Institute Fellow in computational chemistry.

Dr. Temelso's research mainly focuses on the application of efficient computational methods to understand the structure and dynamics of hydrogen-bonded systems ranging from water clusters to atmospheric aerosols. He collaborates with experimental groups to solve interesting problems like the structure of small water clusters and the formation rates of sulfate atmospheric aerosols whose cooling effect on the global climate is significant, but poorly understood.

Bio of George C. Shields

Click here to access/download;Supplemental File (Figures, Permissions, etc.);GCS Bio.docx

An award winning scientist and a national leader in undergraduate research, Shields has collaborated with 118 undergraduates in meaningful projects in the fields of computational chemistry, structural biochemistry and science education. His current research involves using computational methods to gain insights into biochemistry and environmental chemistry.

Since 1990, Shields has received approximately $5.6 million in external research grants from many foundations and funding agencies, including the National Science Foundation and National Institutes of Health. He has published 88 scientific articles, four educational papers, four book chapters, and a book, including 56 scientific papers with 57 undergraduates working in his research group since 1991. His undergraduates have received 34 national awards, and 85% of his research alumni have matriculated to graduate or professional programs of study.

A native of Marcellus, NY, Shields received his bachelor's, master's, and doctoral degrees from the Georgia Institute of Technology. His postdoctoral research on protein-DNA interactions at Yale University and the Howard Hughes Medical Institute was conducted in the laboratory of Professor Thomas Steitz, the 2009 Chemistry Nobel Laureate. Shields received the 2015 American Chemical Society (ACS) award for Research at an Undergraduate Institution, and the 2018 Transformational Research and Excellence in Education (TREE) Award from Research Corporation. He currently serves on the editorial advisory board of the ACS Journal of Physical Chemistry, and on the Science and Software Advisory Board of the Molecular Sciences Software Institute (MolSSI). He is a member of Omicron Delta Kappa, Sigma Xi, Phi Beta Kappa, and is a Research Corporation Cottrell Scholar. He was elected a Fellow of the American Association for the Advancement of Science in 2019. He was a Council on Undergraduate Research (CUR) Chemistry Councilor for nine years, and served on the executive board of CUR from 2015-2018. Since 2000, Shields has served as director of MERCURY, the **M**olecular **E**ducation and **R**esearch **C**onsortium in **U**ndergraduate Computational Chemist**ry**, a collaboration of 30 undergraduate research teams at 27 different institutions, which is a major driver for increasing diversity in the field of chemistry.

Supplemental File (Figures, Permissions, etc.)

Click here to access/download;Supplemental File (Figures, Permissions, etc.);README.docx

Clarifications/README

To help users navigate the scripts and make effective use of them, below are some tips and clarifications.

- The scripts assume all the data they will operate on is located in the same directory. To access them from anywhere,
  - move the whole directory with the scripts to somewhere that is in your $PATH
  - in each tcsh script (`*.csh`), update `setenv SCRIPTS_HOME pwd` to `setenv SCRIPTS_HOME __LOCATION_OF_SCRIPTS__`
- The scripts have been used on the following operating systems
  - CentOS 6/7, although they should work on most GNU Linux systems
  - Mac OS X 10.10+, assuming it has XCode and most Linux utilities
- The scripts assume you will be running them on an HPC system using
  - PBS batch queuing system
  - all the scripts and necessary software are accessible to the compute nodes via a shared filesystem
- The scripts assume users have the following software
  - OpenBabel
    - only version 2.4.x was tested
  - tcsh shell
    - only version 6.18.x was tested
    - if the location of your ‘tcsh’ is different from `/bin/tcsh`, you would need to update the scripts with the proper path to ‘tcsh’
  - Python
    - only versions 2.6 and 2.7 were tested

  - Gaussian 09 or 16
    - g09rev[B,D] and g16rev[A-B] were tested
  - Ogolem
    - only the “classic” (x - 2016) version is tested
    - requires Java
  - all packages that interface with Ogolem need to be available. For example,
    - MOPAC to run PM7
    - Orca to run hf3c
    - DFTBplus to run scc-dftb
    - paths to the locations of all these packages need to be specified in the `runogolem.csh` script

- To calculate the symmetry of molecules, one would likely need to compile the attached ‘`symmetry.c`’ file by entering ‘`gcc -o symmetry symmetry.c`’. Since the code was written in 2002, you may get a lot of warning messages if you compile it using recent versions of GNU gcc.

- Rotational constants are calculated using the ‘`rot_conts.py`’ script.
  - if your molecule has elements aside from the 10 or so first and second row elements included in the script, you would need to add the atom and its atomic mass in the script.
  - this Python script is slow for processing a large number of files. If users want to compile a much faster C version, they can ask the authors for a more efficient C code.
- The scripts write intermediate data to /tmp assuming it exists and that the user has permission to write to files in that directory. If that is not the case, please change ‘/tmp’ to a a location you have write permissions to.

Supplemental Coding Files

Click here to access/download

Supplemental Coding Files

make-thermo-gaussian.csh

Click here to access/download
Supplemental Coding Files
calcRotConsts.py

Click here to access/download
Supplemental Coding Files
gaussianE.csh

Click here to access/download
Supplemental Coding Files
gaussianFreqs.csh

Click here to access/download
Supplemental Coding Files
getrotconsts

Click here to access/download
Supplemental Coding Files
getRotConsts-dft-lb-ultrafine.csh

Click here to access/download
Supplemental Coding Files
getRotConsts-dft-lb.csh

Click here to access/download
Supplemental Coding Files
getRotConsts-dft-sb.csh

Click here to access/download
Supplemental Coding Files
getRotConsts-GA.csh

Click here to access/download
Supplemental Coding Files
getsymmetry.csh

Click here to access/download
Supplemental Coding Files
run-pw91-lb-ultrafine.csh

Click here to access/download
Supplemental Coding Files
run-pw91-lb.csh

Click here to access/download

Supplemental Coding Files

run-pw91-sb.csh

Click here to access/download

Supplemental Coding Files

run-thermo-pw91.csh

Click here to access/download
Supplemental Coding Files
runogolem.csh

Click here to access/download
Supplemental Coding Files
similarityAnalysis.py

Click here to access/download
Supplemental Coding Files
symmetry.c

Click here to access/download
Supplemental Coding Files
thermo.pl

Cartesian coordinates of global minima

Click here to access/download
Supplemental Coding Files
global-minimum-coords.xyz

Click here to access/download
Supplemental Coding Files
combine-GA.csh

Click here to access/download
Supplemental Coding Files
combine-QM.csh

example coordinates of water

Click here to access/download
Supplemental Coding Files
water.xyz

sample OGOLEM input file

Click here to access/download
Supplemental Coding Files
ogolem-input-file.ogo

sample PBS scripts for OGOLEM

Click here to access/download
Supplemental Coding Files
ogolem-submit-script.pbs