Populations for 4-Way H-type Bonding

128.192.9.183/eln/lachele/2020/11/10/populations-for-4-way-h-type-bonding

November 10, 2020

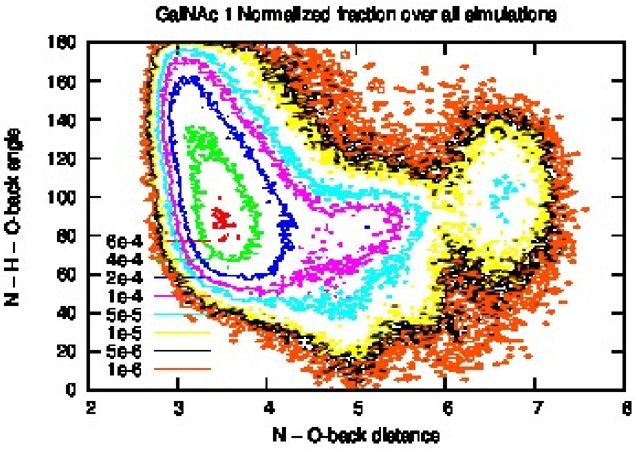
A reviewer asked for specific numbers for the populations of the 4-Way bifurcated-hydrogen bonding interaction. This post describes how I calculated those. It also provides a little back-story to further explain the process and conclusions.

Contents [hide]

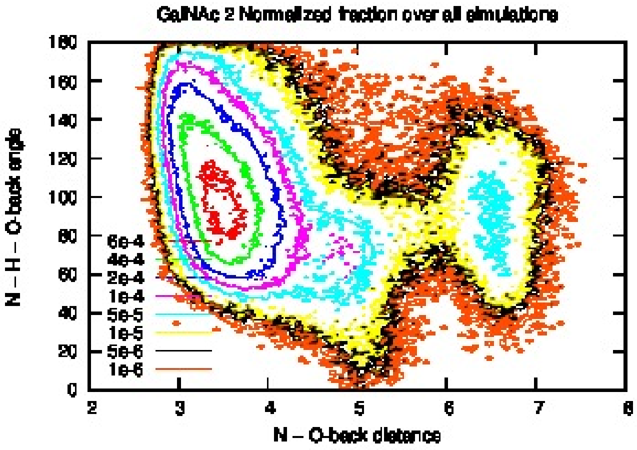
Back-Story: How I Found the Interaction

Previous literature (see paper) had suggested that a hydrogen bond was responsible for holding the GalNAc in a specific orientation. So, I went looking for a hydrogen bond. Specifically, I was looking for a hydrogen bond between the NAc's H2N (hydrogen attached to the nitrogen) and the peptide backbone's oxygen.

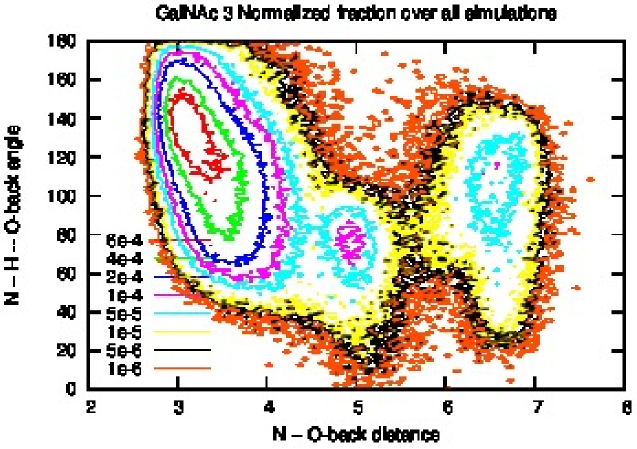
To do this, I plotted the N(NAc)-H(NAc)-O(backbone) angle versus the N(NAc)-O(backbone) distance. I used contour plots initially, with the contours representing relative population, and got these plots:



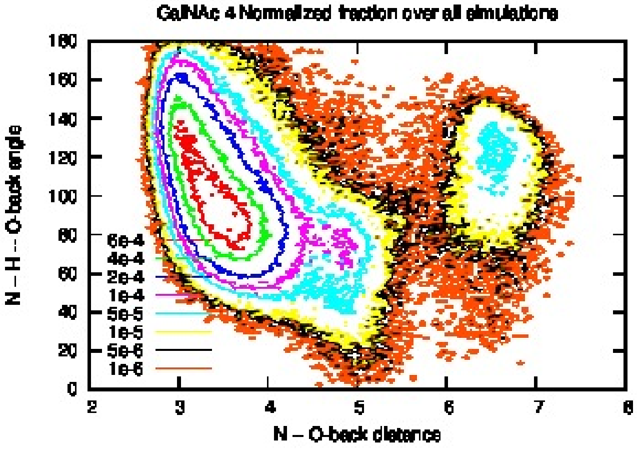
GalNAc 1 Normalized fraction over all simulations



GalNAc 2 Normalized fraction over all simulations



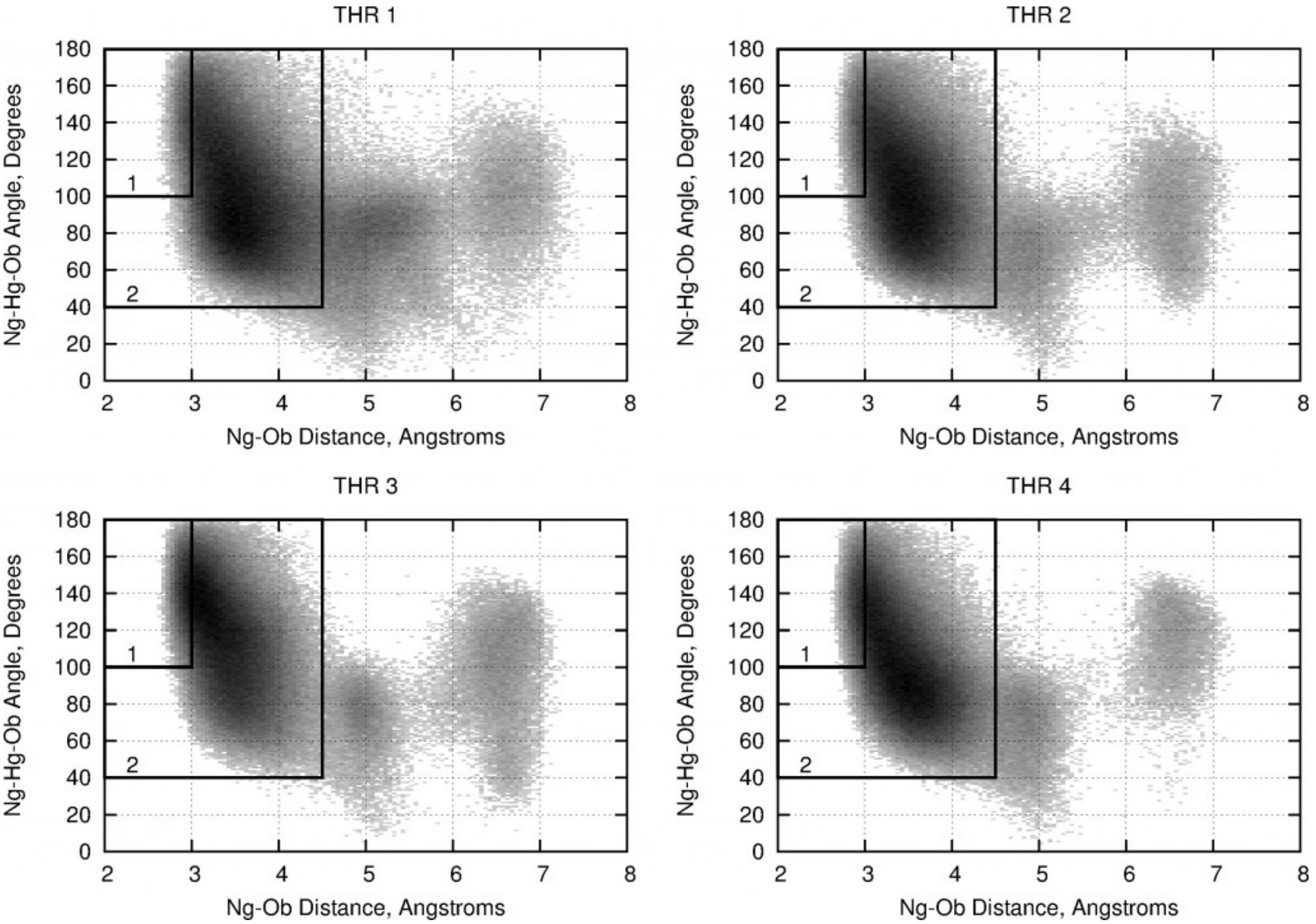
GalNAc 3 Normalized fraction over all simulations



GalNAc 4 Normalized fraction over all simulations

The most obvious feature is that, although there is motion, the overwhelming tendency is for the three atoms involved to be in a specific interaction.

But, the plots above don’t make it easy to explain what I mean by that, so I made the plots just below. These are the same data as above. The only difference is that instead of plotting populations as contours, I used a heat map, where the darker the gray, the greater the population. In these plots, I also added two rectangles:



Rectangle 1 encloses the geometries that fall into a standard definition of a hydrogen bond. Although H-bonding does occur, it is obviously not the dominant geometry. Rectangle 2 encloses the latter. Obviously, there was some reason that the Ng(NAc), Hg(NAc) and Ob(backbone) were being held in what appears to be an oddly-shaped but very stable state with the Ng-Ob distance being between about 2.5 and 4.5 Angstroms and the Ng-Hg-Ob angle being between about 40 and 180 degrees.

The next thing I did was to separate out all the MD frames with a geometry that falls within rectangle 2 in the plots. Because the four sites might be in different orientations, I separated based only on one site. I believe I used the third of the four threonine sites because its geometries in this area seemed more 'concentrated' and because it had more frames in the standard H-bond geometry than the others.

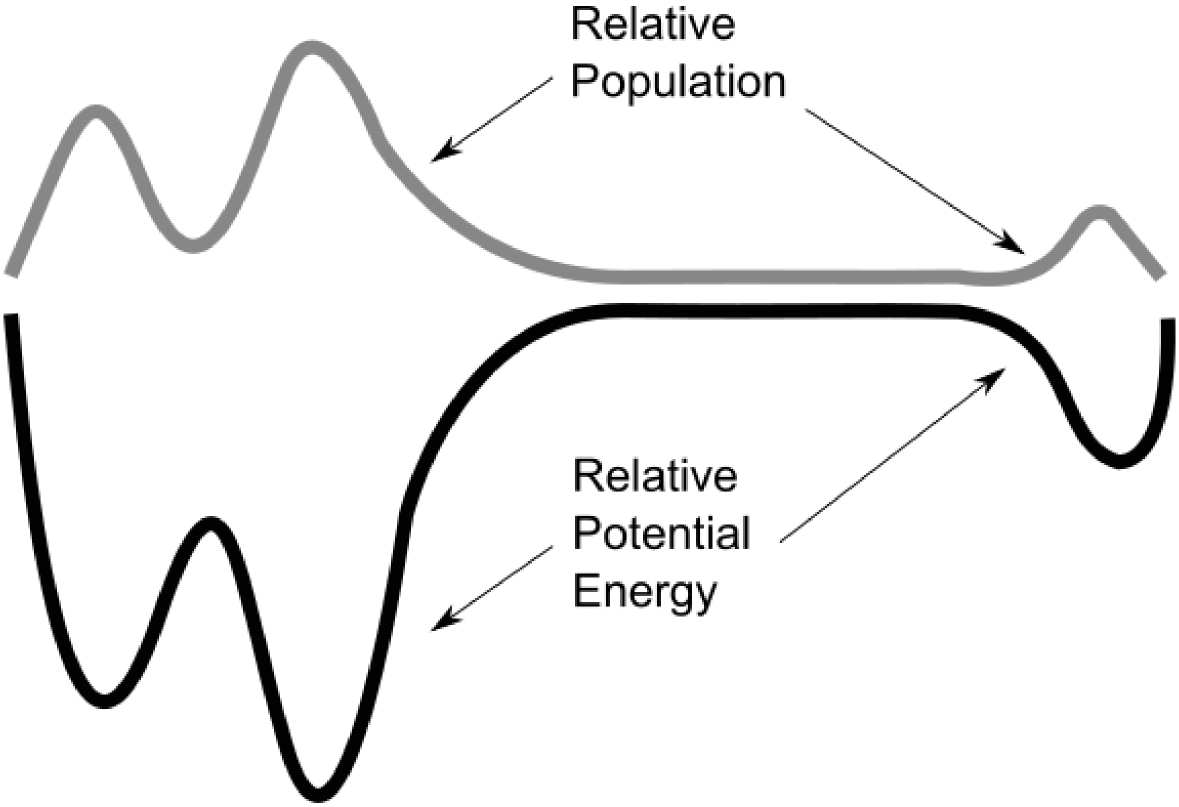
Terminology: State, Conformation, Geometry

It is useful to decide on a term for the set of geometries enclosed by rectangle 2. My natural tendency is to use the term *state*, but that word can mean almost anything: it is no more specific than *situation*. Use of that term is certainly correct, but a better term is nicer. The term *conformer* (or *conformational isomer*) generally applies to geometries accessible via rotations about single bonds. Certainly, that term also applies here in some way, but it doesn't feel right. I think something like *stable geometry* or *geometric conformation* is a little better. From protein geometries, *structural motif* is also close, but not quite right. From physics, the term *local density of states* is relevant here, and this is certainly a *monomodal cluster of states* in that regard, or maybe a *monomodal distribution of locally dense states*. But, again, no satisfying terms arise.

I’m going to stick with *geometric conformer* for now.

A Note About States and Populations

It is appropriate to group geometries based on observed populations. Whenever there is a monomodal population distribution, there is necessarily an underlying energetic landscape with an inverse shape. The image below illustrates this point.



A very rough figure showing the relationship between potential energy and population. The populations are not always a perfectly inverted version of the potential energy. For example, the population distributions are often narrower than the potential energy troughs (depending on what, exactly, is being measured). But, the basic relationship holds.

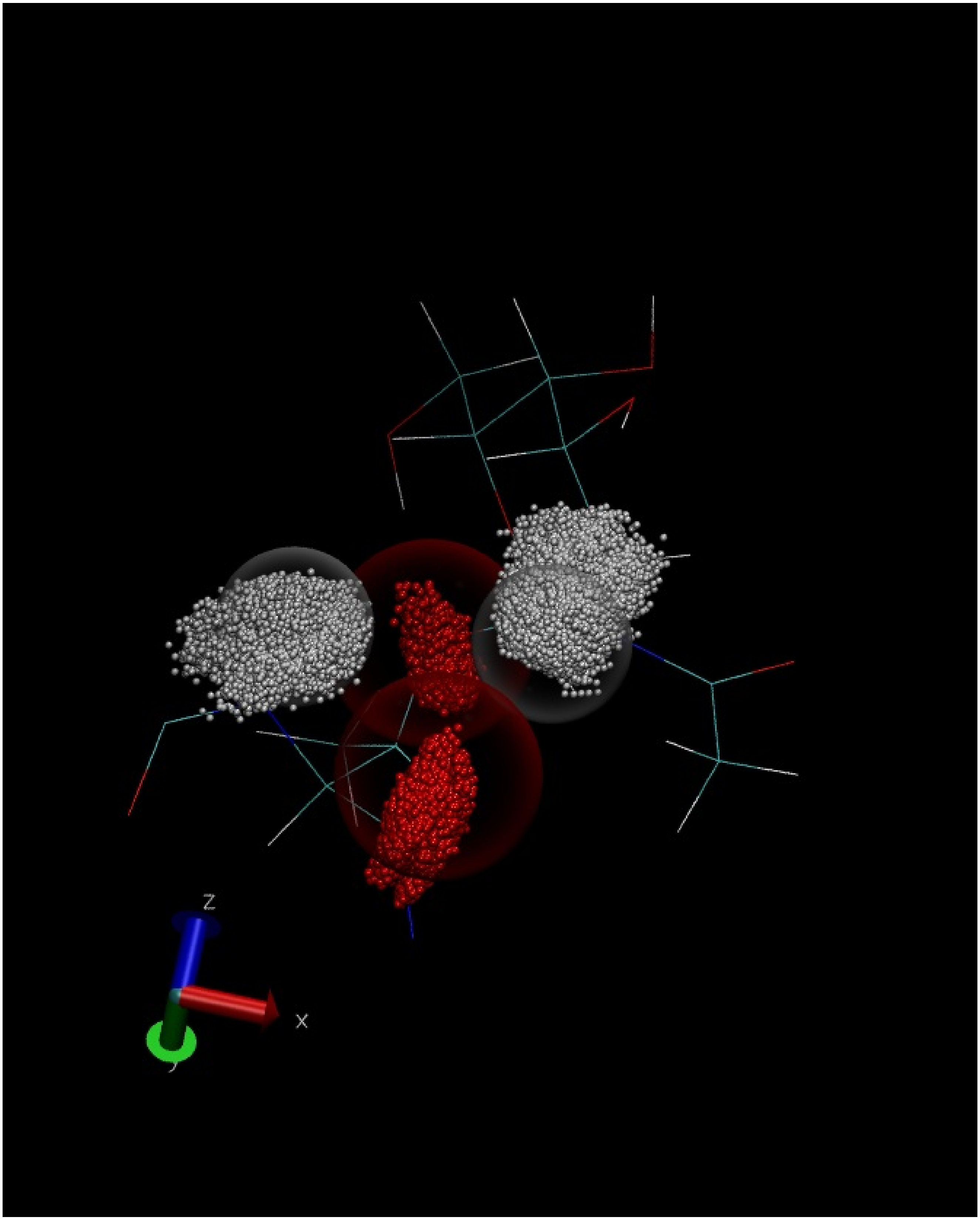
In other words, every time you have a peak in a population, you also have a corresponding trough in potential energy. When there are not other reasons to group structures together, grouping them by population peaks (potential energy troughs) is convenient and appropriate. The minima between population peaks represent transition states between the potential energy troughs.

If you compare the heat maps above to the more colorful contour plots, you can see that the region enclosed by rectangle 2 is, within the resolution of this data, a single, large, asymmetric peak. Therefore, there exists a complementary, asymmetric ‘bowl’ in the potential energy landscape. This is why I considered all the MD frames corresponding to the geometries enclosed by rectangle 2.

Identifying the Geometric Conformer in Rectangle 2

This post details my thoughts at the time. Briefly, I aligned all the relevant atoms and took a close look using VMD. After some doing, I noticed the interactions.

Here is the relevant image from that post:

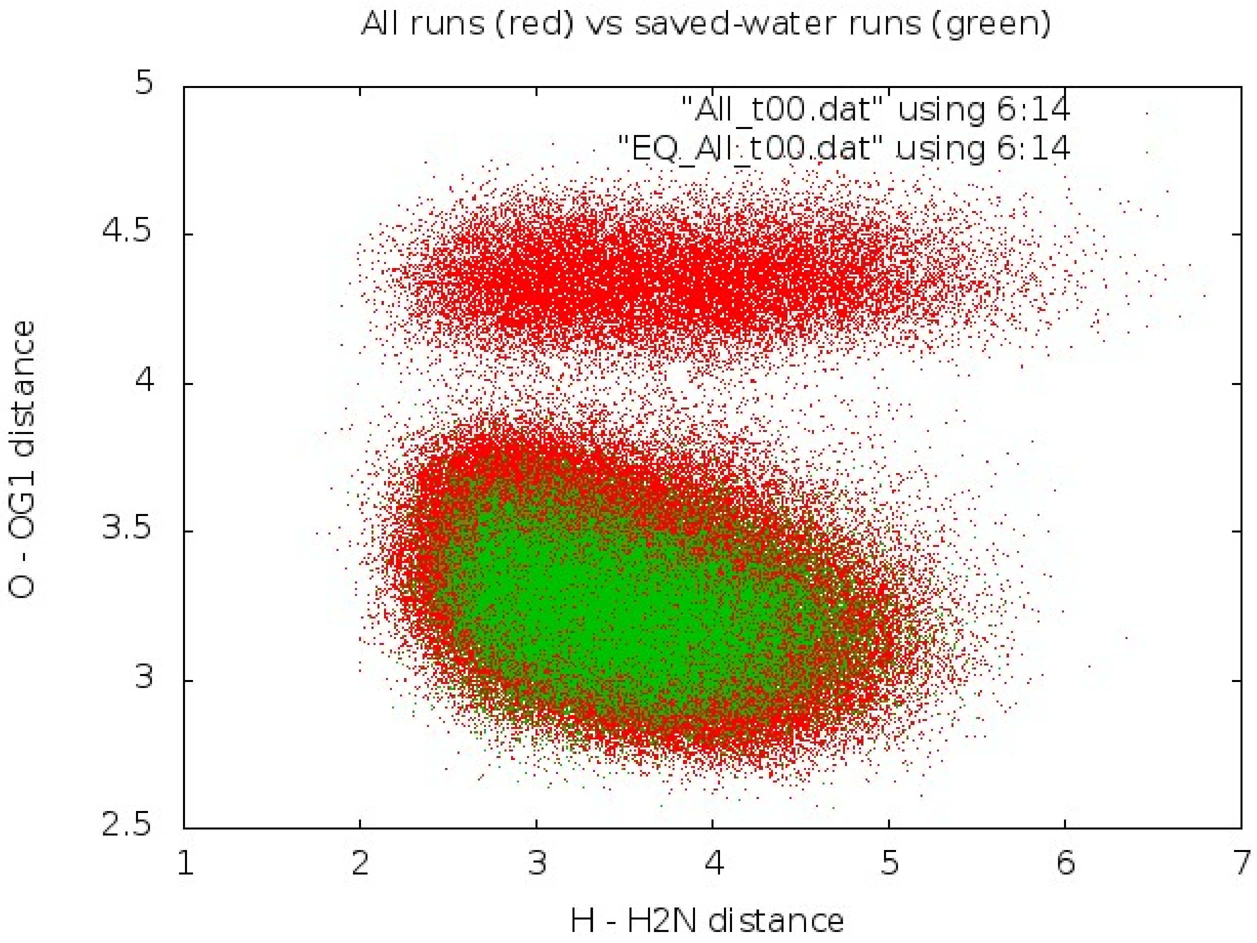


A persistent interaction between four atoms, two oxygens (red) and two hydrogens (white). The large transparent spheres show van der Waals sizes for the four atoms at a geometry close to the center of the population for this interaction. The smaller spheres show the individual locations for those four atoms in each frame from the MD simulation where the atoms were in this relative geometry as defined by the Ng-Hg-Ob angle and the Ng-Ob distance. The interaction between the top oxygen and the two hydrogens is significantly more persistent than the complete 4-way interaction.

Determining Relative Populations

The next thing to do was to determine how to calculate the relative populations. My initial calculations were based on notions related to populations and potential energy, and I had considered everything within rectangle 2 to be a single geometric entity. In fact, the image above shows relative positions of the two H's (small white balls) and the two O's (small red balls) for all the MD frames falling within that rectangle. When this is considered, most of the simulation is included (90-ish%, depending on site).

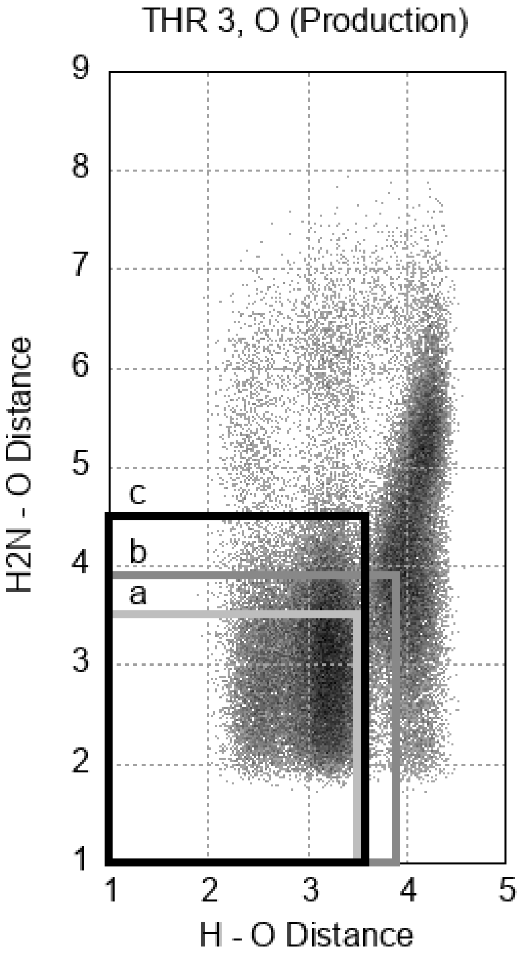
Later, I decided that a definition based just on the hydrogens and the oxygens would be useful. Trying to minimize the number of plots to looks at and/or data points to consider, I chose to plot the distances between the two oxygens versus the distance between the two hydrogens. That gave me this plot (ignore the green for now or see this post for more info):



This isn't a heat-map, but there are obviously two continuums of population. Again, by this measurement, most of the frames fall into the region bounded, here, by 4 A between the oxygens and about 6 A between the hyrogens.

However, there are previously suggested measures for bifurcated hydrogen bonding, and people are familiar with plots of H1—O versus H2—O distances, so David asked me to make plots like that. So, I did. You can see them all here (look at the PNG images). Here are a couple that are reasonably representative. Here, I am choosing the first site (THR 3 in the paper, or THR 4 in the simulations) because that is the site with the most variability in these plots. The other sites are more constrained to lower H—O distances.

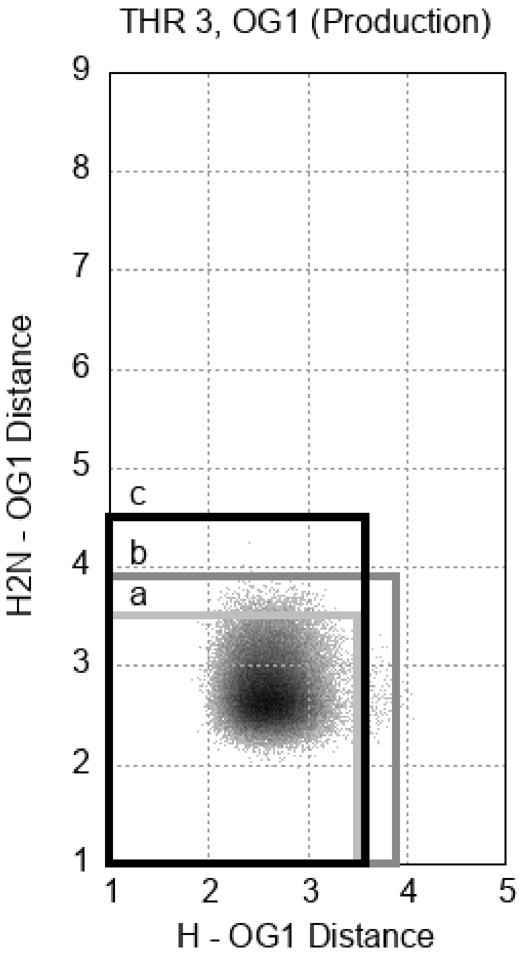
This is for the oxygen in the peptide backbone, interacting with the two hydrogens:



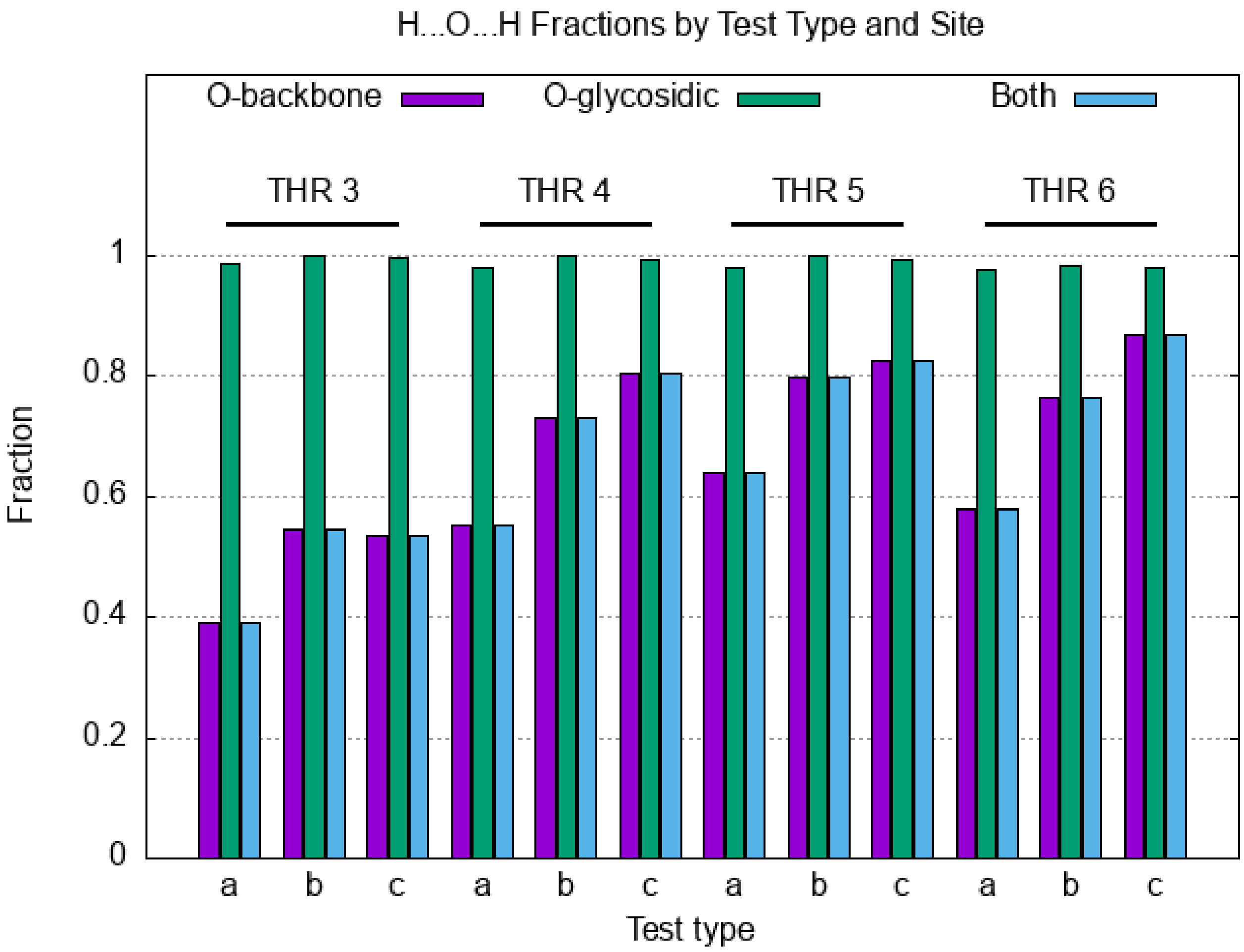
Distances from the backbone oxygen for the NAc hydrogen (y-axis) and the backbone hydrogen (x-axis). (a) A typical suggested cutoff for distances in a bifurcated hydrogen bond; (b) the maximum distance suggested; (c) cutoff corresponding to poputlation-potential considerations.

This is for the glycosidic oxygen. Axes and rectangles are the same as above.

Here is a big plot of the fractions of populations for all four sites, for all three cutoffs (a,b and c in the images), considering just the backbone O being inside each rectangle, considering just the glycosidic O, and considering both oxygens. It’s a lot to take in:



Distances from the glycosidic oxygen for the NAc hydrogen (y-axis) and the backbone hydrogen (x-axis). (a) A typical suggested cutoff for distances in a bifurcated hydrogen bond; (b) the maximum distance suggested; (c) cutoff corresponding to poputlation-potential considerations.



Calculation Method Details

First, I wrote a script (get_only_files_for_4-way-water-style.bash) that would copy down all the simulation files containing the raw data. I linked to the script from within the TREATED/d4g directory, and ran the script from there. The script made some directories and populated them. Notably for this, it made an ANALYSIS directory under d4g.

Next, I wrote a script (make_cpptraj_files_for_4-way-water-style_analysis.bash) to generate cpptraj input files and then run cpptraj. This script instructed cpptraj to extract the relevant distances and save them into files in a manner that I considered to be reasonably convenient. This script should be run from within the ANALYSIS directory, so I made a symbolic link to it from there.

After that, i wrote a program (Make_2D_bins_one_file.c) to do the 2D binning for me. Then, I wrote a script (bin_4-way-water-style_data.bash) that called the compiled version of the program, producing the needed binned output. This script should also be run from within the ANALYSIS directory, so I made a symbolic link to it from there.

At some point, I decided to move all the PLOTS directories out of the ANALYSIS directories and up one level (to the d4g, d4m or protein directories). I did this so that most of the plotting information would live in the git repo. This breaks some links from the plots directories, but I made notes about that in relevant locations.

The 2D binning program referenced above requires an input file. I saved all the input files in the TREATED/d4g/PLOTS/4WAY_HOH directory. I also put various scripts and other files related to making the plots in that directory. I used gnuplot to make all the plots. The gnuplot I used declares itself to be “Version 5.2 patchlevel 8 last modified 2019-12-01”.

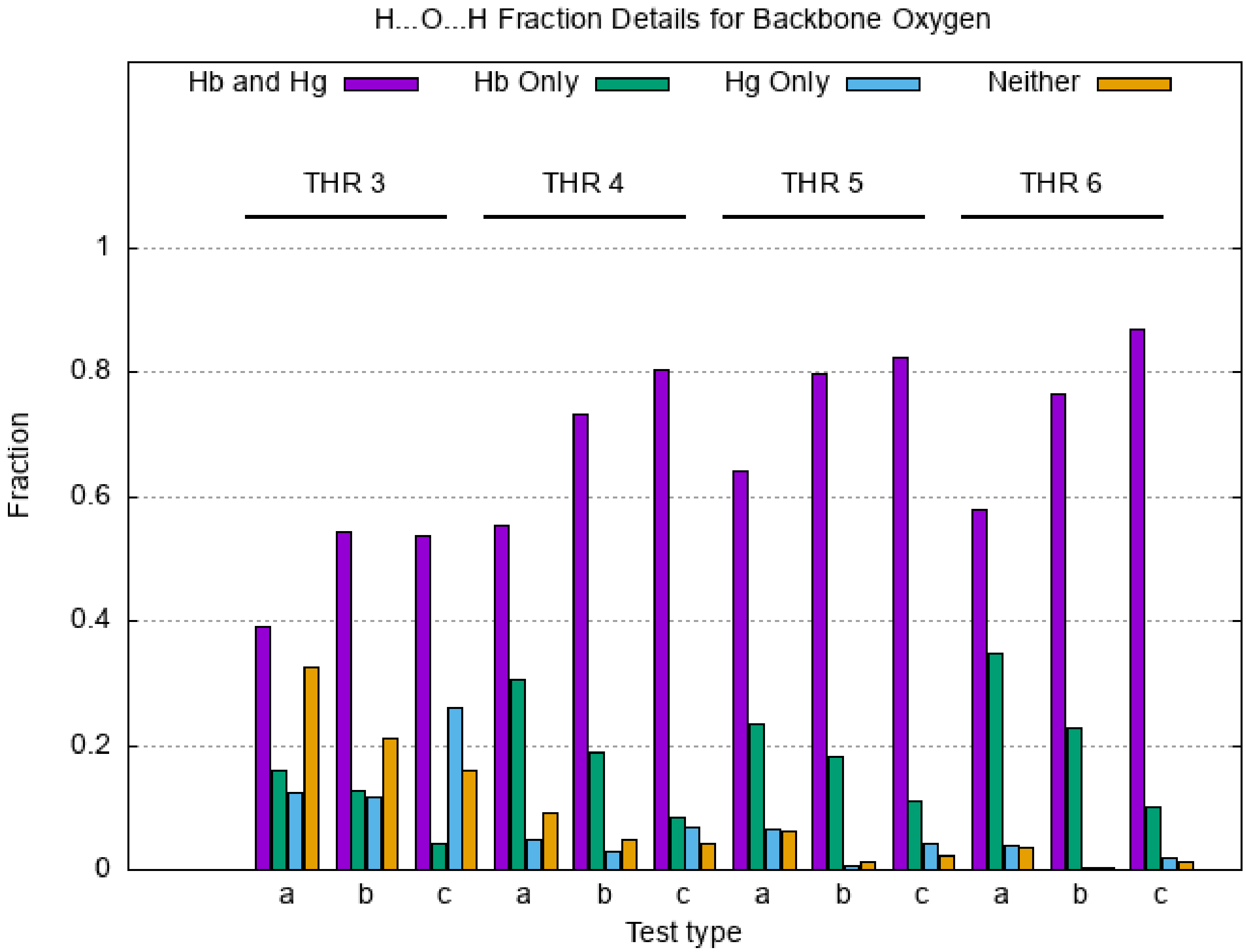
For the final figures, I plan to export to postscript and then use GIMP to convert the images into whatever resolution and/or format is required.

Interpretations

The most obvious interpretation is that the glycosidic oxygen is pretty firmly planted between the NAc’s H2N atom and the backbone H. It occasionally strays a little, but not by much. The next interpretation is that this interaction is causing a point of attraction for the backbone oxygen. It doesn’t participate in the bifurcated H-bond as much as the glycosidic oxygen, but it spends a lot of time in association with at least one of the two hydrogens. Based on the latter, the backbone oxygen is in and extended association with the two hydrogens. While the interaction does not always meet the suggested definition of a bifurcated hydrogen bond, it is obviously quite stable.

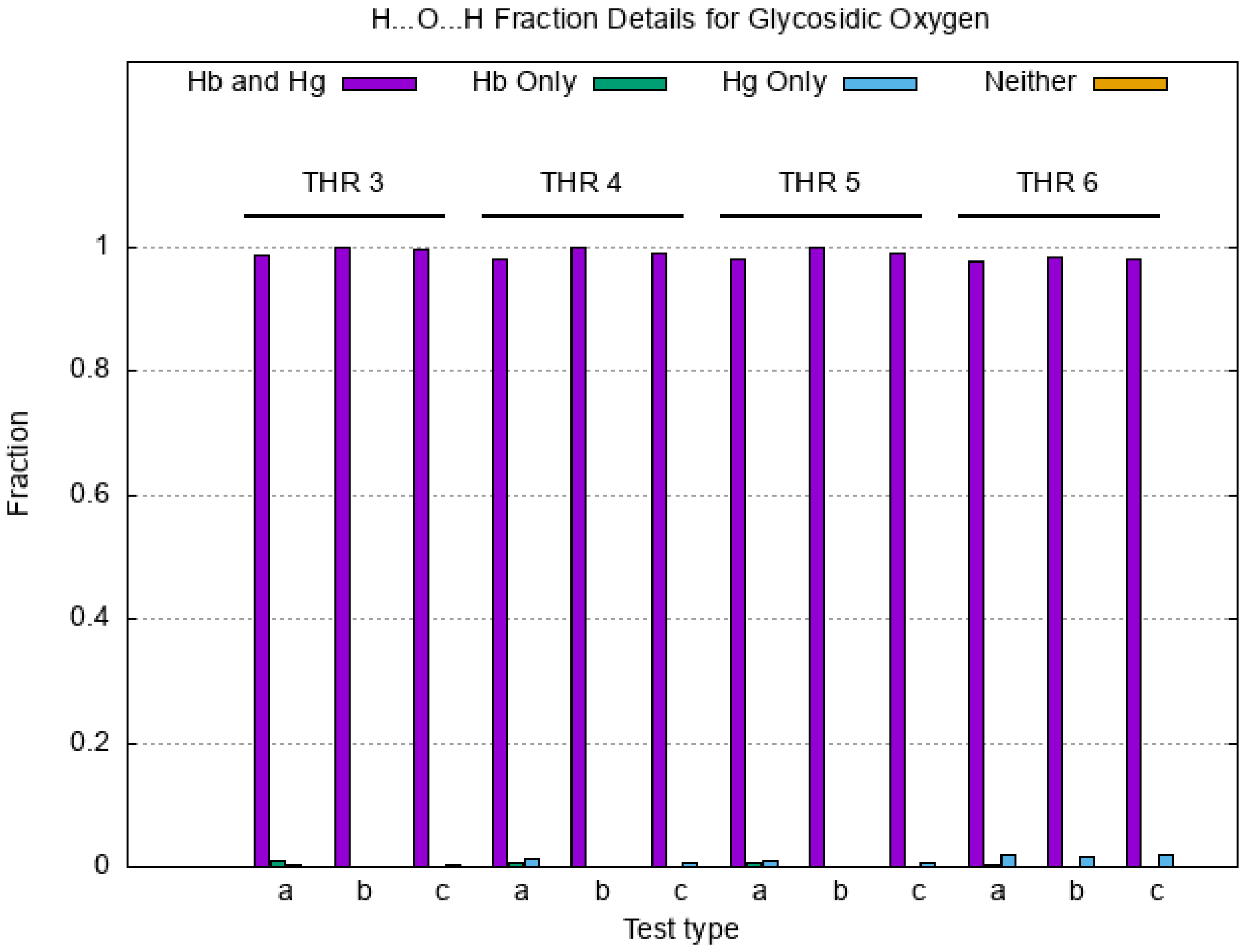
To quantify this, I calculated the fraction of time the backbone and glycosidic oxygens were in association with Hb, Hg, both of those or neither. I also did this for all three cutoff ranges (test type – ‘a’ ‘b’ and ‘c’, above) and at each glycosylation site.

Here is the plot for the backbone oxygen:



Note the significant difference between the first site and the others. There also seems to be a trend across the sites. It will take further investigation to determine the cause of this.

And, for the glycosidic oxygen, which is a much more boring plot:


H...O...H Fraction Details for Glycosidic Oxygen
Hb and Hg
Hb Only
Hg Only
Neither
THR 3
THR 4
THR 5
THR 6
Fraction
1
0.8
0.6
0.4
0.2
0
a
b
c
a
b
c
a
b
c
a
b
c
Test type