Application of Statistical Methods to the Prediction of B3LYP-Optimized Polyhedral Water Cluster Geometries

David J. Anick MD PhD
Harvard Medical School
Water Clusters are Difficult to Study

- H-bonds are weak (5 – 8 kcal / mol)
- Wide range of lengths (243 to 305 pm, in optimized \((\text{H}_2\text{O})_N\)'s)
- Wide range of \(\text{O} – \text{O} – \text{O}\) angles (80.9° to 100.7° in 4-sided rings, 102.2° to 130.8° in hexagons)
- Strong cooperativity effects
- Strong electron correlation effects (Hartree-Fock is inadequate)
- MP2/6-311++g** is a good model
- Benchmarks: B3LYP/6-311++g** geometries come within \(~ 1\%
- To optimize an \((\text{H}_2\text{O})_{20}\) on Pentium III 733 MHz takes 10 – 12 hours per iteration and 15 – 25 iterations
Polyhedral Water Clusters

- Each O is involved in exactly 3 H-bonds
- Applicable to clathrate cells and to the units comprising fused structures
- \# H-bonds = 3N/2 \ (\Rightarrow N \ is \ even)\n- Exactly N/2 O’s have a pendent (non-H-bonded) H ("F-type")
- Exactly N/2 O’s have a non-H-bonding lone pair ("L-type")
- Further restriction for this project: all faces have 4, 5, or 6 sides
- The ability to designate each O as "F-type" or "L-type" yields a great simplification from a database point of view
- A database of 69 B3LYP-optimized polyhedral clusters, ranging from N=8 to N=20, was generated (1,221 H-bonds in all)
- GOAL OF PROJECT: To mine the database for patterns which would allow for the rapid prediction of new cluster geometries
- Wanted to come within ~ 50 cal/mol per H₂O of the B3LYP optimum
Polyhedral Water Clusters: Representations
Connectivity Representation for Polyhedral Water Cluster:

- H-Bond
- F-type Oxygen (with pendent H)
- L-type Oxygen (no pendent H)
- Position of H within H-bond
Outline of Approach

• **STEP 1**: Correlate O – O distances with types (i.e. F or L) of donor, acceptor, and other nearby O’s. Other descriptors include adjacent O – O – O angles and the cluster’s total dipole moment.

• **STEP 2**: Correlate O – O – O angles with local O types, O – O distances, and various combination quantities that may be relevant.

• **STEP 3**: Given a connectivity representation, predict O – O lengths and compute “target” O – O – O angles. A constrained optimization is done, which finds the geometry which satisfies the length constraints exactly and which minimizes the RMS difference between the actual angles and the targets. This yields an “oxygen skeleton.”

• **STEP 4**: Correlate the O – H distances, H – O – H angle, and three H₂O orientation parameters, with distances and angles in the O skeleton. Use these to decorate the O skeleton with H’s, making (H₂O)ₙ.

• **STEP 5**: Apply method to a test set of 16 structures (381 H-bonds, 12 ≤ N ≤ 24) and then find each true B3LYP optimum, to evaluate.
Analysis of O – O Distances

- Split data set of H-bonds into 5 subsets according to the types (i.e. F or L) of donor and acceptor
- Using a template for each of the five bond types, correlate the O – O distances with parameters descriptive of each bond’s local context

**Templates for type FF and type LL bonds**

- **TYPE FF**
  - O1 = index bond donor
  - O2 = index bond acceptor
  - OP2 = acceptor of O2
  - OP1 = cis to OP2
  - OQ1 = other donor to O1
  - OQ2 = donor to O2 (not O1)

- **TYPE LL**
  - O1 = index bond donor
  - O2 = index bond acceptor
  - OQ1 = donor to O1
  - OQ2 = cis to OQ1
  - OQ1 = other acceptor of O1
  - OQ2 = other acceptor of O2
Analysis of O – O distances, cont.

Symmetry of the FL template means that OP1/OP2 cannot be distinguished from OQ1/OQ2. Only symmetric functions of OP1, OQ1, OP2, OQ2 may be used as descriptors.

Different templates for 'cis' vs. 'trans' setups necessitate splitting the set of LF bonds into two smaller data sets to be analyzed separately.

Descriptors include the types of OP1, OQ1, OP2, OQ2; the types of the tertiary O’s; local angles e.g. OP1-O1-O2 and OP2-O2-OQ2; and the proj. of the total cluster’s dipole moment onto O1-O2.
**Descriptor Pruning Algorithm**

- Choose cutoff p-value $p_{cut}$ (0.005 was used for O – O distances and O – O – O angles; 0.001 was used for H placement parameters).

- For each descriptor, the distance from the “observed values vector” to the subspace spanned by the other descriptors, and the distance to the subspace spanned by all descriptors, are computed. The distance will always be shorter to the latter subspace. A p-value is computed which measures the likelihood that the size of the difference in distances could occur due to chance (i.e. with a random descriptor).

- If any p-value is $> p_{cut}$, eliminate the descriptor having the largest p-value.

- Repeat the pruning procedure, deleting one descriptor at a time, until all remaining descriptors have p-values $< p_{cut}$.

- Perform a linear regression using the set of kept descriptors.
### O – O Distances: Data Set Statistics

<table>
<thead>
<tr>
<th>Bond Type</th>
<th>Training Set</th>
<th></th>
<th></th>
<th></th>
<th>Test Set</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># pts</td>
<td>Mean ± SD</td>
<td>Min</td>
<td>Max</td>
<td># pts</td>
<td>Mean ± SD</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>FF</td>
<td>234</td>
<td>276.3 ± 6.1</td>
<td>259.0</td>
<td>290.2</td>
<td>56</td>
<td>275.2 ± 6.4</td>
<td>256.3</td>
<td>287.0</td>
</tr>
<tr>
<td>FL</td>
<td>173</td>
<td>262.5 ± 4.4</td>
<td>248.0</td>
<td>270.3</td>
<td>71</td>
<td>262.0 ± 5.2</td>
<td>243.2</td>
<td>273.5</td>
</tr>
<tr>
<td>LL</td>
<td>234</td>
<td>278.3 ± 6.4</td>
<td>258.8</td>
<td>289.6</td>
<td>56</td>
<td>276.1 ± 6.5</td>
<td>256.6</td>
<td>288.5</td>
</tr>
<tr>
<td>LFc</td>
<td>213</td>
<td>291.1 ± 5.4</td>
<td>275.4</td>
<td>303.4</td>
<td>108</td>
<td>286.6 ± 4.8</td>
<td>275.2</td>
<td>297.2</td>
</tr>
<tr>
<td>LFt</td>
<td>367</td>
<td>288.7 ± 6.3</td>
<td>273.2</td>
<td>304.6</td>
<td>90</td>
<td>285.6 ± 5.3</td>
<td>274.5</td>
<td>302.9</td>
</tr>
</tbody>
</table>

Where:

- **Max**: Maximum value.
- **Min**: Minimum value.
- **Mean ± SD**: Mean value ± Standard Deviation.
# O – O Distances: Regression Results

<table>
<thead>
<tr>
<th>Bond Type</th>
<th>Initial # Descriptors</th>
<th># Descriptors Kept</th>
<th>Maximum p-value</th>
<th>Training Set RMS Error</th>
<th>Test Set RMS Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF</td>
<td>35</td>
<td>19</td>
<td>0.0001</td>
<td>0.84</td>
<td>1.53</td>
</tr>
<tr>
<td>FL</td>
<td>21</td>
<td>14</td>
<td>0.0000</td>
<td>0.52</td>
<td>0.84</td>
</tr>
<tr>
<td>LL</td>
<td>35</td>
<td>20</td>
<td>0.0027</td>
<td>0.73</td>
<td>1.13</td>
</tr>
<tr>
<td>LFc</td>
<td>35</td>
<td>21</td>
<td>0.0034</td>
<td>0.98</td>
<td>1.60</td>
</tr>
<tr>
<td>LFt</td>
<td>35</td>
<td>16</td>
<td>0.0042</td>
<td>0.96</td>
<td>1.39</td>
</tr>
</tbody>
</table>
“Complexity” of Polyhedral Clusters

- A crude measure of the difficulty of predicting a cluster’s properties
- Defined as the size of the largest connected component of F- or L-type O’s in the connectivity diagram
- For a given polyhedron, the lowest energy structures have $C = 1$ or $2$
- Theorem: For any polyhedron, there exist clusters with $C = 1$ or $2$

Test Cluster 8: $N = 16, C = 2$

Test Cluster 11: $N = 16, C = 6$
Performance Evaluation: Comparing Predicted and Optimized Geometries

<table>
<thead>
<tr>
<th>Cluster #</th>
<th>N</th>
<th>C</th>
<th>RMS $\Delta$ O–O</th>
<th>RMS $\Delta$ O–O–O</th>
<th>$\Delta E / N$ cal/mol</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>1</td>
<td>0.98</td>
<td>0.74</td>
<td>12.0</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>3</td>
<td>1.38</td>
<td>1.30</td>
<td>19.1</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>5</td>
<td>0.84</td>
<td>0.75</td>
<td>31.8</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>2</td>
<td>1.00</td>
<td>0.94</td>
<td>21.7</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>3</td>
<td>1.81</td>
<td>2.32</td>
<td>25.2</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>6</td>
<td>0.93</td>
<td>1.43</td>
<td>14.8</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>1</td>
<td>1.25</td>
<td>1.07</td>
<td>11.8</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>2</td>
<td>0.99</td>
<td>0.69</td>
<td>19.2</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>4</td>
<td>1.27</td>
<td>1.77</td>
<td>23.1</td>
</tr>
<tr>
<td>10</td>
<td>16</td>
<td>6</td>
<td>1.54</td>
<td>2.08</td>
<td>45.1</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>6</td>
<td>1.55</td>
<td>2.50</td>
<td>51.4</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>7</td>
<td>2.17</td>
<td>1.62</td>
<td>32.2</td>
</tr>
<tr>
<td>13</td>
<td>18</td>
<td>2</td>
<td>1.35</td>
<td>1.33</td>
<td>26.2</td>
</tr>
<tr>
<td>14</td>
<td>18</td>
<td>3</td>
<td>1.03</td>
<td>1.23</td>
<td>28.0</td>
</tr>
<tr>
<td>15</td>
<td>20</td>
<td>8</td>
<td>1.65</td>
<td>1.77</td>
<td>38.6</td>
</tr>
<tr>
<td>16</td>
<td>24</td>
<td>2</td>
<td>0.91</td>
<td>1.11</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Distances in pm, angles in degrees, energy in cal/mol per H$_2$O
Conclusions about Water Clusters

• In polyhedral \((H_2O)_N\)’s, H-bond lengths and angles, O–H distances, and \(H_2O\) orientation parameters correlate strongly with local and global cluster descriptor parameters.

• B3LYP-optimized \((H_2O)_N\) geometries can be predicted to within \(\sim 50\) cal/mol per \(H_2O\) (and often better) using statistical methods.

• There is a rough correlation between the complexity of a cluster and the accuracy of the prediction.

• Analysis of the data base provides several new insights into water clusters, including the pronounced variation of H-bond properties with bond type and the influence of total cluster dipole on the local geometry.
Structural Database Mining Principles

- If possible, restrict the problem to a uniform class of objects.
- Use templates to split the data set into subsets, within which the descriptors will have consistent meaning.
- Where template symmetry is unavoidable, use symmetric functions of the descriptors.
- Pruning algorithm gave good results in this study.
- Generate redundant predictions (i.e. overdetermine the system) and let the predicted structure represent an optimum fitting.
- Look to outliers to suggest overlooked descriptors.