Molecule Unfolding with Quantum Annealing

Intro to Quantum Computing

25 June 2021

Riccardo Mengoni, Kevin Mato, Daniele Ottaviani, Gianluca Palermo
**Molecular Docking** is a method to calculate the **preferred position and shape** of one molecule to a second when bound to each other

- Shape Complementarity
- Scoring function to evaluate the binding affinity
3 Phases Process

Ligand Expansion
- MOL2 ligand elaboration
- Identification of the rotatable bonds
- Internal distances maximization
- Removes tool related bias (e.g. smile-to-3D)

Initial Placement
- Ligand main fragments decomposition
- Ligand initial poses identification
- Placement of the ligand into the pocket with rigid roto-translations

Shape Refinement
- Use of the rotatable bonds to modify the ligand shape and to match the protein pocket
- Docking Score Maximization
3 Phases Process

**Ligand Expansion**
- MOL2 ligand elaboration
- Identification of the rotatable bonds
- Internal distances maximization
- Removes tool related bias (e.g. smile-to-3D)

**Initial Placement**
- Ligand main fragments decomposition
- Ligand initial poses identification
- Placement of the ligand into the pocket with rigid roto-translations

**Shape Refinement**
- Use of the rotatable bonds to modify the ligand shape and to match the protein pocket
- Docking Score Maximization
**Problem Definition**

**Objective:** find the unfolded torsion configuration that maximizes the molecular volume, or equivalently, that maximizes the distances between fragments.

\[ \theta = [\theta_1, ..., \theta_M] \]
To each torsion is associated is a rotation matrix $R$.

\[ \hat{a}(\theta_1) = R(\theta_1)a_0 \]

\[ \hat{a}(\theta_1, \theta_2) = R(\theta_2)R(\theta_1)a_0 \]
Overview of the problem size (ComplexDB)
Molecule Unfolding: Rotatable Betweenness Centrality

**Betweenness centrality:**

\[ g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \]
Molecule Unfolding: Rotatables Influence Set

Rotatables Influence set:

\[ I_S = E_{C_\alpha, a_k} \cap E_R \]

\[ E_R = \text{Rotable bonds}; \]
\[ E_{C_\alpha, a_k} = \text{Bonds on the shortest path } \sigma_{C_\alpha, a_k} \]
Molecule Unfolding

Original 2D molecule:

Without Hydrogen:

Double and amide bonds are not rotatable
**Problem Definition**

**Objective:** find the unfolded *torsion configuration* that maximizes the molecular volume, or equivalently, that maximizes the distances between fragments.

\[ \theta = [\theta_1, \ldots, \theta_M] \]
Molecule Conformation Exploration: Random Search

Random Search

\[ F(\theta_1, \theta_2) \]

\[
\begin{array}{cccc}
P0 & P1 & P2 & P3 \\
& & & \\
P4 & P5 & P6 & P7 \\
& & & \\
P8 & P9 & P10 & P11 \\
\end{array}
\]

\( A_0 \)

\( \theta_1 \)

\( \theta_2 \)

\( A_D \)
Molecule Conformation Exploration: Greedy

\( M = \text{Measure total sum of internal distances} \)

\( \theta^\# = \text{physical rotation of torsion}\# \text{ for all possible angles} \)

**Greedy:**

- **GeoDock** = \( \theta_1 - M - \theta_2 - M - \theta_3 - M - \theta_4 - M \)
- **Batch** =  
  i) \( \theta_1, \theta_2 - M - \theta_3, \theta_4 - M \)  
  ii) \( \theta_1, \theta_2, \theta_3 - M - \theta_4 - M \)  
  iii) \( \theta_1, \theta_2, \theta_3, \theta_4 - M \)
It is convenient to identify a **conformation of a molecule** with M torsions by a **torsion vector**

\[
[\theta_1, \ldots, \theta_M]
\]

Where each torsion \(\theta_N\) can assume values in \([0, 2\pi)\).

**Objective:** find the unfolded **torsion configuration** that maximizes the molecular volume, or equivalently, that maximizes the distances between fragments.
Objective: find the unfolded torsion configuration

\[ [\theta_1^{unfold}, \ldots, \theta_M^{unfold}] \]

that maximizes the sum of distances \( D_{ab}(\theta) \) between fragments \( a \) and \( b \)

\[
D(\theta) = \sum_{a,b} D_{ab}(\theta)^2
\]

where \( D_{ab}(\theta)^2 = \|\vec{a}_0 - R(\theta)\vec{b}_0\|^2 \)
Constructing the Binary Optimization problem

Consider a **discretization of the torsion angle** $\theta_i$ into $d$ possible values

$$\theta_i = [\theta_i^1, \theta_i^2, \theta_i^3, ..., \theta_i^d]$$

And introduce a **binary variable** $x_{ik}$ with $1 \leq k \leq d$, such that

$$x_{ik} = \begin{cases} 1 & \text{if } \theta_i = \theta_i^k; \\ 0 & \text{otherwise.} \end{cases}$$

with the constraint

$$\sum_{k=1}^{d} x_{ik} = 1$$
This also induces a **discretization of the sine and cosine** for each torsion

\[
\sin(\theta_i) = \sum_{k=1}^{d} \sin(\theta_i^k) \ x_{ik} \quad \text{and} \quad \cos(\theta_i) = \sum_{k=1}^{d} \cos(\theta_i^k) \ x_{ik}
\]

With such encoding, the **rotation matrix** \( R(\theta_i) \) associated the **torsion angle** \( \theta_i \) becomes a **function of** all the **binary variables** \( x_{ik} \) needed to represent the angle \( \theta_i \)

\[
R(\theta_i) = R(x_{i1}, x_{i2}, \ldots, x_{id})
\]
The general form of the HUBO optimization function is

\[ O(x_{ik}) = A \sum_i \left( \sum_{k=1}^{d} x_{ik} - 1 \right)^2 - \sum_{a,b} D_{ab}(\theta)^2 \]

where the pairwise distances are expressed using the binary variables

\[ D_{ab}(\theta)^2 = \| \vec{a}_0 - R(\theta)\vec{b}_0 \|^2 \]

In general, if \( D_{ab}(\theta) \) depends on \( m \) torsions, \( D_{ab}(\theta) \) contains terms up to the \( m \)-th order, hence the highest order in the HUBO is \( 2m \)
In order to obtain a precision of $\Delta\theta_i$, the number of variables needed for each torsion is

$$d = \frac{2\pi}{\Delta\theta_i} = \frac{2\pi}{\theta_{i}^{k+1} - \theta_{i}^{k}}$$

Given a molecule with $M$ torsions, the total number of binary variables $x_{ik}$ in the HUBO

$$n = d \times M = \frac{2\pi}{\Delta\theta_i} \times M$$

Molecules: 20 to 50 atoms - 10 torsions
Angle subsampling effect on the unfolding degradation

Granularity Degradation Effect 15%

Unfolded vs Initial

AVG Volume gain 17%
HUBO Problem Structure at $\Delta \theta_i = \pi/4$

- **HUBO linear terms**
  - Mean number of linear terms by rotables and chop
  - $n = d \times M = \frac{2\pi}{\Delta \theta_i} \times M$

- **HUBO high order terms (Log-scale)**
  - Mean number of high-order by rotables and chop

---

**Graphs:**
- Bar charts showing the number of linear terms and high-order terms for different numbers of torsions and parameters $\epsilon_0$, $\epsilon_20$, $\epsilon_60$.
Delete HUBO terms below a certain threshold. Applied in two phases:

1. Speed up the construction of the HUBOs;

2. Speed up the transformation of HUBOs into QUBOs (done via dimod.make_quadratic);

Approximated HUBO problems solvable with DW2000Q and Advantage
Form HUBOs to QUBOs

QUBO linear terms
Average number of linear terms by rotables and chop

<table>
<thead>
<tr>
<th>Number of torsions</th>
<th>c0.1</th>
<th>c100</th>
<th>c200</th>
<th>c300</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>23.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>56.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>66.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>40.57</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>100.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>90.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>58.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>219.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>174.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>75.33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

QUBO quadratic terms
Average number of quadratic terms by rotables and chop

<table>
<thead>
<tr>
<th>Number of torsions</th>
<th>c0.1</th>
<th>c100</th>
<th>c200</th>
<th>c300</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>198.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>331.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>544.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>607.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>533.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>245.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2541.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1306.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>341.67</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Embeddings DW2000Q & Advantage

**Number of physical qubits**

Average number of qubits for each topology

- avg. 51% less physical qubits

**AVG chain length**

Average chain length for each topology

- avg. 52% shorter chains
Results, 4 Torsions : Volume Gain in Time (seconds)
Results, 6 Torsions : Volume Gain in Time (seconds)

Maximal Volume Gain in Time, with distance simplification, R.6

Percentage of Volume Gain

Time

SA, RAN, ADV, 2000Q, GEO
Results, 8 Torsions : Volume Gain in Time (seconds)
Results, 8 Torsions: Volume Gain in Time (seconds)

Maximal Volume Gain in Time, with distance simplification, R:8

Percentage of volume Gain

SA, RAN, ADV, GEO, 2000Q

Sampling time ($T_s$), Delay time ($T_d$), Readout time ($T_r$), Anneal time ($T_a$), Programming time ($T_p$)

CINECA
POLITECNICO MILANO 1863
Results: Time To Solution (TTS) & Volume Gain

**TimeToSolution:**
the lower the better

**Volume Gain:**
the higher the better

\[
TTS = \frac{\text{total execution time}}{\text{occurrence of the best solution}}
\]
Results: Normalized Volume Gain per TTS

Normalized Volume Gain per TTS:

- Takes into account both quality of solution and TTS
- Measures how fast the method fails to provide good solutions
- Advantage has lower avg. slope with respect to SA and DW2000Q

<table>
<thead>
<tr>
<th></th>
<th>sa</th>
<th>two</th>
<th>adv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.609948</td>
<td>6.0189816</td>
<td>1.36206524</td>
</tr>
</tbody>
</table>
• We tackled the problem of Molecular Unfolding, an important step in molecular docking.

• New HUBO formulation that can be solved on D-WAVE annealers has been developed.

• We have observed that by increasing the approximation threshold with the problem size, it is possible to embed formulations that couldn’t be otherwise.

• Embedding our problems on Advantage, compared to the DW2000Q, cost 51% less in terms of physical qubits and with chains 52% shorter.

• In terms of absolute time (seconds), SA is the fastest method to provide close to optimal solutions.

• Advantage significantly outperforms DW2000Q in terms of TTS and VolumeGain by increasing torsions. Advantage also show a better NormalizedVolumeGain/TTS scaling w.r.t. SA.