Supporting Information II for the Movable Type Method Applied to Protein-Ligand Binding

*Zheng Zheng, Melek N. Ucisik and Kenneth M. Merz, Jr.**

Department of Chemistry and the Quantum Theory Project, 2328 New Physics Building, P.O. Box 118435, University of Florida, Gainesville, Florida 32611-8435

Different approximate algorithms have been explored for the Movable Type (MT) matrix multiplication approach presented in the manuscript. In order for our algorithm to have good accuracy and computational performance we require: (1) much smaller sub-matrices for computational convenience, (2) avoid creation of large matrices during MT computation, (3) inclusion of as many as possible of the significant Boltzmann factor and probability values in the respective sub-matrices. It is difficult to avoid a tensor product generating a large matrix for any large molecular system, thus we utilized a Hadamard (pointwise) product through fixed-size matrices for all atom pairs. This means the size of the final matrix is pre-determined at the beginning of the computation, and the Boltzmann factor and probability matrices for each atom pair should have the same size. We call these fixed-size matrices for all atom pairs "Standard Matrices". Obviously, sizes of the original atom pair vectors with tens to hundreds of elements shown in Equation 1 below are far from enough for the final matrix size. Construction of the "Standard Matrices"

relies on replication and tiling of the original atom pair vectors. Wherein, the vectors for each individual atom pairwise Boltzmann factor and probability are replicated in the "Standard Matrices" through all atom pairs. In order to perform the vector-to-matrix conversion, randomly scrambled permutations of the original vectors are needed. By introducing permutations to the original vector increases the diversity of atom pair combinations at different discrete distance values (r_a) in the MT computation, thereby increasing the sample size. We offer a detailed explanation in the latter paragraphs.

Using the bond probability vector as an example, permutations were made based on a vector with elements in order.

$$q_k^{bond} = \begin{bmatrix} q_k^{bond}(r_1) \\ q_k^{bond}(r_2) \\ q_k^{bond}(r_3) \\ \vdots \\ q_k^{bond}(r_t) \end{bmatrix} \Rightarrow scram\left(q_k^{bond}\right)_i = \begin{bmatrix} q_k^{bond}(r_{t-2}) \\ q_k^{bond}(r_3) \\ q_k^{bond}(r_1) \\ \vdots \\ q_k^{bond}(r_2) \end{bmatrix} \quad (1)$$

q_k^{bond} is the unscrambled probability vector of the kth atom pair with a bond constraint. t is the number of discrete probabilities with significant values. $scram(X)_i$ represents a randomly scrambled permutation of matrix X with i as the index number. The enlarged matrix of q_k^{bond} is represented as follow:

$$Q_k^{bond} = \begin{bmatrix} scram\left(q_k^{bond}\right)_1 & scram\left(q_k^{bond}\right)_{\alpha+1} & \cdots & scram\left(q_k^{bond}\right)_{\beta-\alpha+1} \\ scram\left(q_k^{bond}\right)_2 & scram\left(q_k^{bond}\right)_{\alpha+2} & \cdots & scram\left(q_k^{bond}\right)_{\beta-\alpha+2} \\ \vdots & \vdots & \ddots & \vdots \\ scram\left(q_k^{bond}\right)_\alpha & scram\left(q_k^{bond}\right)_{2\alpha} & \cdots & scram\left(q_k^{bond}\right)_\beta \end{bmatrix} \quad (2)$$

$\mathcal{Q}_k^{bond}$ in Equation 2 is the "Standard Matrix" we built for the *k*th atom pair with a bond constraint. Although the sizes (the number *t* in Equation 1) of different vectors q_k vary under different constraints (*i.e.* $q_k^{bond}, q_k^{angle}, q_k^{torsion}, q_k^{long-range}$), $\mathcal{Q}_k$ for all atom pairs was fixed to the same size with a predetermined permutation number α and β. The size of the Standard Matrix (SM) *e.g.* *g* rows ×*h* columns, must satisfy that the row number *g* is divisible by the sizes *t* of all the atom pair vectors q_k, so that each discrete probability $q_k(r_i)$ has an equal number of appearance in each SM $\mathcal{Q}_k$. This definition is important to make sure the replication numbers for all Boltzmann factors and probabilities are identical in each SM, so that their relative probabilities are the same as in the original probability vector. Hadamard products of all the protein probability SMs (*n* as the total number of atom pairs) are then performed:

$$\mathcal{Q}_P^{bond} = \mathcal{Q}_1^{bond} \circ \mathcal{Q}_2^{bond} \circ \mathcal{Q}_3^{bond} \circ \cdots \circ \mathcal{Q}_k^{bond} \circ \cdots \circ \mathcal{Q}_n^{bond} \tag{3}$$

$$\mathcal{Q}_P^{final} = \mathcal{Q}_P^{bond} \circ \mathcal{Q}_P^{angle} \circ \mathcal{Q}_P^{torsion} \circ \mathcal{Q}_P^{long-range} \tag{4}$$

Similarly, SMs for the ligand and the complex are given as:

$$\mathcal{Q}_L^{final} = \mathcal{Q}_L^{bond} \circ \mathcal{Q}_L^{angle} \circ \mathcal{Q}_L^{torsion} \circ \mathcal{Q}_L^{long-range} \tag{5}$$

$$\mathcal{Q}_{PL}^{final} = \mathcal{Q}_P^{bond} \circ \mathcal{Q}_P^{angle} \circ \mathcal{Q}_P^{torsion} \circ \mathcal{Q}_P^{long-range} \circ \mathcal{Q}_L^{bond} \circ \mathcal{Q}_L^{angle} \circ \mathcal{Q}_L^{torsion} \circ \mathcal{Q}_L^{long-range} \circ \mathcal{Q}_{PL}^{long-range} \tag{6}$$

The Boltzmann factor SMs are obtained similarly:

$$\mathcal{Z}_P^{bond} = \mathcal{Z}_1^{bond} \circ \mathcal{Z}_2^{bond} \circ \mathcal{Z}_3^{bond} \circ \cdots \circ \mathcal{Z}_k^{bond} \circ \cdots \circ \mathcal{Z}_n^{bond} \tag{7}$$

$$\mathcal{Z}_P^{final} = \mathcal{Z}_P^{bond} \circ \mathcal{Z}_P^{angle} \circ \mathcal{Z}_P^{torsion} \circ \mathcal{Z}_P^{long-range} \tag{8}$$

$$\mathcal{Z}_L^{final} = \mathcal{Z}_L^{bond} \circ \mathcal{Z}_L^{angle} \circ \mathcal{Z}_L^{torsion} \circ \mathcal{Z}_L^{long-range} \tag{9}$$

$$\mathcal{Z}_{PL}^{final} = \mathcal{Z}_P^{bond} \circ \mathcal{Z}_P^{angle} \circ \mathcal{Z}_P^{torsion} \circ \mathcal{Z}_P^{long-range} \circ \mathcal{Z}_L^{bond} \circ \mathcal{Z}_L^{angle} \circ \mathcal{Z}_L^{torsion} \circ \mathcal{Z}_L^{long-range} \circ \mathcal{Z}_{PL}^{long-range} \tag{10}$$

From $\mathcal{Q}_k^{bond}$, the bond probability matrix of one specific atom pair *k,* through $\mathcal{Q}_P^{bond}$ the probability matrix of all protein atom pairs with bond constraints to $\mathcal{Q}_{PL}^{final}$ the probability matrix of all atom pairs in the protein-ligand system, all the matrices have the same size such that the size of the SMs is the sample size of the atom pair combinations of the protein-ligand system. The advantage of using a pointwise product instead of a tensor product is that the size of the final matrix can be controlled at the beginning of the computation.

We use the two sp^3 carbon -sp^3 carbon bond terms in propane as an example to further explain the construction of the SMs using replication and tiling of the randomized vectors. The Boltzmann factor and probability vectors for each of the two bonds were modeled as unscrambled arrays:

$$\mathcal{t}_k^{bond} = \begin{bmatrix} z_k^{bond}(r_1) \\ z_k^{bond}(r_2) \\ \vdots \\ z_k^{bond}(r_a) \\ \vdots \\ z_k^{bond}(r_t) \end{bmatrix} = \begin{bmatrix} e^{-\beta E_k^{bond}(r_1)} \\ e^{-\beta E_k^{bond}(r_2)} \\ \vdots \\ e^{-\beta E_k^{bond}(r_a)} \\ \vdots \\ e^{-\beta E_k^{bond}(r_t)} \end{bmatrix} \tag{11}$$

and

$$q_k^{bond} = \begin{bmatrix} q_k^{bond}(r_1) \\ q_k^{bond}(r_2) \\ \vdots \\ q_k^{bond}(r_a) \\ \vdots \\ q_k^{bond}(r_t) \end{bmatrix} \tag{12}$$

k indicates one sp^3 carbon-sp^3 carbon bond in propane and the discrete distance a goes from 1 through t and represent the distance increments. Disordered vectors are generated using random scrambling of the original vectors. An example of the randomly scrambled vector of the Boltzmann factor with the index number i is shown in Equation 13. For a vector with t elements in it, the maximum number of permutation is $t!$ (the maximum value of i). Each index number i in the scramble operation $scram(X)_i$ represents one certain arrangement order of elements in the vector.

$$scram(z_k^{bond})_i = \begin{bmatrix} z_k^{bond}(r_3) \\ z_k^{bond}(r_{t-4}) \\ \vdots \\ z_k^{bond}(r_{t-1}) \\ \vdots \\ z_k^{bond}(r_5) \end{bmatrix} \tag{13}$$

With $scram(z_k^{bond})_i$ created, the SM for the kth sp^3 carbon-sp^3 carbon bond Boltzmann factor can be created by replication and tiling of the $scram(z_k^{bond})_i$ s. For instance, to create a SM with 20 rows and 30 columns using a vector containing 5 Boltzmann factors (t=5), replication of the $scram(z_k^{bond})_i$ in the SM would be generated as:

$$\mathcal{Z}_k^{bond} = \begin{bmatrix} scram\left(\mathcal{t}_k^{bond}\right)_1 & scram\left(\mathcal{t}_k^{bond}\right)_5 & \cdots & scram\left(\mathcal{t}_k^{bond}\right)_{117} \\ scram\left(\mathcal{t}_k^{bond}\right)_2 & scram\left(\mathcal{t}_k^{bond}\right)_6 & \cdots & scram\left(\mathcal{t}_k^{bond}\right)_{118} \\ scram\left(\mathcal{t}_k^{bond}\right)_3 & scram\left(\mathcal{t}_k^{bond}\right)_7 & \ddots & scram\left(\mathcal{t}_k^{bond}\right)_{119} \\ scram\left(\mathcal{t}_k^{bond}\right)_4 & scram\left(\mathcal{t}_k^{bond}\right)_8 & \cdots & scram\left(\mathcal{t}_k^{bond}\right)_{120} \end{bmatrix} \tag{14}$$

with

$$scram(\mathcal{t}_k^{bond})_1 = \begin{bmatrix} z_k^{bond}(r_3) \\ z_k^{bond}(r_1) \\ z_k^{bond}(r_2) \\ z_k^{bond}(r_4) \\ z_k^{bond}(r_5) \end{bmatrix} \tag{15}$$

as an example of one randomly scrambled vector.

Each vector $\mathcal{t}_k^{bond}$ is assumed to contain 5 elements thus 4 scrambled vectors tiled in a column makes 20 as the row number thus 120 scrambled vectors assemble the SM $\mathcal{Z}_k^{bond}$ with 600 elements.

For the probability SM, the same scramble and replication processes are performed. Furthermore, a probability value and a Boltzmann factor value corresponding to the same discrete distance (r_a) are mapped to each other in the probability and Boltzmann factor SMs. In other words, scrambled vectors of probabilities and Boltzmann factors scrambled in the same way (the same index number *i*) are in the same position in both the probability and Boltzmann factor SMs ($\mathcal{Q}_k^{bond}$ and $\mathcal{Z}_k^{bond}$). The mapping of probability and Boltzmann factor vectors in the SM $\mathcal{Q}_k^{bond}$ and SM $\mathcal{Z}_k^{bond}$ is illustrated in Figure 1.

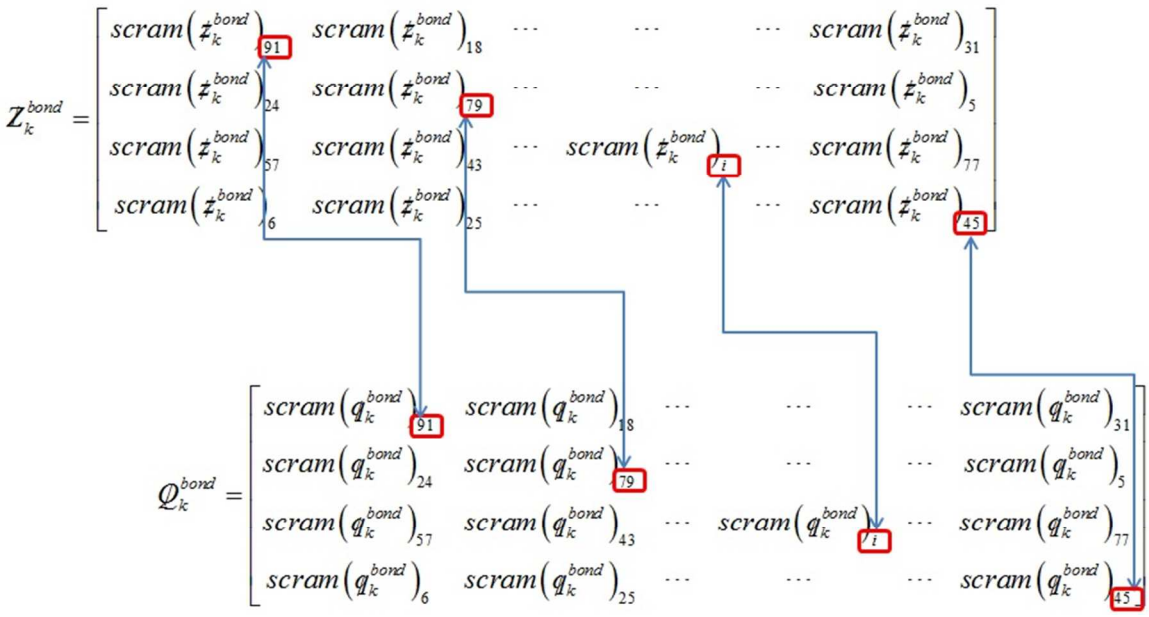

Figure 1. An example of the SM Z_k^{bond} and its corresponding SM Q_k^{bond}. The scramble operator index numbers with red circles connected by blue arrows indicate that scrambled vectors with the same scramble manner (the same index number *i*) in both Z_k^{bond} and Q_k^{bond} are tiled in the same position in both SMs.

Due to the mapping of $scram(z_k^{bond})_i$ and $scram\left(q_k^{bond}\right)_i$ in Z_k^{bond} and Q_k^{bond}, probabilities and Boltzmann factors of the same discrete distance (r_a) encounter each other in the pointwise product, because each index number *i* in the $scram(X)_i$ operation indicates a certain arrangement order of the elements in vector *X*. So that in the final SM, probabilities for each discrete distance (r_a) are assigned to the corresponding Boltzmann factors for the same (r_a).

Since the SM for the *k*th sp^3 carbon-sp^3 carbon bond Boltzmann factor is designed as in Equation 14, correspondingly, the probability SM is modeled as:

$$\mathcal{Q}_k^{bond} = \begin{bmatrix} scram\left(q_k^{bond}\right)_1 & scram\left(q_k^{bond}\right)_5 & \cdots & scram\left(q_k^{bond}\right)_{117} \\ scram\left(q_k^{bond}\right)_2 & scram\left(q_k^{bond}\right)_6 & \cdots & scram\left(q_k^{bond}\right)_{118} \\ scram\left(q_k^{bond}\right)_3 & scram\left(q_k^{bond}\right)_7 & \ddots & scram\left(q_k^{bond}\right)_{119} \\ scram\left(q_k^{bond}\right)_4 & scram\left(q_k^{bond}\right)_8 & \cdots & scram\left(q_k^{bond}\right)_{120} \end{bmatrix} \quad (16)$$

Thus Boltzmann factor and probability SMs for the kth (one of the three) sp^3 carbon-sp^3 carbon bond in propane are modeled. In $\mathcal{Z}_k^{bond}$ and $\mathcal{Q}_k^{bond}$, 120 scrambled vectors represent 120 different scrambled permutations of vector z_k^{bond} and q_k^{bond} tiled in a pattern from 1 through 120. SMs for lth ($1 \le l \le 2$, $l \in \mathrm{N}$, $l \ne k$) are modeled in a similar way while with different tiling sequences for the $scram(X)_i$ vectors in both SMs. For $\mathcal{Z}_l^{bond}$ and $\mathcal{Q}_l^{bond}$, tiling of the $scram(X)_i$ should use a different pattern. For instance a possible $\mathcal{Q}_l^{bond}$ with 120 scrambled vectors could be:

$$\mathcal{Q}_l^{bond} = \begin{bmatrix} scram\left(q_k^{bond}\right)_{124} & scram\left(q_k^{bond}\right)_{240} & \cdots & scram\left(q_k^{bond}\right)_{157} \\ scram\left(q_k^{bond}\right)_{168} & scram\left(q_k^{bond}\right)_{231} & \cdots & scram\left(q_k^{bond}\right)_{216} \\ scram\left(q_k^{bond}\right)_{200} & scram\left(q_k^{bond}\right)_{197} & \ddots & scram\left(q_k^{bond}\right)_{178} \\ scram\left(q_k^{bond}\right)_{145} & scram\left(q_k^{bond}\right)_{135} & \cdots & scram\left(q_k^{bond}\right)_{132} \end{bmatrix} \quad (17)$$

As we have mentioned, the maximum permutation number for a vector with t elements is $t!$. So there are around 10^{30} scrambled permutations with a bond vector containing about 30 elements, with which we could easily design thousands of $\mathcal{Z}_l^{bond}$ and $\mathcal{Q}_l^{bond}$ of a certain atom type pair using different tiling patterns. Using different tiling patterns for different atom type pairs increases the mix and match diversity of atom pairs at different discrete distance values (r_a) in the MT computation, and maximizes the degrees of

freedom of the elements (shown in Equation 18) in the pointwise product of the SMs of these two atom pairs (k and l).

$$\mathcal{Q}_k^{bond} \circ \mathcal{Q}_l^{bond} \circ \mathcal{Z}_k^{bond} \circ \mathcal{Z}_l^{bond} = \begin{bmatrix} q_k^{bond}(r_5) q_l^{bond}(r_3) z_k^{bond}(r_5) z_l^{bond}(r_3) & \cdots & q_k^{bond}(r_1) q_l^{bond}(r_4) z_k^{bond}(r_1) z_l^{bond}(r_4) \\ q_k^{bond}(r_2) q_l^{bond}(r_4) z_k^{bond}(r_2) z_l^{bond}(r_4) & \cdots & q_k^{bond}(r_3) q_l^{bond}(r_1) z_k^{bond}(r_3) z_l^{bond}(r_1) \\ \vdots & \ddots & \vdots \\ q_k^{bond}(r_3) q_l^{bond}(r_5) z_k^{bond}(r_3) z_l^{bond}(r_5) & \cdots & q_k^{bond}(r_4) q_l^{bond}(r_2) z_k^{bond}(r_4) z_l^{bond}(r_2) \end{bmatrix} \quad (18)$$

Using this replication and tiling scheme, the chance of element duplication in the final SM is extremely small due to the pointwise product of all atom pairwise SMs, thus maximizing the sampling size with a predetermined SM size. However, the size of the SMs is not arbitrary. Probability and Boltzmann factor vectors in each SM are randomly permutated and tiled, so that each element in the final matrices (i.e. $\mathcal{Q}_L^{final}$, $\mathcal{Q}_P^{final}$, $\mathcal{Q}_{PL}^{final}$, $\mathcal{Z}_L^{final}$, $\mathcal{Z}_P^{final}$, $\mathcal{Z}_{PL}^{final}$ in the equations above) is a probability or Boltzmann factor of one energy state in the protein-ligand system from a random combination of the chosen atom pairwise probabilities and Boltzmann factors. This indicates that with a fixed SM size, each time the pointwise product calculation is carried out it would generate different free energy values due to the random combination employed. Hence a SM size that ensures the convergence of the final free energy values is necessary for this pointwise approximation computation scheme to work effectively.

With the SM row number g fixed at 700 in order to be divisible by all vectors, 1, 1000, 10^8, 10^{13} were selected as the SM column number h in order to generate 700, 7×10^5, 7×10^{10} and 7×10^{15} sampling sizes for the final SMs. In order to test the convergence of

the free energy calculation using the SM pointwise product with different sizes, the binding free energy of one protein-ligand complex was calculated 100 times for each SM size, and RMSDs of the resultant binding free energies were collected for the four SM sizes. The protein-ligand complex with PDB ID 1LI2 was chosen for the test calculation. Binding affinity (pK_d) RMSDs for the four SM sizes are listed in Table 1.

Table 1. pK_d RMSDs for 100 rounds binding affinity calculations against the protein-ligand complex 1LI2 using the SM pointwise product with four different SM sizes.

SM sizes	700	7×10^5	7×10^{10}	7×10^{15}
pK_d RMSD	0.059	0.012	0.011	0.011

The test result shows that the pK_d RMSD for SM sizes of 7×10^5, 7×10^{10} and 7×10^{15} only differ by 0.001 and they all generate very low RMSDs (0.012, 0.011 and 0.011). We concluded that MT calculations with sample sizes of 7×10^5 is sufficient to ensure free energy convergence.

Using the pointwise product approximation, a protein-ligand complex would create several thousand SMs on average. For a laptop with a Intel(R) Core(TM) i7 CPU with 8 cores at 1.73GHz and 8Gb of RAM, it takes 6 seconds to calculate the pose and binding affinity for the protein-ligand complex 1LI2 and on average less than a minute to calculate the pose and binding free energy of one of the 795 protein-ligand complexes studied herein. If the SM size is increased to 7×10^{10}, the computation time required for 1LI2 increases to 8 minutes and on average it increases to around 20 minutes using the same laptop. Hence, this approach is faster than using MD or MC simulations to collect

the energies of 7×10^{5} to7×10^{10} protein-ligand poses. Future speed-ups are clearly possible using state of the art CPUs and GPUs and this is work that is underway.