# Accurate Molecular-Orbital-Based Machine Learning Energies via Unsupervised Clustering of Chemical Space

Lixue Cheng,[1] Jiace Sun,[1] and Thomas F. Miller III[2]
[1)]*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA*
[2)]*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA, tfm@caltech.edu*

(Dated: 6 July 2022)

## I. MO TYPE DETERMINATION

**Algorithm S1** MO type determination algorithm

**Input:** MO coordinates $M_1,\ldots,M_N$, atoms coordinates $A_1,\ldots,A_n$, expected number of bonds for atoms $\hat{S}_1,...,\hat{S}_n$
**Output:** Success or Failure of the process, a set of MO type descriptors $\{(I_{k,1},I_{k,2},BO_k)\}_{k=1,...,N}$ if the process is successful

**for** $k \leftarrow 1$ to $N$ **do**
  Compute $\{D_k^i\}_{i=1,\ ...,\ n}$
  $\alpha_k,\beta_k \leftarrow$ First two $i$ sorted by increasing order of $\{D_k^i\}_{i=1,\ ...,\ n}$
  **if** $\theta_k < 72^\circ$ **then**
    $T_k \leftarrow$ L
  **else**
    $T_k \leftarrow$ B ▷ Temporarily classify them to be bonds, may be changed to lone pair later
  **end if**
**end for**
Compute $S_1,\ ...,\ S_n$ from $T_1,\ ...,\ T_N$ ▷ Initialization of $\{S_i\}$
**while** $(\exists S_i > \hat{S}_i)$ **do**
  **if** $(S_i > \hat{S}_i)$ **then**
    **return** Failure
  **end if**
  randomly pick $u \in \{i|S_i > \hat{S}_i\}$
  $p \leftarrow \arg\max_k \theta_k$, subject to $T_k = B, u \in \{\alpha_k,\ \beta_k\}$
  $(T_p,\ S_{\alpha_p},\ S_{\beta_p}) \leftarrow (L, S_{\alpha_p}-1,\ S_{\beta_p}-1)$
**end while**
**for** $k \leftarrow 1$ to $N$ **do**
  **if** $T_k = B$ **then**
    $I_{k,1} \leftarrow \alpha_k,\ I_{k,2} \leftarrow \beta_k,\ BO_k = \#\{j|\{\alpha_j,\beta_j\} = \{\alpha_k,\beta_k\},\ T_j = B\}$ ▷ Unordered pair, i.e. $\{a,b\} = \{b,a\}$
  **else**
    $I_{k,1} \leftarrow \alpha_k,\ I_{k,2} \leftarrow$ None, $BO_k = 0$
  **end if**
**end for**
**return** Success, $\{(I_{k,1},I_{k,2},BO_k)\}_{k=1,...,N}$

The raw atomic connectivity of each MO is identified by searching the two atoms which have the smallest Euclidean distances to the centroid of the corresponding MO. For each MO, we assume that its final atomic connectivity can only be in two cases. If this MO is a bond, then two selected atoms are connected; and if this MO is a lone pair, it only belongs to the atom with a smaller distance to its centroid. To judge the MO identity (a bond or a lone pair), we define "atom-bond angle", i.e., $\angle ACB$, where $C$ is the centroid position of this bond, and $A,B$ are its two nearest atoms. Ideally, the center of the bond between two atoms should be collinear with these two atoms, i.e., the atom-bond angle is 180°. The final atomic connectivity is determined by iteratively classifying the MOs with small atom-bond angles as lone pairs until all atoms satisfy the octet rule in chemistry. The bond order of each MO is computed by the number of bonds between the two corresponding detected atoms.

Algorithm S1 states the details of determination process of MO types of a closed-shell molecule using the MO centroid coordinates $\{M_1,...,M_N\} \in \mathbb{R}^3$ and the atom coordinates $\{A_1,...,A_n\} \in \mathbb{R}^3$. Additionally, for each atom $a_i$ with a certain number of connected bonds, we define $\hat{S}_i$ as the expected number of bonds connected to each atom $i$ (i.e., 1 for H, 2 for O, 3 for N, 4 for C, 1 for Cl, and not defined for S) also as part of the algorithm input. The output of this algorithm is the atomic connectivity

represented as tuple $(I_{k,1}, I_{k,2}, BO_k)$ for each MO $k$, where $I_{k,1}$, $I_{k,2}$ are the two connected atoms of the MO $k$, and $BO_k$ is its bond order.

For each MO $k$, a boolean variable $T_k$ is introduced to determine if the MO $k$ is a bond ($B$) or a lone pair ($L$). We initialize the atomic connectivity of the MO $k$ as the atoms $\alpha_k$ and $\beta_k$, which are equal to the indices of the first and second smallest elements in $\{D_k^i | i = 1, ..., n\}$, where $D_k^i$ is the euclidean distance between $M_k$ and $A_i$. We define the atom-bond angle of the MO $k$ as $\theta_k = \angle \alpha_k M_k \beta_k$, which tends to be large for bond because it is 180° in the ideal case. For the MO $k$, we initialize $T_k = L$ if $\theta_k < 72°$, and $T_k = B$ if $\theta_k > 72°$, because 72° is small enough that, for any MO $k$, $\theta_k < 72°$ guarantees $T_k = L$. The number of the bonds connected to each non-sulfur atom $i$, i.e., $S_i$, is computed. We iteratively converge $\{S_i\}$ by decreasing the values until $S_i = \hat{S}_i$ for each non-sulfur $i$ (which leads to "success"), or there is at least one $S_i < \hat{S}_i$ so that $\{S_i\}$ is no longer possible to agree with $\{\hat{S}_i\}$ (which leads to "failure"). We note that sulfur is not checked because it is more complicated than the rest types of atoms. In each iteration, an atom $u$ satisfying $S_u > \hat{S}_u$ is selected randomly, and then we find the set of bonds connected to $u$. We change $T_p$ with the smallest $\theta_p$ in this set to lone pair, i.e., $T_p = L$, and update $S_{\alpha_p}$ and $S_{\beta_p}$ by decreasing one. After the iteration finishes, each $T_k$ has been successfully determined, so we can now determine $(I_{k,1}, I_{k,2})$ as $(\alpha_k, \beta_k)$ if $T_k = B$, or as $(\alpha_k, \text{None})$ if $T_k = L$. Finally, for each MO $k$ that $T_k = B$, the bond order $BO_k$ can be determined by the number of bonds having the same unordered pair $(\alpha_k, \beta_k)$ with it, which finishes the algorithm.

Since randomness is introduced in the algorithm, we repeat the algorithm several times until success, or it fails more than ten times so that we believe a solution cannot be found. In this work, all the MO types of 99.9% of QM7b-T and 98.7% of GDB-13-T molecules have been successfully recognized without any contradictions to the octet rule. The molecules with at least one atom violating the octet rule are excluded only in the analysis of unsupervised clustering on organic chemical space but included in the energy predictions.

## II. MOLECULAR ENERGY LEARNING WITH UNSUPERVISED CLUSTERING

In Table S1 and S2, we summarize the MAEs of molecular energies in kcal/mol using different regression with clustering protocols plotted in Fig. 4 in the main text for QM7b-T and GDB-13-T, respectively.

TABLE S1. MOB-ML prediction accuracy (kcal/mol) for four regression with clustering protocols applied to MPC/cc-pVTZ energies of QM7b-T. The training and testing sets corresponding to non-overlapping subsets of QM7b-T.

| Training sizes | GMM/GPR/10X | RC/GPR/10X | GMM/LR/10X | RC/LR/10X |
|---|---|---|---|---|
| 50 | 1.187 | 1.289 | – | – |
| 100 | 0.788 | 0.968 | 1.156 | 1.344 |
| 250 | 0.579 | 0.648 | 0.889 | 0.800 |
| 500 | 0.429 | 0.520 | 0.718 | 0.605 |
| 1000 | 0.313 | 0.468 | 0.549 | 0.531 |
| 1500 | 0.270 | 0.416 | 0.499 | 0.479 |
| 2000 | 0.239 | 0.383 | 0.445 | 0.456 |
| 2500 | 0.219 | 0.373 | 0.414 | 0.446 |
| 4000 | 0.189 | 0.338 | 0.375 | 0.420 |
| 5000 | 0.169 | 0.325 | 0.362 | 0.413 |
| 6500 | 0.157 | 0.315 | 0.359 | 0.407 |

## III. COMPARISON BETWEEN TRAINING COSTS OF SUPERVISED CLUSTERING AND UNSUPERVISED CLUSTERING

Figure S1 plots the wall-clock timing of RC and GMM clustering on 8 NVIDIA Tesla V100-SXM2-32GB GPUs as functions of the number of training sizes. The costs of RC and GMM are similar across all the training sizes. GMM leads to more accurate clustering results but does not require higher computational costs.

## IV. SOFT CLUSTERING FROM GMM

GMM provides probabilities of every possible cluster for a test point, the predictions of molecular energies with soft clustering are possible. The prediction $\varepsilon_{ij,soft}^{ML}$ of each pair energy is a weighted average of the predictions $\varepsilon_{ij,n}^{ML}[\mathbf{f}_{ij}]$ from all $K$ possible

TABLE S2. MOB-ML prediction accuracy (kcal/mol) for four regression with clustering protocols applied to MP2/cc-pVTZ energies of GDB-13-T. Models are the same as the ones in Table S1

| Training sizes | GMM/GPR/10X | RC/GPR/10X | GMM/LR/10X | RC/LR/10X |
|---|---|---|---|---|
| 50 | 1.286 | 1.428 | – | – |
| 100 | 0.945 | 1.145 | 1.383 | 1.477 |
| 250 | 0.873 | 0.980 | 1.120 | 1.086 |
| 500 | 0.702 | 0.795 | 0.968 | 0.830 |
| 1000 | 0.613 | 0.781 | 0.873 | 0.752 |
| 1500 | 0.577 | 0.743 | 0.865 | 0.709 |
| 2000 | 0.549 | 0.715 | 0.809 | 0.696 |
| 2500 | 0.521 | 0.701 | 0.809 | 0.685 |
| 4000 | 0.498 | 0.608 | 0.770 | 0.637 |
| 5000 | 0.466 | 0.591 | 0.763 | 0.626 |
| 6500 | 0.462 | 0.573 | 0.760 | 0.613 |

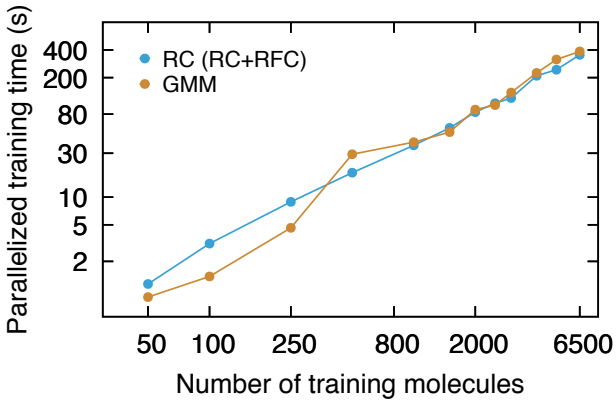


FIG. S1. Wall-clock timings for RC+RFC and GMM for different number of training molecules from the QM7b-T set with 8 NVIDIA Tesla V100 GPUs. The LR regression of each cluster is performed independently on a different core in the RC step, and RFC is trained using parallel implementation of SCIKIT-LEARN. GMM is trained using parallel implementation of EM algorithm

clusters evaluated as Eq. S1.

$$\varepsilon^{ML}_{ij,soft} = \sum_{n=1}^{K} P_n[\mathbf{f}_{ij}] * \varepsilon^{ML}_{ij,n}[\mathbf{f}_{ij}], \tag{1}$$

where $\mathbf{f}_{ij}$ is the set of MOB features for pair $ij$, $n = 1,2,...,K$ is the cluster ID; and $P_n$ is the corresponding probability of the pair $ij$ being classified into cluster $n$ satisfying $\sum_{n=1}^{K} P_n = 1$.

Figure S2 displays the Comparison of the prediction accuracies of molecular energies regressed by LR on top of hard and soft clustering from the same set of GMMs for QM7b-T and GDB-13-T. In both panels, the results from the soft clustering method (GMM, soft/LR) overlap with the ones from hard clustering (GMM/LR) with accuracy differences smaller than 0.003 kcal/mol, which suggests that soft clustering does not provide any extra benefits in this application. Since we create an interpolation to a weak extrapolation problem, the cluster identities of the tests points are therefore unambiguous. Table S3 shows the percentages of pairs with more than one cluster with a probability higher than 0.0001, namely, how many pairs can be influenced by adapting the soft clustering method during the predictions of energies. For QM7b-T, under 10% of pairs in both diagonal and off-diagonal feature spaces have more than one possible cluster. The numbers of pairs with more than one possible cluster identity increase for GDB-13-T but are still not significant enough to change the predicted energies in Fig. S2. We note that soft clustering from GMM might provide some accuracy improvements in some future applications.

Table S3 shows the ratio of the pairs in the entire QM7b-T and GDB-13-T having predicted probability over 1e-4 for at least two clusters by the GMM models. For QM7b-T, the percentages of points that might have different predictions by using soft clustering are very low, and thus hard and soft clustering provide similar results. For example, there are over 10% of the off-diagonal pairs that have second probable clusters, but as shown in Fig. S2, the prediction accuracy is not changed by these points too. This could be attributed to the fact that multiple clusters might map to the same types of MOs, and the predictions will not be affected by changing the cluster identity from one to the other.

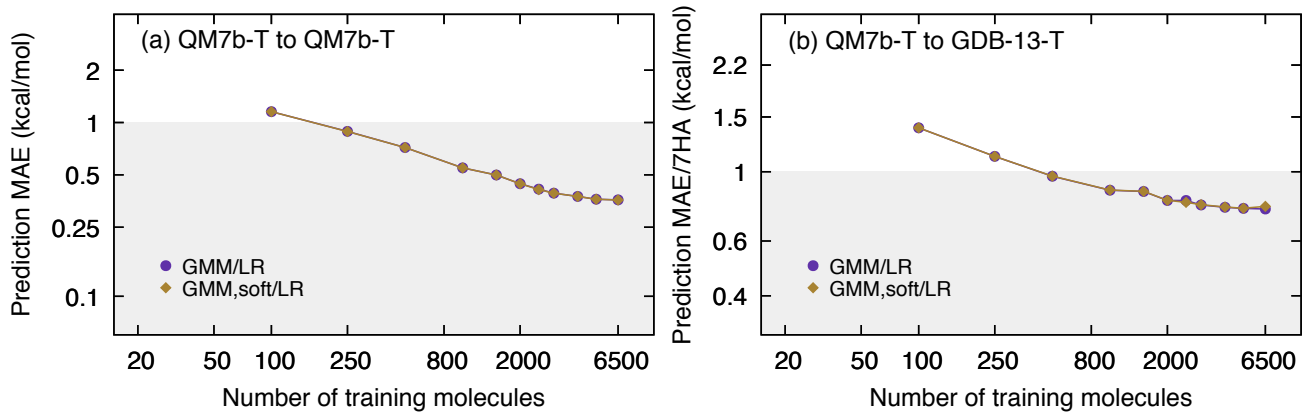


FIG. S2. Comparison between hard clustering and soft clustering on molecular energy learning regressed by LR for (a) QM7b-T and (b) GDB-13-T. The results from hard clustering (purple circles) and the ones from soft clustering (dark gold diamond) overlap well. All the data are plotted on a logarithmic scale, and the shaded areas correspond to an MAE/7HA of 1 kcal/mol.

TABLE S3. Percentages of pairs in QM7b-T and GDB-13-T having n number of clusters that their predicted probability over 0.0001 by GMMs

| Pair type | GMM training size | QM7b-T | | GDB-13-T | |
|---|---|---|---|---|---|
| | | n =2 | n≥3 | n =2 | n≥3 |
| Diagonal | 250 | 0.37% | 0 | 0.69% | 0 |
| | 1000 | 6.29% | 0 | 8.83% | 0 |
| Off-diagonal | 250 | 8.74% | 0.83% | 20.42% | 1.85% |
| | 1000 | 6.40% | 1.21% | 10.63% | 3.99% |