Supporting Information

A Neural Network Protocol for Electronic excitations of N-Methylacetamide

Sheng Ye[a,1], Wei Hu[b,1], Xin Li[a,1], Jinxiao Zhang[a], Kai Zhong[a], Guozhen Zhang[a], Yi Luo[a], Shaul Mukamel[c,d,2] and Jun Jiang[a,2]

[a]Hefei National Laboratory for Physical Sciences at the Microscale, CAS Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China

[b]Shandong Provincial Key Laboratory of Molecular Engineering, School of Chemistry and Pharmaceutical Engineering, Qilu University of Technology, Jinan, Shandong 250353, P. R. China

[c]Departments of Chemistry, University of California, Irvine, CA 92697

[d]Department of Physics and Astronomy, University of California, Irvine, CA 92697

[1]S.Y., W.H. and X.L. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: jiangj1@ustc.edu.cn

Table of Contents

Computational details

Molecular dynamics simulations. Molecular dynamics Simulations with 1 fs time step at temperatures of 200K, 300K, and 400K were performed using the GROMACS code with NPT ensemble and OPLS-AA force fields. Periodic boundary conditions were imposed on a 30.2 Å cubic box containing one NMA and 875 TIP4P water molecules (1). Coulomb interactions were truncated at 12.0 Å and a shift function was used for vdW forces with the same cutoff. The bond lengths were constrained by the LINCS methods (2). Electrostatic interactions were treated by the Particle mesh Ewald method (3). At T=300K, we generated two independent trajectories of 2 ns and 10 ns, from which 10000 and 50000 configurations were extracted with a 200 fs interval, respectively. At T=200K and 400K, we generated 2 ns trajectories, from which 10000 configurations were harvested with the same time interval. The equilibrated structures were used for including quantum chemistry calculations and NN learning.

The machine learning protocol. The NN consists of one input layer, three hidden layers and one output layer. For each hidden NN layer we used the Rectified Linear Unit activation function (4). The 50000 sets of data at 300K were randomly divided into two subsets: 40000 were used for training and the rest (10000) were used for testing (Additional 10000 sets at 300K were randomly divided into 7000 and 3000 for training and testing, respectively). And the 10000 sets of data at 200K and 400K were randomly selected 5000 for testing the NN model obtained in 300K. In order to compare the prediction ability of the map method and NN, we take some strongly-deviated structures to predict, in which the red dots represents the normal NMA structure, and the blue dots represents the strongly-deviated structures (Fig. S2 C-F). Then the NN were subjected to a supervised training scheme using a back propagation algorithm implemented in TensorFlow frame. (5).

We have taken the following steps to mitigate the over-fitting issue in the neural network training process:
(1) The size of dataset was increased from 10,000 to 50,000 data points.
(2) For the selection of the descriptors, we calculated Pearson correlation coefficient (r) among the descriptors. We find that most descriptors have a low linear correlations (Fig. S4), which significantly reduces the over-fitting problem.
(3) For the training, we added L2 regularization (6) to the structure of the neural network to prevent overfitting.

To avoid the use of raw variables with different range of values which may undermine the robustness NN results, we firstly normalized the input features i to reduce the dimensional inconsistency, *i.e.*, converted data to the dimensionless data in range 0 to 1. This is because different raw variables with remarkably different range of values can severely undermine the robustness of the result generated by neural network. Therefore, to eliminate the dimensional impact between the input data, data normalization is required to resolve the comparability of data.

The data was transformed with $x' = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$, where x_i are input data, x' are normalized data,

and x_{min} and x_{max} are minimum and maximum values of the input data, respectively.

The mean relative error (*MRE*) was computed with $MRE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$, where A_t is the actual value and F_t is the predicted value.

The cross-validation technique (7) was employed to verify the accuracy and robustness of the final NN results. In the cross-validation procedure, *N* sets of data were randomly and evenly distributed into 10 bins. Each bin was used as a test set with the remaining nine bins as training sets.

Importance analysis: Random forest is a popular machine learning algorithm. A multitude of decision trees are constructed for classification, regression and other tasks (8). In addition, it can be used for analyzing the importance of each descriptor which is calculated as follows (8):
As each tree evolves, predictions are made based on the Out-Of-Bag (OOB) data for that tree. At the same time, each descriptor in the OOB data is randomly permuted, one at a time, and each such modified data set is also predicted by the same tree. At the end of the model training process, the margins for each sample are calculated based on the OOB prediction as well as the OOB predictions with each descriptor permuted. Let *M* be the average margin based on the OOB prediction and M_j the average margin based on the OOB prediction with the *j* th descriptor permuted. The difference between *M* and M_j (M-M_j) reflects the importance for the *j* th descriptor. For regression problems, addressed here.

The average maximum of the frequency. The average maximum of the frequency was used to compare the difference between of the UV absorption spectra at different temperatures of NMA by TDDFT and NN, it was computed with $\overline{\Omega} = \frac{\int d\Omega \bullet f(\Omega)\Omega}{\int d\Omega \bullet f(\Omega)}$, where Ω is the frequency and f is the oscillator strength.

Map (1)**:**

$$\omega^{\mu} = \omega_0^{\mu} + \sum_{i=1}^{4}\left(\alpha_i^{\mu} d_i + \beta_i^{\mu} d_i^2\right) + \sum_{i=5}^{8}\left(\alpha_i^{\mu}\theta_i + \beta_i^{\mu}\theta_i^2\right) + \alpha_9^{\mu}\cos(\phi) + \beta_9^{\mu}\cos^2(\phi)$$

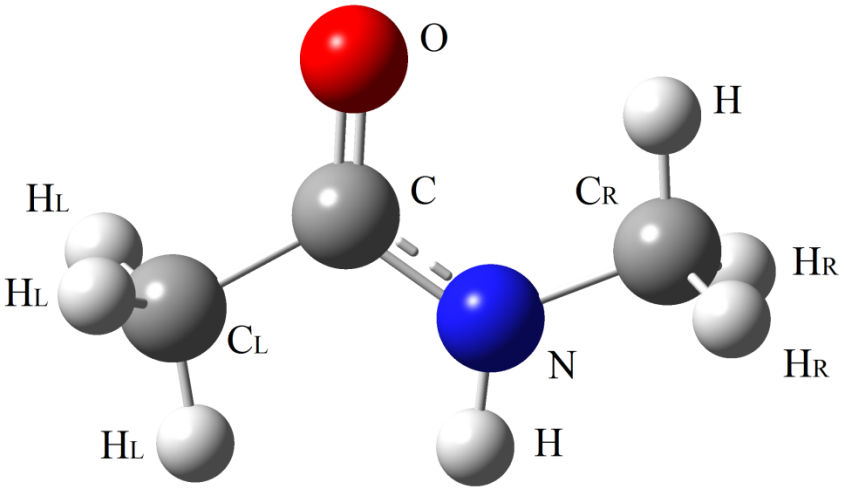
$$\mu = n\pi^*, \pi\pi^*$$

where the four bond lengths d_i are d_{CO}, d_{CN}, d_{C_LC}, and d_{NC_R}, the four angles θ_i are $\angle OCN$, $\angle CNH$, $\angle NCC_L$, and $\angle CNC_R$, and the dihedral angle ϕ is $\angle OCNH$.

The spectra was obtained by $f(\omega)=\sum_{i=1,n}\frac{f_{1,n}}{(\omega-\Omega_{1,n})^2+\gamma^2}$,where $f_{1,n}$ and $\Omega_{1,n}$ denotes the oscillator strength and frequency of electronic excitations respectively, and the n denotes the numbers of structures of NMA molecules, in our work n=5000.

Table S1. Time required to compute transition energies, dipole moments, transition dipole moments of 5000 frames for TDDFT (PBE0/cc-pVDZ) and NN.

Method	Transition Energy	Dipole Moment	Transition Dipole Moment
DFT (nπ*)	65000s	65000s	65000s
NN (nπ*)	24.63s	29.48s	231.16s
DFT (ππ*)	65000s	65000s	65000s
NN (ππ*)	61.33s	29.48s	65.96s



Bond lengths: $d_1=d_{CO}$, $d_2=d_{CN}$, $d_3=d_{CLC}$, $d_4=d_{NCR}$;

Angles: $\alpha_1=\angle OCN$, $\alpha_2=\angle CNH$, $\alpha_3=\angle NCC_L$, $\alpha_4=\angle CNC_R$, $\alpha_5=\angle HNC_R$, $\alpha_6=\angle OCC_L$;

Dihedral angles: $\beta_1=\angle OCNH$, $\beta_2=\angle OCNC_R$, $\beta_3=\angle C_LCNH$, $\beta_4=\angle C_LCNC_R$;

Fig. S1. Descriptors used for predicting transition energies.

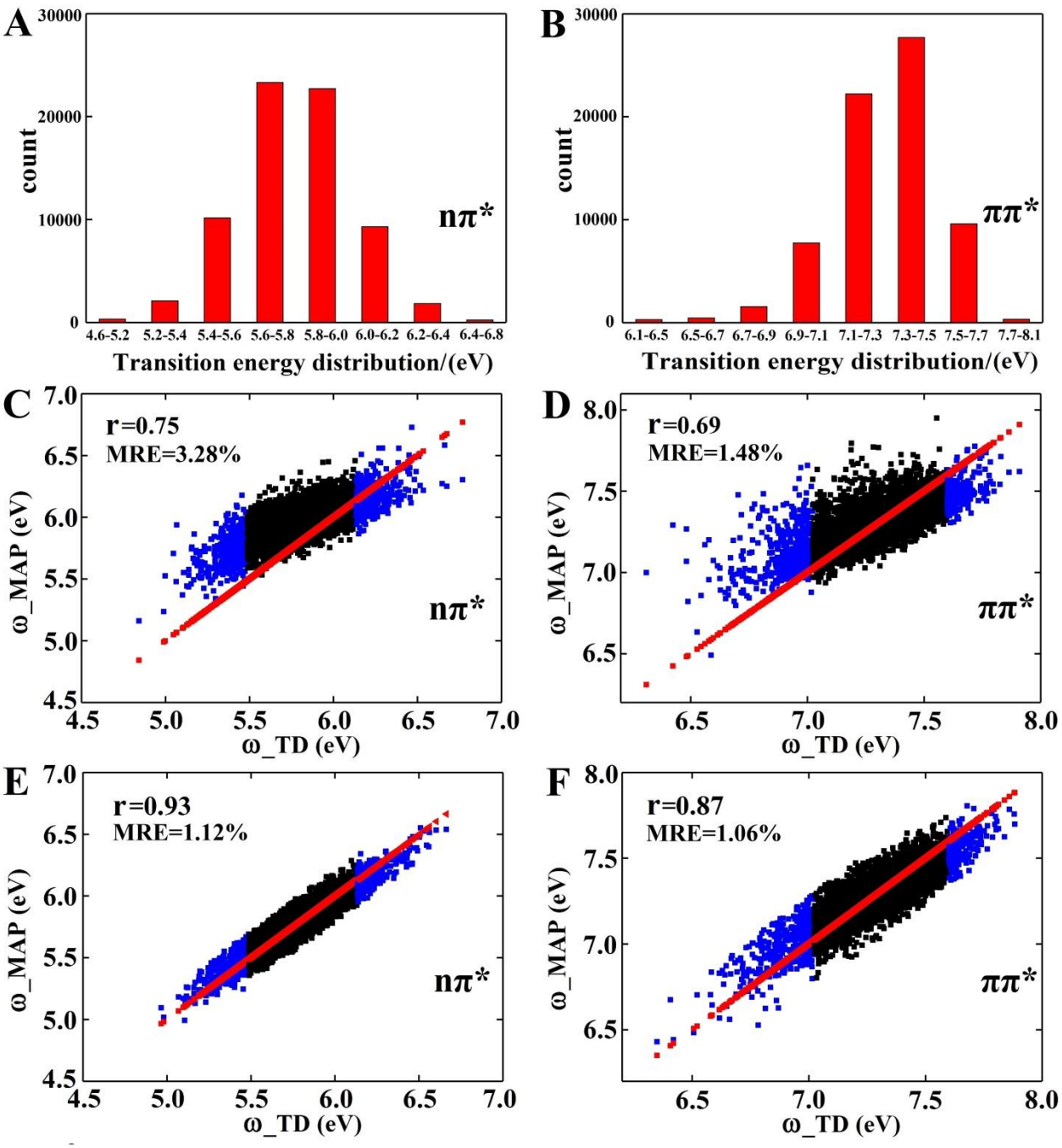


Fig. S2. (A) Distribution of NMA nπ* transition energies calculated by TDDFT. (B) Same as (A) but for the ππ* transition (C) Correlation plots of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP).(1) (D) Same as (C) but for the ππ* transition. (E) Comparison of nπ* transition energies by TDDFT (ω_TD) and neural network (ω_NN). (F) Same as (E) but for the ππ* transition. The red lines/dots represent the transition energies (ω_TD) of NMA calculated by TDDFT which performed at the PBE0/cc-pvdz level. Black points refer to those NMA structures close to minima, while blue points refer to those far from minima.

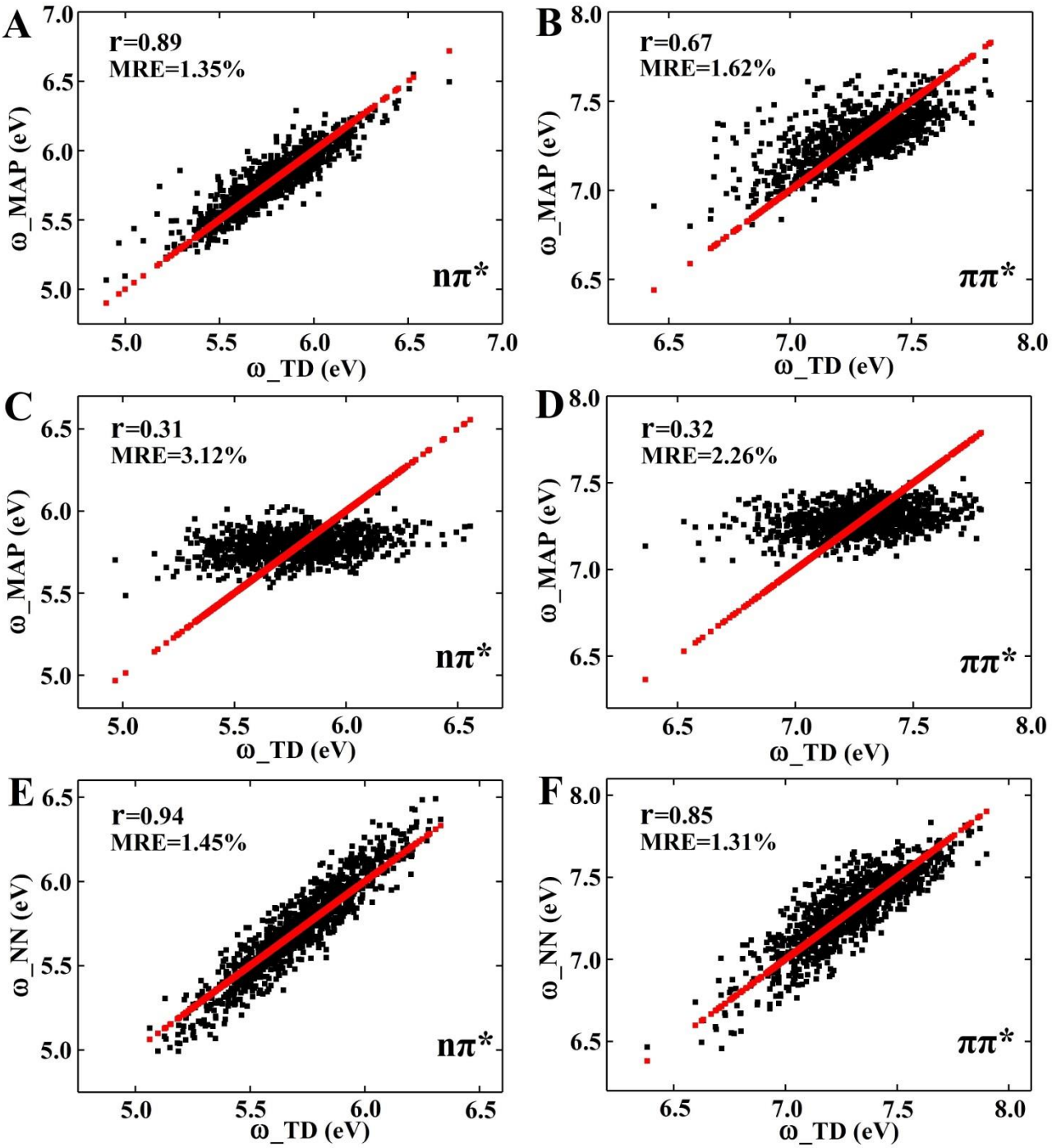


Fig. S3. (A) Correlation plots of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP).[1] (B) Same as (a) but for the ππ* transition. (C) Comparison of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP)[1] which fitted by different data sets. (D) Same as (C) but for the ππ* transition. (E) Comparison of nπ* transition energies by TDDFT (ω_TD) and neural network (ω_NN). (F) Same as (E) but for the ππ* transition. The red lines/dots on the figures represent the transition energies (ω_TD) of NMA calculated by TDDFT which performed at the (B3LYP/6-31G(d,p)) level.

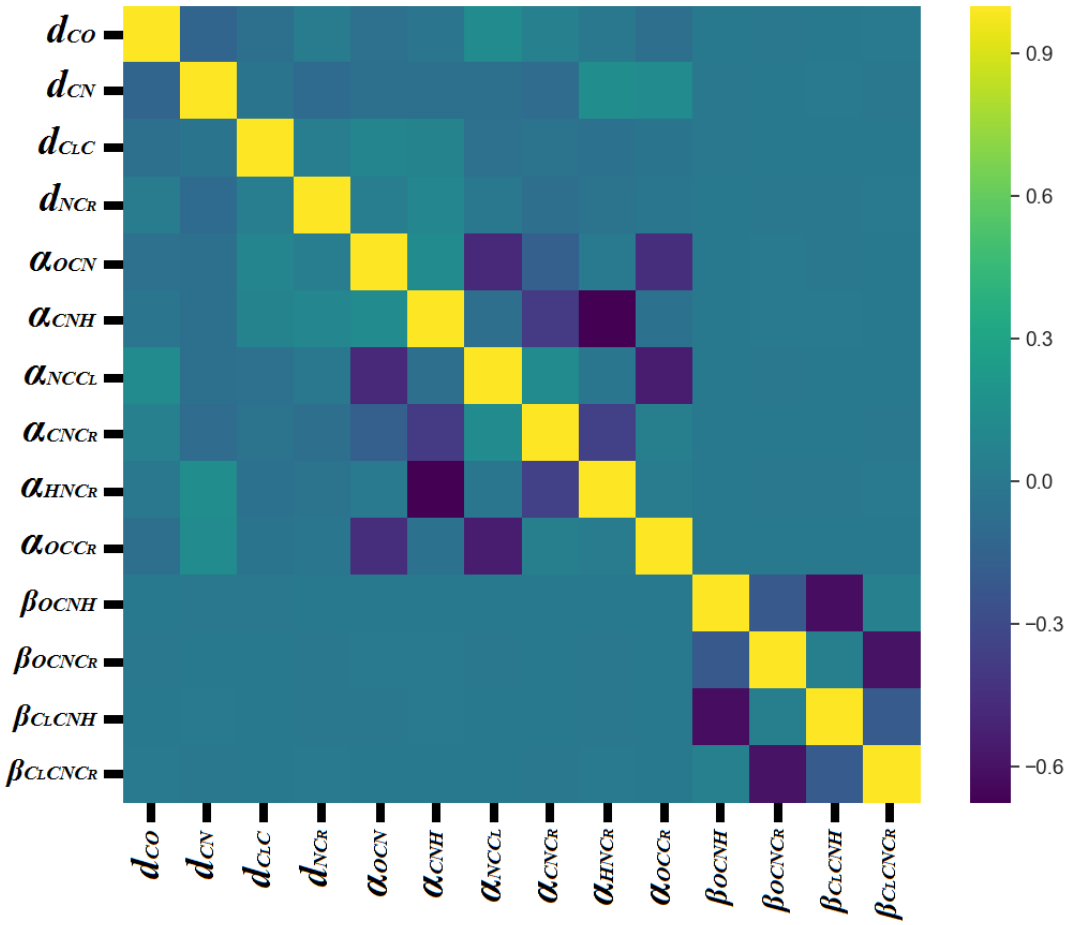


Fig. S4. Heat map of the Pearson correlation coefficient (r) among the descriptors for predicting the nπ* and ππ*transition energy.

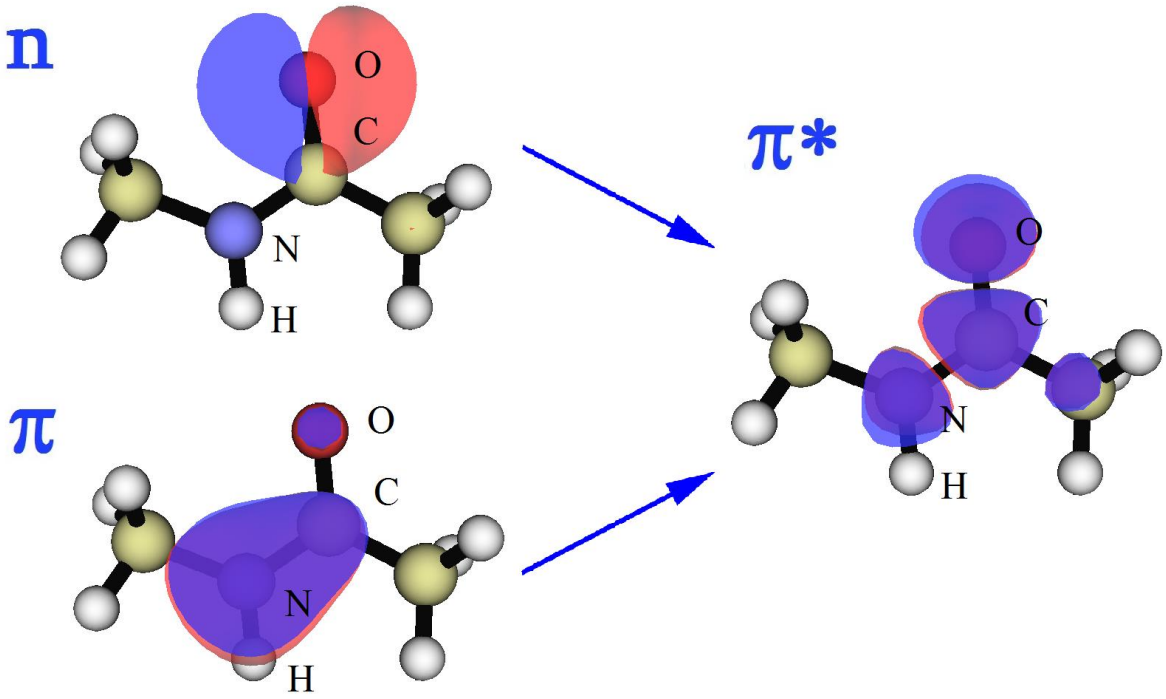


Fig. S5. The NMA molecular orbitals which after localizing analysis were included in the two transitions: nπ* and ππ* transition.

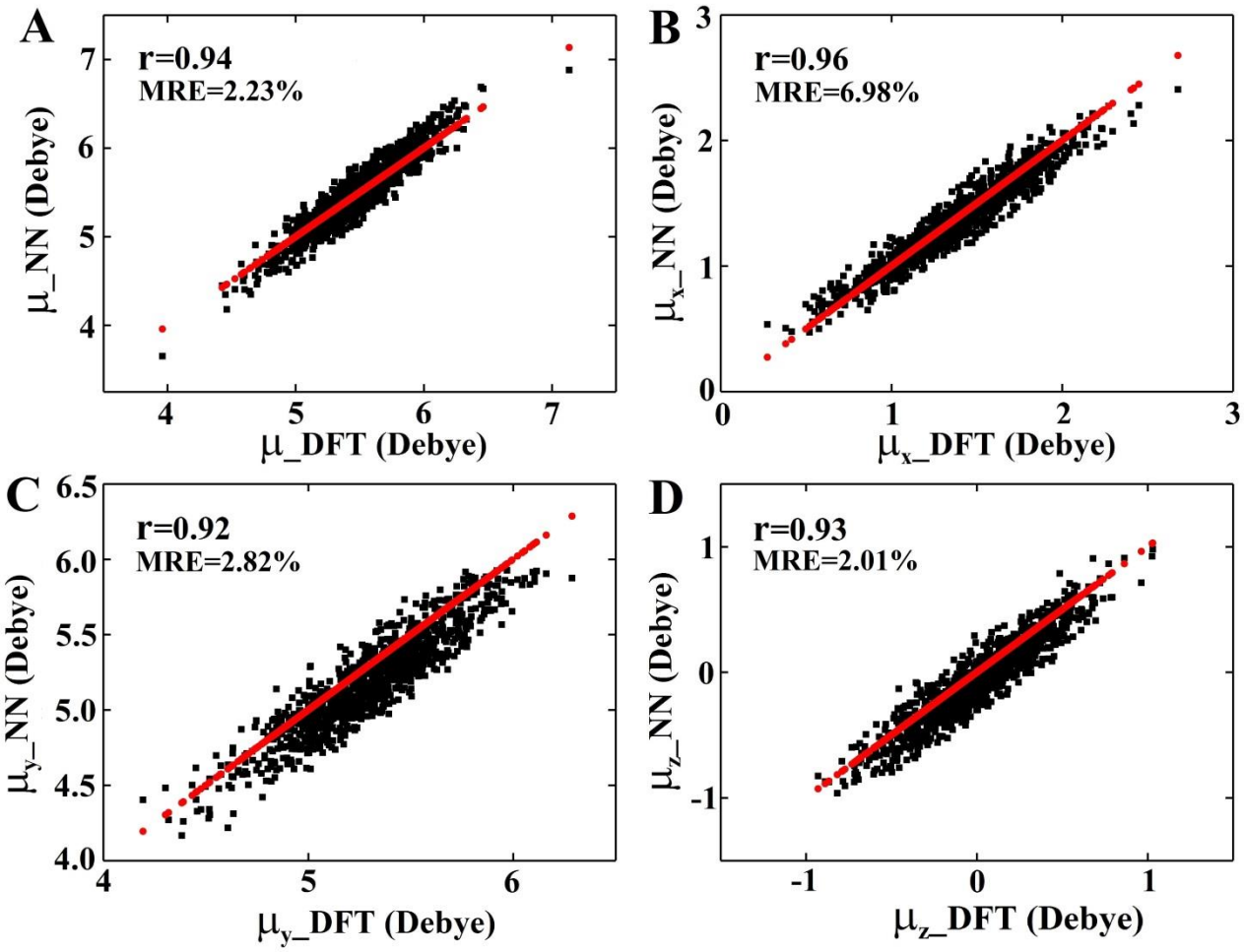


Fig. S6. (A) Correlation of dipole moment by DFT (μ_DFT) and NN (μ_NN). (B) Comparison of dipole moment in the *x* direction by DFT (μ_x_DFT) and NN (μ_x_NN). (C) Comparison of dipole moment in the *y* direction by DFT (μ_y_DFT) and NN (μ_y_NN). (D) Comparison of dipole moment in the z direction by DFT (μ_z_DFT) and NN (μ_z_NN). The red lines/dots on the figures represent μ_DFT, μ_x_DFT, μ_y_DFT and μ_z_DFT of NMA calculated by DFT which performed at the (B3LYP/6-311G++(d,p)) level, respectively.

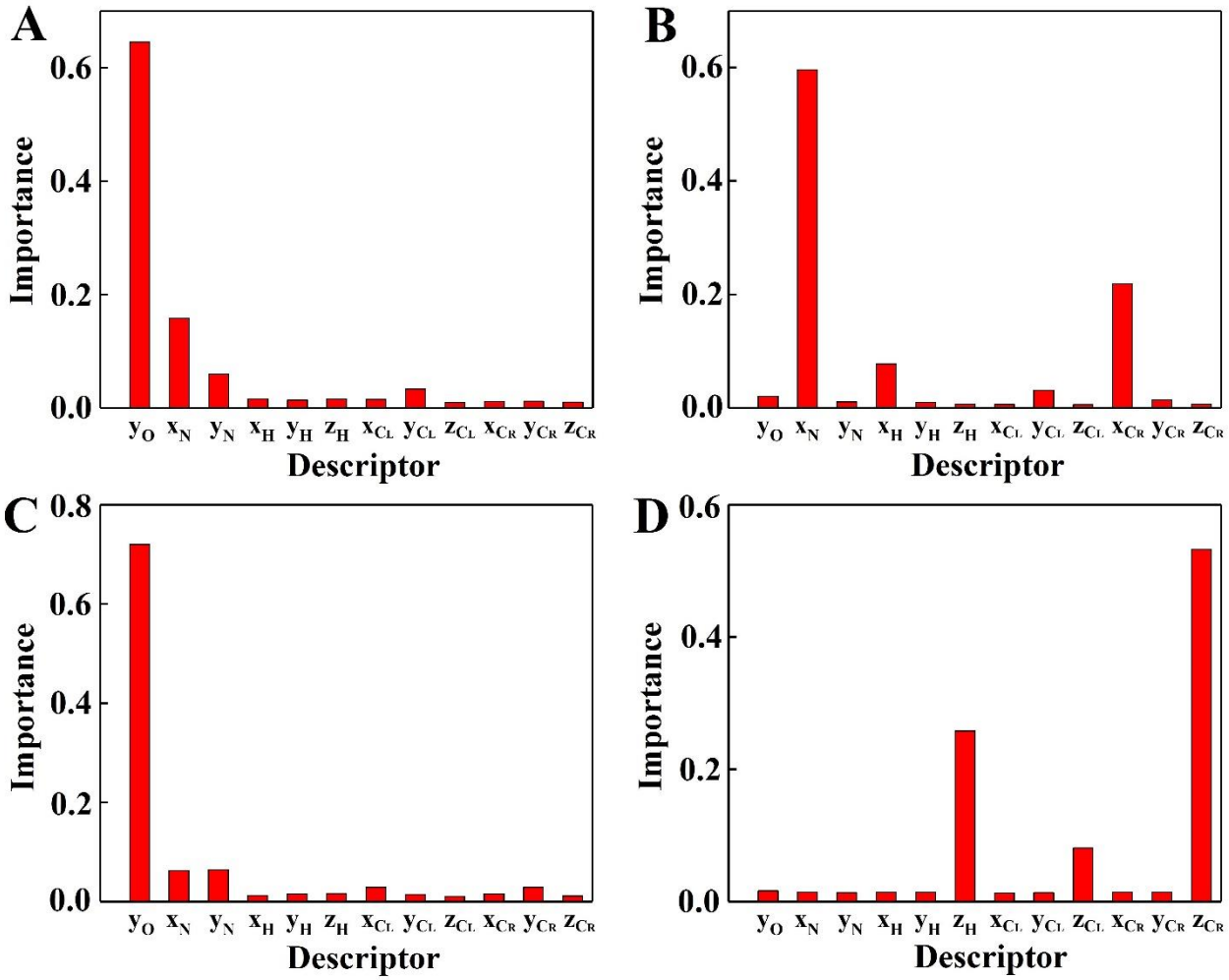


Fig. S7. (A) Descriptor importance analysis of dipole moment. (B) Descriptor importance analysis of dipole moments in *x* direction. (C) Descriptor importance analysis of dipole moments in *y* direction. (D) Descriptor importance analysis of dipole moments in *z* direction.

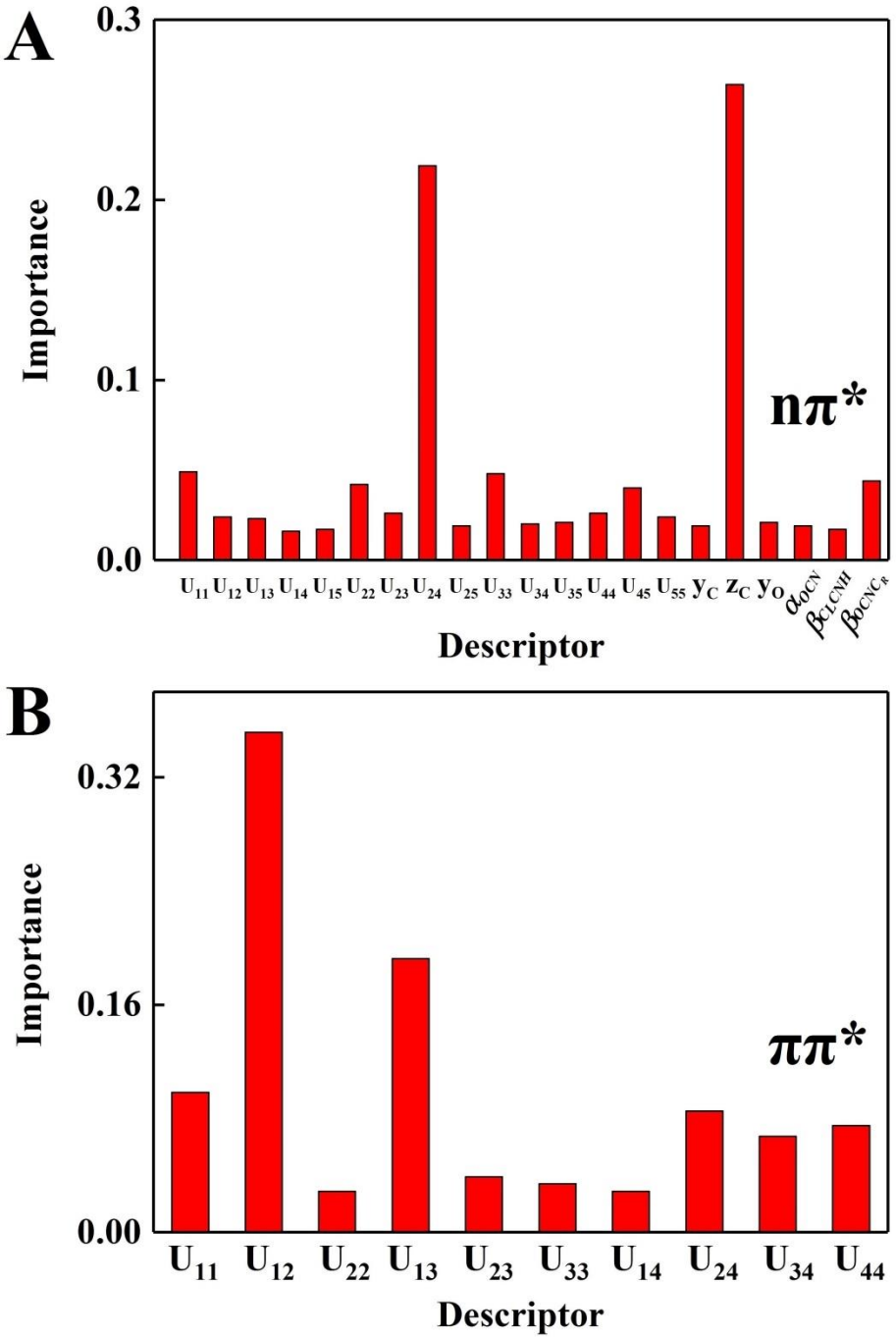


Fig. S8. (A) The importance of transition dipole moment descriptors for nπ*transition. (B) Same as (A) but for the ππ* transition.

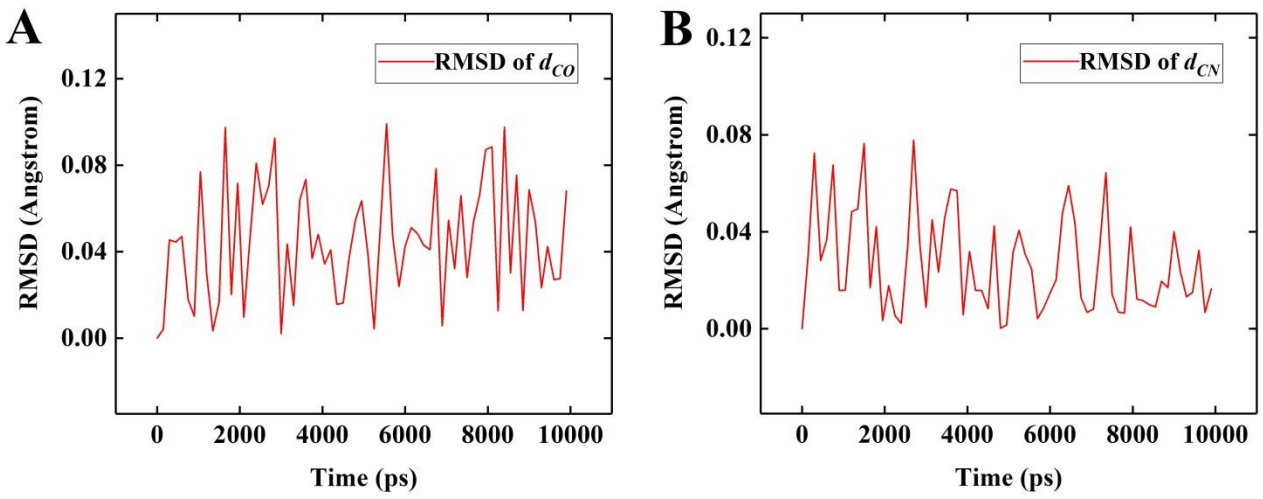


Fig. S9. (A) The root mean square deviation (RMSD) of CO bond. (B) Same as (A) but for the CN bond.

References

1. Li Z, Yu H, Zhuang W, & Mukamel S (2008) Geometry and excitation energy fluctuations of NMA in aqueous solution with CHARMM, AMBER, OPLS, and GROMOS force fields: Implications for protein Ultraviolet spectra simulation. *Chem Phys Lett* 452:78-83.
2. Hess B, Bekker H, Berendsen HJ, & Fraaije JG (1997) LINCS: a linear constraint solver for molecular simulations. *J Comput Chem* 18:1463-1472.
3. Darden T, York D, & Pedersen L (1993) Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *J Chem Phys* 98:10089-10092.
4. Maas AL, Hannun AY, & Ng AY (2013) Rectifier nonlinearities improve neural network acoustic models. *In Proc. ICML* vol. 30.
5. Abadi M, et al. (2016) Tensorflow: a system for large-scale machine learning. *OSDI*, pp 265-283.
6. Ng AY (2004) Feature selection, L 1 vs. L 2 regularization, and rotational invariance. in Proceedings of the twenty-first international conference on Machine learning.
7. Hansen K, et al. (2013) Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J Chem Theory Comput* 9:3404-3419.
8. Svetnik V, et al. (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 43:1947-1958.