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, Chinese Academy of Sciences Center for Excellence in Nanoscience, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China; [b]Shandong Provincial Key Laboratory of Molecular Engineering, School of Chemistry and Pharmaceutical Engineering, Qilu University of Technology, Jinan, 250353 Shandong, People's Republic of China; [c]Department of Chemistry, University of California, Irvine, CA 92697; and [d]Department of Physics and Astronomy, University of California, Irvine, CA 92697

Contributed by Shaul Mukamel, February 8, 2019 (sent for review December 13, 2018; reviewed by Jonathan D. Hirst and Jin Wang)

UV absorption is widely used for characterizing proteins structures. The mapping of UV spectra to atomic structure of proteins relies on expensive theoretical simulations, circumventing the heavy computational cost which involves repeated quantum-mechanical simulations of excited-state properties of many fluctuating protein geometries, which has been a long-time challenge. Here we show that a neural network machine-learning technique can predict electronic absorption spectra of *N*-methylacetamide (NMA), which is a widely used model system for the peptide bond. Using ground-state geometric parameters and charge information as descriptors, we employed a neural network to predict transition energies, ground-state, and transition dipole moments of many molecular-dynamics conformations at different temperatures, in agreement with time-dependent density-functional theory calculations. The neural network simulations are nearly 3,000× faster than comparable quantum calculations. Machine learning should provide a cost-effective tool for simulating optical properties of proteins.

UV photoabsorption | protein peptide bond | machine learning | neural network

Structure determination is crucial for understanding protein activity and function (1). Their secondary and tertiary structure can be characterized by the UV absorption spectra of its backbone peptide bonds (2, 3). Interpreting these signals requires extensive electronic structure simulations, since the time-averaged optical response is affected by conformational and environmental fluctuations. The repeated application of high-level quantum-mechanical tools to represent ensembles of structures of systems of thousands of atoms is computationally prohibitive. It makes the understanding of complex systems like proteins at atomic precision a formidable task. More cost-effective approaches are called for.

The map method has long been used to estimate key excited-state parameters, avoiding expensive quantum-mechanical calculations (4–9). Empirical formulas are employed to obtain transition energies from given ground-state structures of the target molecule. Empirical fitting of peptides and proteins containing hundreds of atoms by the map method is not an easy task. *N*-methylacetamide (NMA), which is the simplest molecule that can capture the essence of UV response of the protein backbone, has been extensively used to construct spectroscopic maps and model the spectra of the amide group of the peptide backbone (6). However, this method has a limited predictive power and transferability since it is based on a few-parameter fit of observables to key structural parameters. The inability to predict transition dipoles is another limitation. Developing a cost-effective solution for predicting both excitation energies and transition dipoles of peptides is an open challenge.

Machine learning is a family of statistics-based methods that enable a computer code to make predictions without being explicitly programmed. In computational chemistry, it has the ability to predict properties of molecules, avoiding computationally demanding high-level electronic structure calculations (10). These include the band gap for inorganic compounds (11), molecular atomization energies (12), atomization and total energies of molecules (13), and intrinsic bond energies (14).

We shall employ a subclass of machine-learning algorithms, known as neural network (NN). A standard NN consists of input, hidden, and output layers connected by artificial neurons. Input signals are passed through a weighted connection and processed by an activation function to produce the neuron output (15). Unlike the map method which relies on an empirical polynomial fit with a limited set of parameters, NN can create the structure–property relationship by iterative learning based on a complex high-dimensional function in a much larger, essentially unlimited parameter space. These make it a much more adaptable, flexible, and accurate tool compared with simple maps.

In this work, we employ NN to establish a quantitative relationship between the electronic excited-state properties of NMA and its ground-state geometry and charge distribution. Based on iterative learning of quantum-chemistry calculations for 70,000 molecular-dynamics conformations, we show that NN can satisfactorily predict the $n\pi^*$ and $\pi\pi^*$ transition energies and transition dipole moments. The UV spectra of NMA at different temperatures predicted by NN are in good agreement with results from time-dependent density-functional theory (TDDFT). We demonstrate that machine learning can provide an efficient tool for simulating spectra.

Significance

UV absorption spectroscopy is an effective technique for characterizing protein structure. However its theoretical interpretation requires expensive first-principles simulations. We employ a neural network strategy to predict UV electronic spectra of peptide bonds. The protocol establishes structure–property relations and predicts ground-state dipole moments, as well as transition dipole moments. We establish machine learning as a useful spectroscopy simulation tool.

Author contributions: J.J. designed research; S.Y., W.H., J.Z., and K.Z. performed research; S.Y., W.H., X.L., G.Z., S.M., and J.J. analyzed data; and S.Y., W.H., X.L., Y.L., S.M., and J.J. wrote the paper.

Reviewers: J.D.H., University of Nottingham; and J.W., State University of New York at Stony Brook.

The authors declare no conflict of interest.

Published under the PNAS license.

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

[2]To whom correspondence may be addressed. Email: smukamel@uci.edu or jiangj1@ustc.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1821044116/-/DCSupplemental.

Published online May 30, 2019.

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

Results and Discussion

The nπ* and ππ* excitations of NMA (Fig. 1 *A* and *B*), regarded as a simplified model for the UV absorption of protein backbone, have been well studied by various electronic structure methods, including TDDFT and wavefunction-based electron correlation methods (16, 17). Compared with expensive electron correlation methods such as coupled-cluster single and double excitation equation of motion approach (EOM-CCSD) and complete active space with second-order perturbation theory (CASPT2), TDDFT is able to obtain satisfactory electronic excitation energies of various types of molecules at a modest computation cost. For NMA, Perdew–Burke–Ernzerhof hybrid functional (PBE0) functional has been employed in previous study by Gordon and coworkers (17). PBE0 has an error of about 0.3 eV in calculating transition energies of various molecules (18). Therefore, using PBE0 results as the reference data for NN training is a reasonable choice.

The distribution of NMA nπ* and ππ* transition energies shows that the snapshot structures of NMA extracted from molecular-dynamics simulation trajectories are highly diverse (*SI Appendix*, Fig. S2 *A* and *B*). This allows us to employ both map and NN method to establish structure–property relationship for the UV absorption of NMA. Specifically, the map method uses formulas obtained by least-square fitting of data to establish the relationship between the transition energies and ground-state geometric parameters. This formula was then used to predict vibrational or electronic excitations. The electronic transitions of a molecule are complex functions of its parameters. However, maps only employ simple empirical formulas (details in *SI Appendix*) that do not fully capture the complexity of electronic transitions. In contrast, the NN technique employed here does not require explicit knowledge or guess of the relationship between input and output. Instead, the structure–property relationship is established by using a high-dimensional complex mapping between input and output. The training process of the model requires identification and screening descriptors for the problem of interest, and optimizing parameters for the NN model is nontrivial. In addition, overfitting of model is a known shortcoming of NN methods that needs to be carefully taken care of in practice. In this work, we have mitigated the overfitting issue in the NN training process using well-established procedures (details in *SI Appendix*).

We first examined the conventional map method (4) for the transition energies (ω) (*SI Appendix*, Fig. S2 *C* and *D*). Different datasets yield different maps, suggesting lack of transferability (*SI Appendix*, Fig. S3 *A–D*). Predictions by the map method have large errors and poor correlation with TDDFT. The selected descriptors for NN only require readily available ground-state

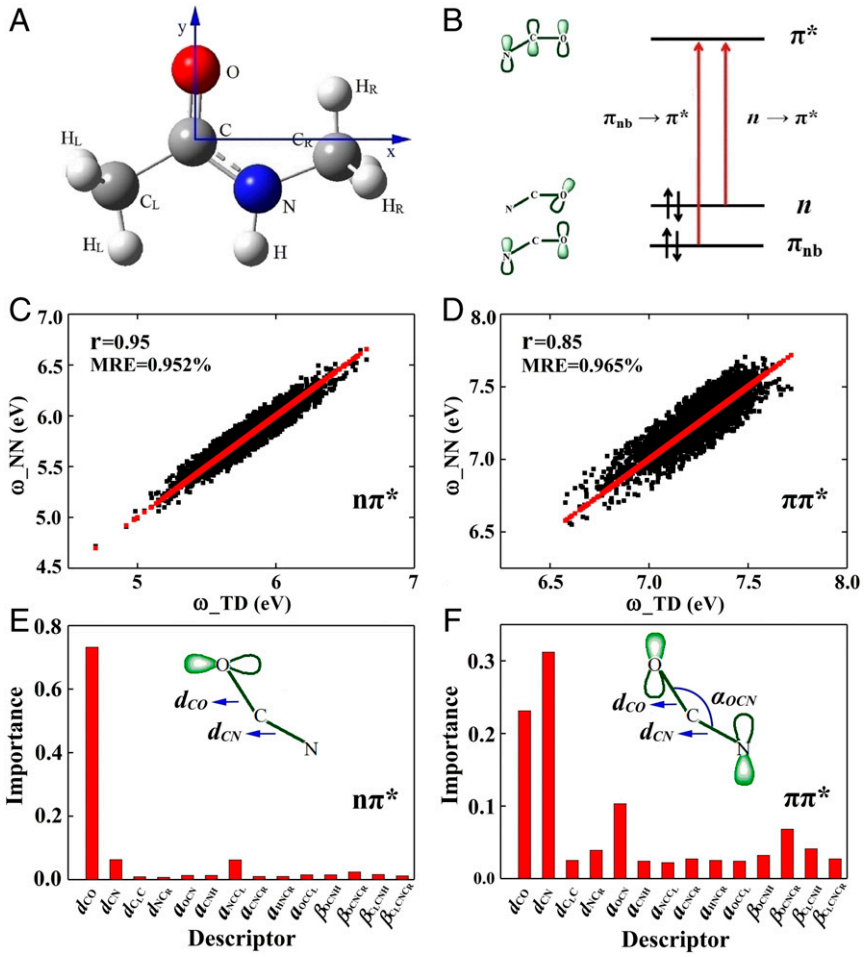


Fig. 1. The structure and electronic transitions of NMA and prediction of the NMA transition energy by NN and descriptor importance analysis by random forest. (*A*) NMA gas-phase geometry optimized at the B3LYP/6–311++G** level of DFT. (*B*) Valence molecular orbitals and electronic transitions of amides. (*C*) Comparison of TDDFT nπ* transition energies (ω_TD) and NN (ω_NN) on test data. (*D*) Same as *C* but for the ππ* transition. The red lines/dots represent the transition energies (ω_TD) of NMA calculated by TDDFT at the PBE0/cc-pvdz level. Descriptor importance analysis of transition energies for nπ* (*E*) and ππ* (*F*) transitions.

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

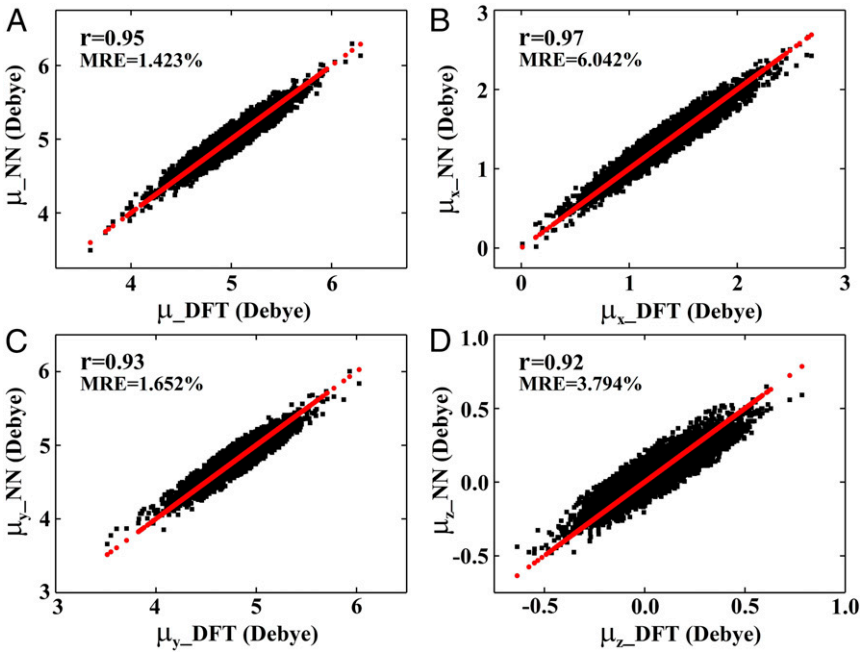


Fig. 2. Prediction of the NMA ground-state dipole moment by NN. (*A*) Correlation plots of the DFT dipole moment (μ_DFT) and NN (μ_NN) on test data. (*B–D*) Comparison of the DFT dipole moment in the *x*, *y*, *z* direction (μ_x_DFT, μ_y_DFT, μ_z_DFT) and NN (μ_x_NN, μ_y_NN, μ_z_NN) on test data. The red lines/dots represent μ_DFT, μ_x_DFT, μ_y_DFT, and μ_z_DFT of NMA calculated by DFT at the PBE0/cc-pvdz level, respectively.

information. Fourteen internal coordinates (*SI Appendix*, Fig. S1) were then used as input for NN to predict transition energies, and the produced data are then compared with TDDFT calculations (Fig. 1 *C* and *D*). The Pearson correlation coefficient (*r*) between pairs of descriptors show that most descriptors have low linear correlations (*SI Appendix*, Fig. S4), which significantly improve the performance of NN prediction. The mean relative error (*MRE*) of NN on the test set are 0.95% for nπ* and 0.96% for ππ*, and the Pearson correlation coefficient (*r*) are 0.95 for nπ* and 0.85 for ππ*, demonstrating highly accurate NN predictions with nearly 3,000× faster than traditional quantum calculations once the NN model was established (*SI Appendix*, Table S1). We note that the optimized NN model can always reproduce the result from a specific method that has been used for generating

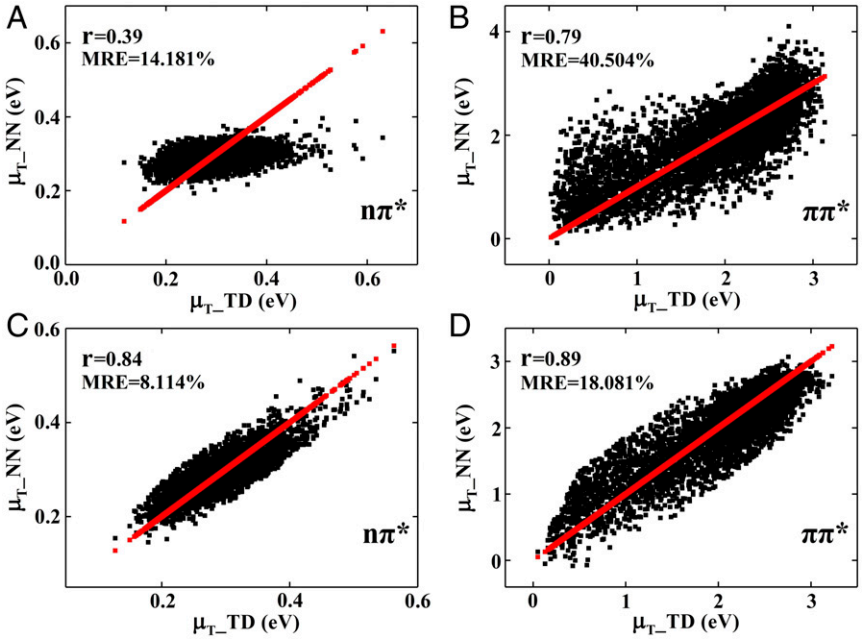


Fig. 3. NN prediction of the NMA transition dipole moment. (*A*) Correlation plots of nπ* transition dipole moment by TDDFT (μ_T_TD) and NN (μ_T_NN) using descriptors of CM on test data. (*B*) Same as *A* but for the ππ* transition. (*C*) Comparison of nπ* transition dipole moment by TDDFT (μ_T_TD) and NN (μ_T _NN) using descriptors of CM based on NPA charge (CM_Q) on test data. (*D*) Same as *C* but for the ππ* transition. The red lines/dots represent the μ_T_TD of NMA calculated by TDDFT at the PBE0/cc-pvdz level.

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

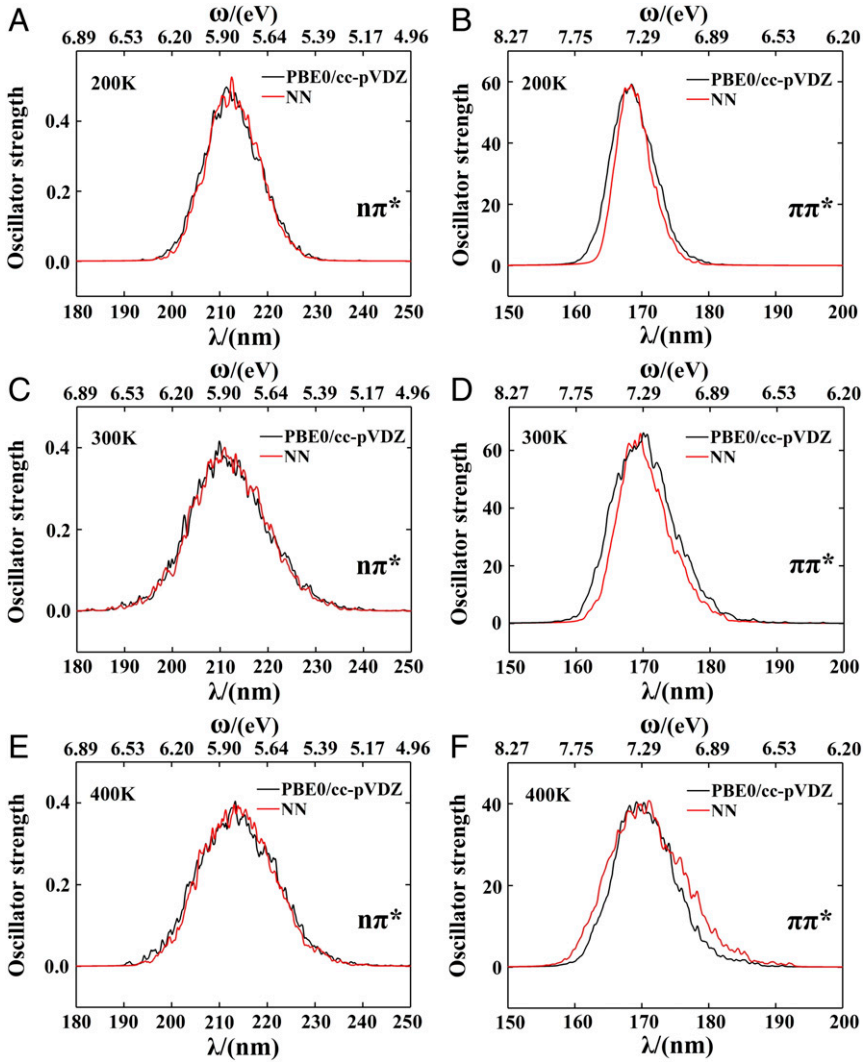


Fig. 4. (*A*) The nπ* UV spectra at 200 K of 5,000 NMA structures calculated by TDDFT and NN. (*B*) Same as *A* but for the ππ* transition. (*C* and *D*) Same as *A* and *B* but at 300 K (*E* and *F*) Same as *A* and *B* but at 400 K. TDDFT calculations are at the PBE0/cc-pvdz level.

training data (Fig. 2 *C* and *D* and *SI Appendix*, Fig. S3 *E* and *F*). Therefore, we only need to take into account the accuracy of the chosen density functional on the transition energy of NMA. And, the NN results predict transition energies better than the maps.

With the random forest algorithm for the descriptor importance analysis, we found that the C–O bond length is the dominant descriptor (Fig. 1*E*) for the nπ* excitation. This reflects the localized nature of nπ* transition. For the ππ* transition, the important descriptors are the C–N and C–O bond lengths and ∠OCN angle (Fig. 1*F*). The diverse nature of important descriptors for the ππ* transition arises from its nonlocal nature with strong dependences on the entire amid group structure. This conclusion is further verified by orbital localization analysis based on Mulliken populations (19). The localized molecular orbitals involved in the nπ* transition (*SI Appendix*, Fig. S5) clearly shows the dominant effect of the C-O bond, while the ππ* transition is determined by the entire amide group (*SI Appendix*, Fig. S5).

We have further applied the NN to predict the ground-state dipole moments. To eliminate orientational differences during NN training, we applied a rotation matrix operation by setting the carbonyl C atom of each NMA as the origin of coordinate, the C–O bond as the positive *y* axis, and the ∠OCN being fixed in the *xy* plane (Fig. 1*A*). The coordinates of five atoms (C_L, O, N, H, and C_R) were used as NN training descriptors. These adequately predict both the magnitude and direction of ground-state dipole moments (Fig. 2 and *SI Appendix*, Fig. S6). The most important descriptors for the total dipole moment are the *y* coordinate of the O atom and the *x* coordinate of the N atom (*SI Appendix*, Fig. S7). For the *x* component of dipole, the most important descriptors are the *x* coordinates of the N and C_R atom. Similarly, the *y* coordinate of the O atom has the largest influence on the *y* component of dipole. The most important descriptors of dipole moment along *z* are the *z* coordinates of the C_R and H atom. We have thus constructed the relationship between the molecular dipole moment and its structure, which allows us to rapidly predict the dipole moment (*SI Appendix*, Table S1) compared with quantum-chemistry calculations.

Then we aimed at the prediction of the transition dipole moment which governs the strength of electronic transition. However, it poses great challenge to the NN training because of the involvement of two different electronic states. We failed to get satisfactory results by using the regular coulomb matrix (CM) (10) as descriptors. We thus replaced the fixed point charge in the force field by the natural population analysis (NPA) charges which is regarded as a reliable parameter for describing charge distribution of atoms in molecules. For the nπ* transition, the peptide bond was used to construct the CM based on NPA charge.

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

The α_1-angle, dihedral angles β_2 and β_3 of NMA (*SI Appendix*, Fig. S1), the components of the y and z coordinates of the C atom [C(y), C(z)], and the y coordinate of the O atom [O(y)] were chosen as descriptors for the nπ* transition dipole moment. For the ππ* transition, the descriptors were the NPA charge-derived CM_Q of the peptide bonds. The CM_Q gave a better fit in NN training and smaller *MRE* compared with the traditional CM (Fig. 3).

Using NN-generated excitation energies and transition dipole moments, we have calculated the oscillator strength $f = \frac{2}{3}\omega|\mu|^2$, where ω and μ represent the transition energy and transition dipole moment, respectively. For each temperature considered (200, 300, 400 K), UV spectra generated using a Lorentzian lineshape with 30-meV width (details in *SI Appendix*) from 5,000 randomly selected structures show good agreements to the TDDFT results (Fig. 4). For both the average maximum of the frequency and full width at half maximum of the UV absorption spectra at different temperatures, NN results are in good agreement with TDDFT results (Tables 1 and 2).

As shown in Table 1, the transition energies of nπ* and ππ* of NMA calculated at PBE0/Dunning's correlation-consistent polarized valence double-zeta basis set (cc-pVDZ) are 5.85 eV (211.94 nm) and 7.26 eV (170.87 nm), in line with the available experiment data—5.85 eV (212.00 nm) for nπ* and 6.67 eV (186.00 nm) for ππ*, respectively (20). The nπ* transition primarily involves the highest occupied to lowest unoccupied molecular orbital transition, for which the agreement between TDDFT and experiment is excellent. The ππ* transition is known to have multireference character which can be mixed with higher excited states; therefore, the prediction of the corresponding transition energy is very challenging even for highly accurate wavefunction-based electron correlation method, like EOM-CCSD (17). Given an ∼0.3 eV error of TDDFT methods for transition energies (18), PBE0/cc-PVDZ provides a reasonably good prediction of the ππ* transition energy of NMA compared with experiment.

The NN model was trained using data at 300 K, and then applied to predict UV spectra of NMA at other temperatures (200 and 400 K). The agreement between NN-predicted and TDDFT-computed UV spectra at different temperatures suggests good transferability. These excellent agreements thus demonstrate the ability of NN to reproduce spectra, based solely on the molecular geometry and charge distribution.

Conclusions

An NN protocol was developed to represent the transition energy, the dipole moment, and the electronic spectra of NMA based solely on ground-state information (structure and charge distribution). NN predictions are more robust and accurate than the conventional map method for describing excited-state properties. It is also cheaper than repeated quantum-chemistry calculations. We have chosen the model structure of NMA as a first step to test the methodology of applying machine-learning technique to the optical response of biomolecules. Our study shows that NN is a versatile practical tool for simulating UV spectra of the NMA molecule. We are currently using NN to map the electronic properties of all 20 amino acids with their respective atomic structures, and use the acquired NN model to study the UV spectra of NMA in other solvents. It is a key step toward the machine-learning prediction of UV spectra of proteins in various solvents. The NN protocol for predicting transition energies and dipole moments in this work is being used to construct model Hamiltonian for specific proteins toward simulation of their UV spectra. The NN model for dipoles developed here may be extended to investigate optical spectroscopy of large biological complexes, and also be applied to many other important properties involving key parameters of electric and magnetic dipole moments such as chemical reactions, optoelectronic conversion, and information processing.

Table 1. Comparison of the average maximum (in nanometers) of the frequency of nπ* and ππ* absorption bands of NMA in aqueous solution at different temperatures between simulation (TDDFT, NN) and experiment

Temperature	200 K	300 K	400 K
nπ*			
TDDFT	212.74	211.94	213.83
NN	212.94	211.95	213.95
Expt	—	212.00[†]	—
ππ*			
TDDFT	168.96	170.87	170.93
NN	169.19	171.28	170.90
Expt	—	186.00[†]	—

[†]The maxima of nπ* and ππ* absorption bands of NMA in aqueous solution measured by Nielsen and Schellman (20).

Table 2. The full width at half maximum (in nanometers) of the UV absorption spectra at different temperatures of NMA by TDDFT and NN

Temperature	200 K	300 K	400 K
nπ*			
TDDFT	13.56	17.85	18.36
NN	12.15	17.19	18.87
ππ*			
TDDFT	7.47	9.32	11.34
NN	6.32	8.24	12.64

Methods

Molecular-dynamics simulations were performed using the GROMACS code (21) for isothermal-isobaric ensemble with all-atom optimized potentials for liquid simulations (22) force fields for one NMA in the solution of 875 Jorgensen's transferable intermolecular potential four point water model water molecules (4) generating 70,000 configurations at three temperatures (T = 200 K, T = 300 K, T = 400 K) (details in *SI Appendix*). The excited-state properties of these structures were calculated by first-principles TDDFT with PBE0/cc-pVDZ implemented in the Gaussian 09 package (23). Solvation effects were modeled implicitly by the integral equation formalism polarizable continuum model (24). Calculations by Becke's three-parameter hybrid functionals with the Becke exchange and the Lee-Yang-Parr correlation functional, in conjunction with Pople's split-valence double-zeta basis set with additional polarization functions [B3LYP/6-31G(d,p)] and B3LYP/6–311G ++(d,p) were performed for another set of 10,000 data points at 300 K, so as to verify if NN results are predictable well under different functional and basis sets. The nπ* transition involves the highest occupied to lowest unoccupied molecular orbital transition, while the ππ* transition has multireference character so that it can involve more than one electronic transition depending on the geometry. To unify the training dataset, we had discarded data points involving much higher excitations in the ππ* transition.

Transition energies, ground-state, and transition dipole moments were targets for the trainings and tests of the NN protocol. Fourteen internal coordinates (*SI Appendix*, Fig. S1) were used as input variables (descriptors) to predict transition energies. The *xyz* representation was used as input for the dipole moment. For the modulus of the transition dipole moment, the CM (10) based on atomic NPA charge (CM_Q) was defined as descriptors of $U_{ij} = \begin{cases} 0.5Q_i^{2.4} & \forall i = j \\ Q_iQ_j/|R_i - R_j| & \forall i \neq j \end{cases}$, where i and j are atomic indices, $|R_i - R_j|$ is the interatomic distance, and Q_i represents NPA charge.

Multilayer perceptrons (25) were used in the NN training to establish the relationship between excited-state properties and ground-state geometry and charge distributions (details in *SI Appendix*). Our protocol includes 50,000 data points at 300 K, and 10,000 at 200 and 400 K. The training set includes 40,000 data points randomly selected from a 300-K simulation. We have further used the random forest algorithm (26) to analyze the importance of each descriptor in NN. The Pearson correlation coefficient (r) is used to evaluate the NN performance, which measures the linear correlation between predicted and actual values. The *MRE* and cross-validation technique (12) were employed to verify the accuracy and robustness of the NN.

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

ACKNOWLEDGMENTS. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of University of Science and Technology of China. This work was financially supported by the Ministry of Science and Technology of the People's Republic of China (2018YFA0208603, 2017YFA0303500, and 2016YFA0400904) and the National Natural Science Foundation of China (21633006, 21473166, and 21703221). S.M. gratefully acknowledges the support of the National Science Foundation Grant CHE-1663822.

1. D. Whitford, *Proteins: Structure and Function* (John Wiley & Sons, 2013).
2. N. Berova, K. Nakanishi, R. W. Woody, R. Woody, *Circular Dichroism: Principles and Applications* (John Wiley & Sons, 2000).
3. G. D. Fasman, *Circular Dichroism and the Conformational Analysis of Biomolecules* (Springer Science & Business Media, 2013).
4. Z. Li, H. Yu, W. Zhuang, S. Mukamel, 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 (2008).
5. N. A. Besley, M. T. Oakley, A. J. Cowan, J. D. Hirst, A sequential molecular mechanics/quantum mechanics study of the electronic spectra of amides. *J. Am. Chem. Soc.* **126**, 13502–13511 (2004).
6. L. Wang, C. T. Middleton, M. T. Zanni, J. L. Skinner, Development and validation of transferable amide I vibrational frequency maps for peptides. *J. Phys. Chem. B* **115**, 3713–3724 (2011).
7. W. Zhuang, T. Hayashi, S. Mukamel, Coherent multidimensional vibrational spectroscopy of biomolecules: Concepts, simulations, and challenges. *Angew. Chem. Int. Ed. Engl.* **48**, 3750–3781 (2009).
8. J. K. Carr, A. V. Zabuga, S. Roy, T. R. Rizzo, J. L. Skinner, Assessment of amide I spectroscopic maps for a gas-phase peptide using IR-UV double-resonance spectroscopy and density functional theory calculations. *J. Chem. Phys.* **140**, 224111 (2014).
9. S. Hahn, K. Park, M. Cho, Two-dimensional vibrational spectroscopy. I. Theoretical calculation of the nonlinear Raman response function of CHCl3. *J. Chem. Phys.* **111**, 4121–4130 (1999).
10. G. Montavon *et al*., Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
11. J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
12. K. Hansen *et al*., Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
13. K. Hansen *et al*., Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
14. K. Yao, J. E. Herr, S. N. Brown, J. Parkhill, Intrinsic bond energies from a bonds-in-molecules neural network. *J. Phys. Chem. Lett.* **8**, 2689–2694 (2017).
15. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
16. N. A. Besley, J. D. Hirst, Ab initio study of the effect of solvation on the electronic spectra of formamide and N-methylacetamide. *J. Phys. Chem. A* **102**, 10791–10797 (1998).
17. N. De Silva, S. Y. Willow, M. S. Gordon, Solvent induced shifts in the UV spectrum of amides. *J. Phys. Chem. A* **117**, 11847–11855 (2013).
18. A. D. Laurent, D. Jacquemin, TD-DFT benchmarks: A review. *Int. J. Quantum Chem.* **113**, 2019–2039 (2013).
19. J. Pipek, P. G. Mezey, A fast intrinsic localization procedure applicable for abinitio and semiempirical linear combination of atomic orbital wave functions. *J. Chem. Phys.* **90**, 4916–4926 (1989).
20. E. B. Nielsen, J. A. Schellman, The absorption spectra of simple amides and peptides. *J. Phys. Chem.* **71**, 2297–2304 (1967).
21. D. Van Der Spoel *et al*., GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
22. W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118**, 11225–11236 (1996).
23. M. Frisch *et al*., *Gaussian 09 Revision D. 01, 2009* (Gaussian Inc., Wallingford, CT, 2009).
24. M. Cossi, G. Scalmani, N. Rega, V. Barone, New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *J. Chem. Phys.* **117**, 43–54 (2002).
25. F. Häse, S. Valleau, E. Pyzer-Knapp, A. Aspuru-Guzik, Machine learning exciton dynamics. *Chem. Sci. (Camb.)* **7**, 5139–5147 (2016).
26. V. Svetnik *et al*., Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).

Downloaded at SCIENCE LIBRARY SERIALS on July 1, 2020

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)**:**

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

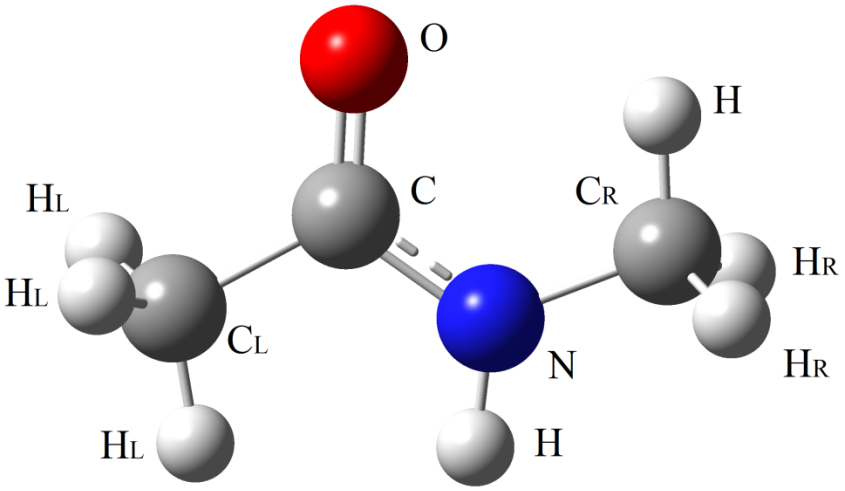
$$\mu = n\pi^*, \pi\pi^*$$

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.

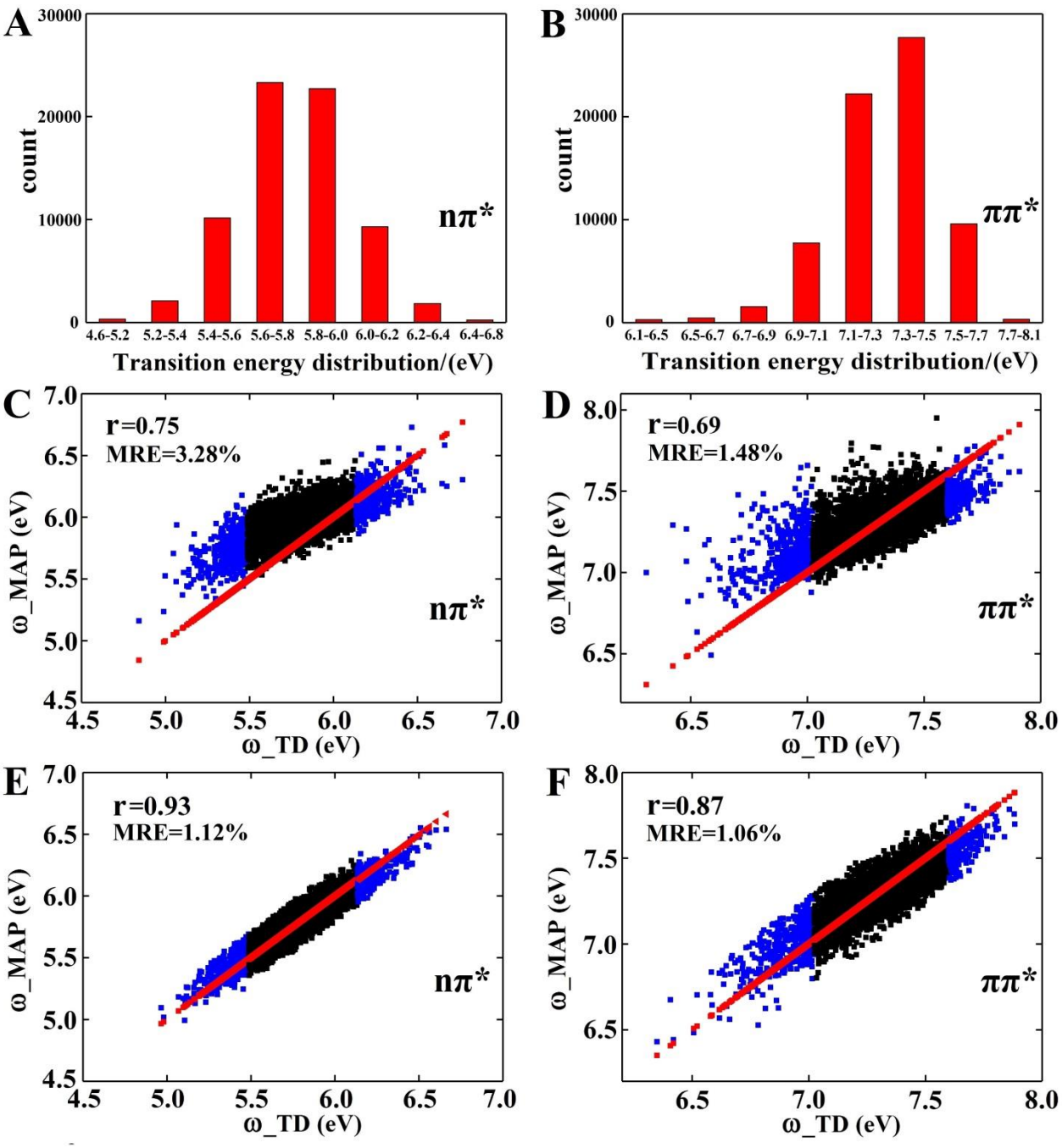


Fig. S2. (A) Distribution of NMA nπ* transition energies calculated by TDDFT. (B) Same as (A) but for the ππ* transition (C) Correlation plots of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP).(1) (D) Same as (C) but for the ππ* transition. (E) Comparison of nπ* transition energies by 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.

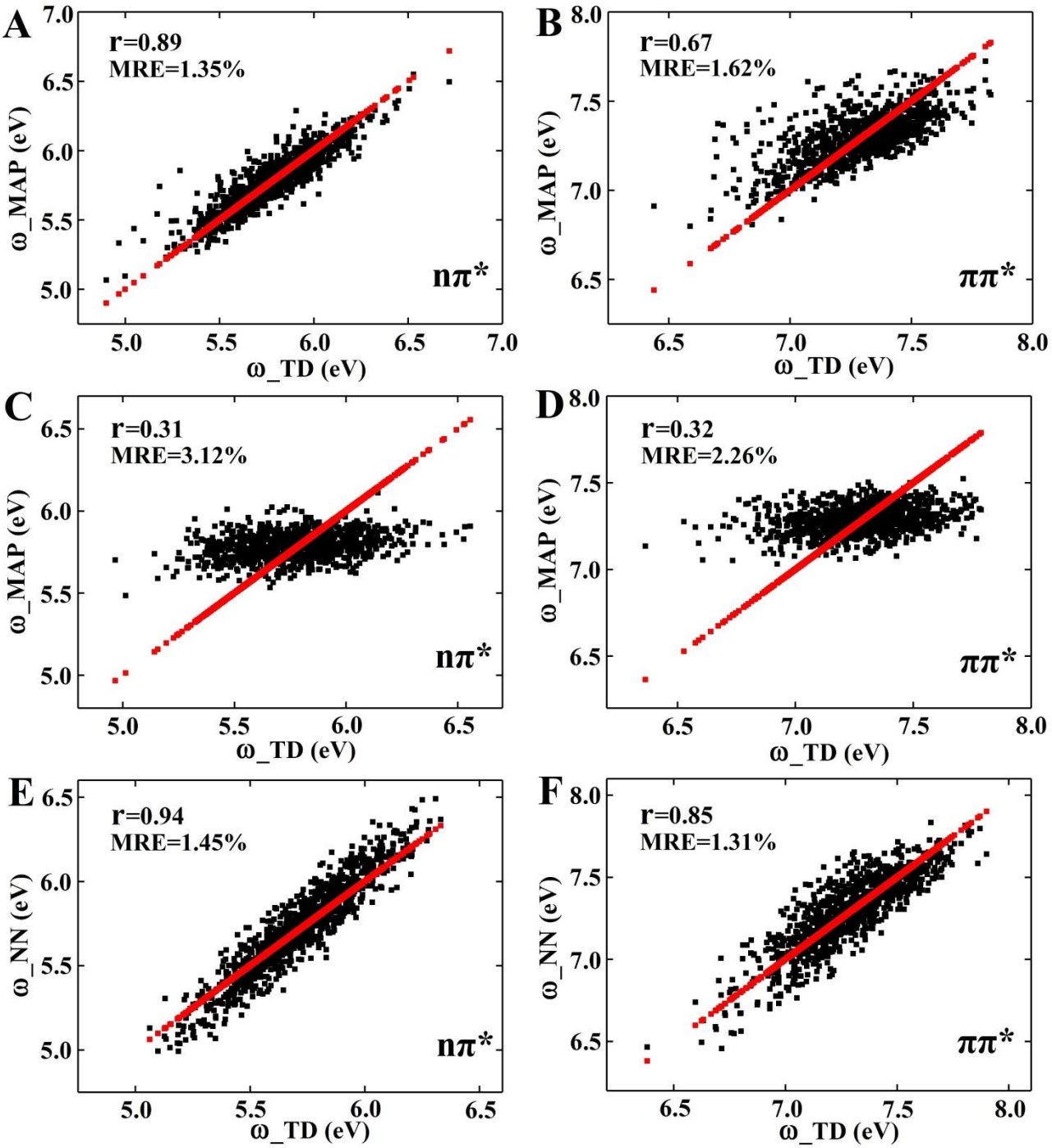


Fig. S3. (A) Correlation plots of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP).[1] (B) Same as (a) but for the ππ* transition. (C) Comparison of nπ* transition energies by TDDFT (ω_TD) and map method (ω_MAP)[1] which fitted by different data sets. (D) Same as (C) but for the ππ* transition. (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.

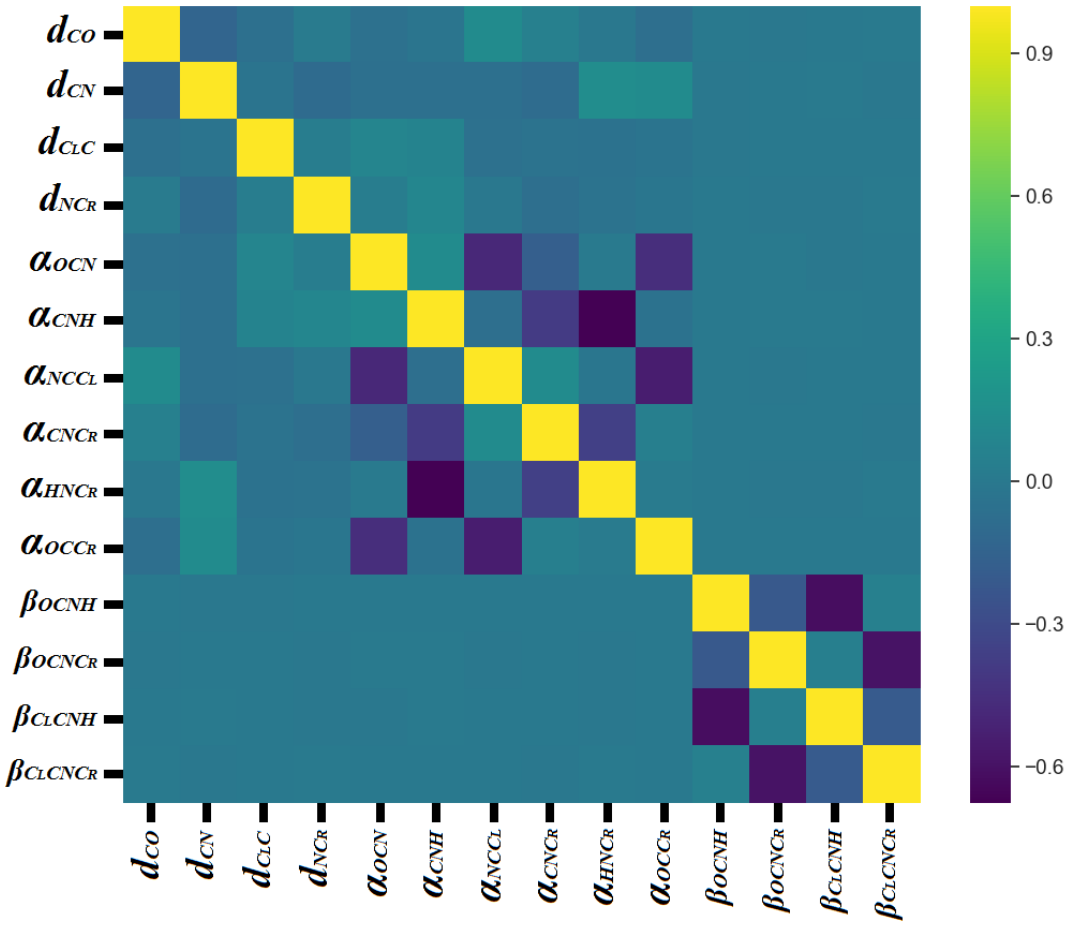


Fig. S4. Heat map of the Pearson correlation coefficient (r) among the descriptors for predicting the nπ* and ππ*transition energy.

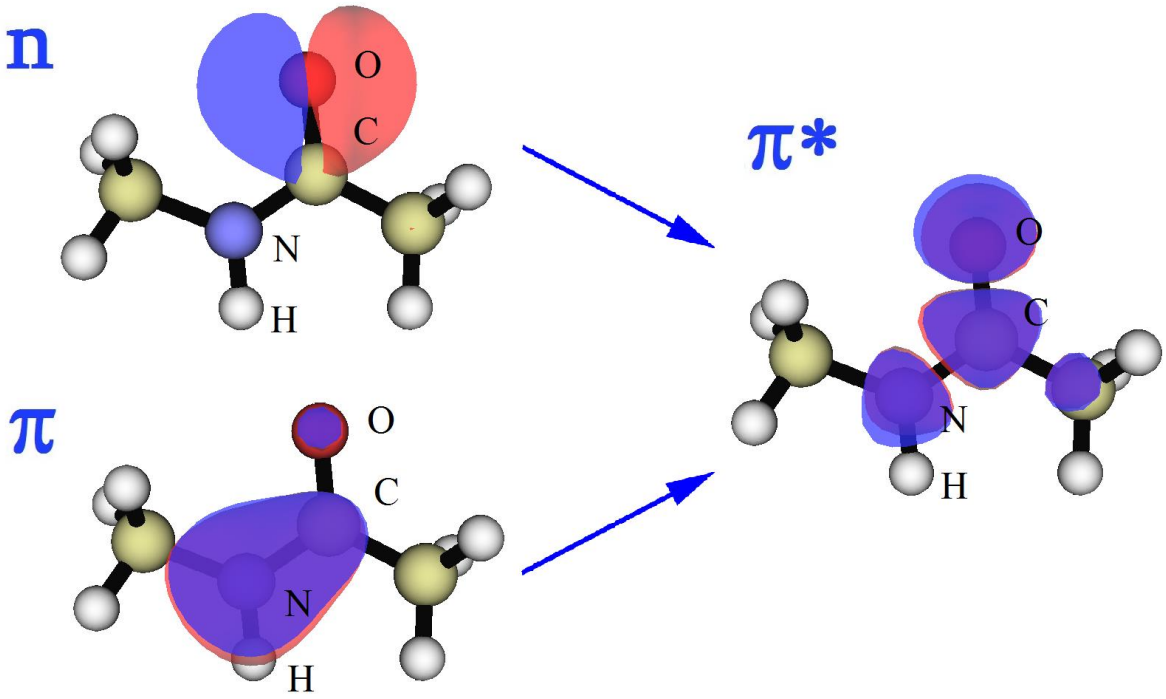


Fig. S5. The NMA molecular orbitals which after localizing analysis were included in the two transitions: nπ* and ππ* transition.

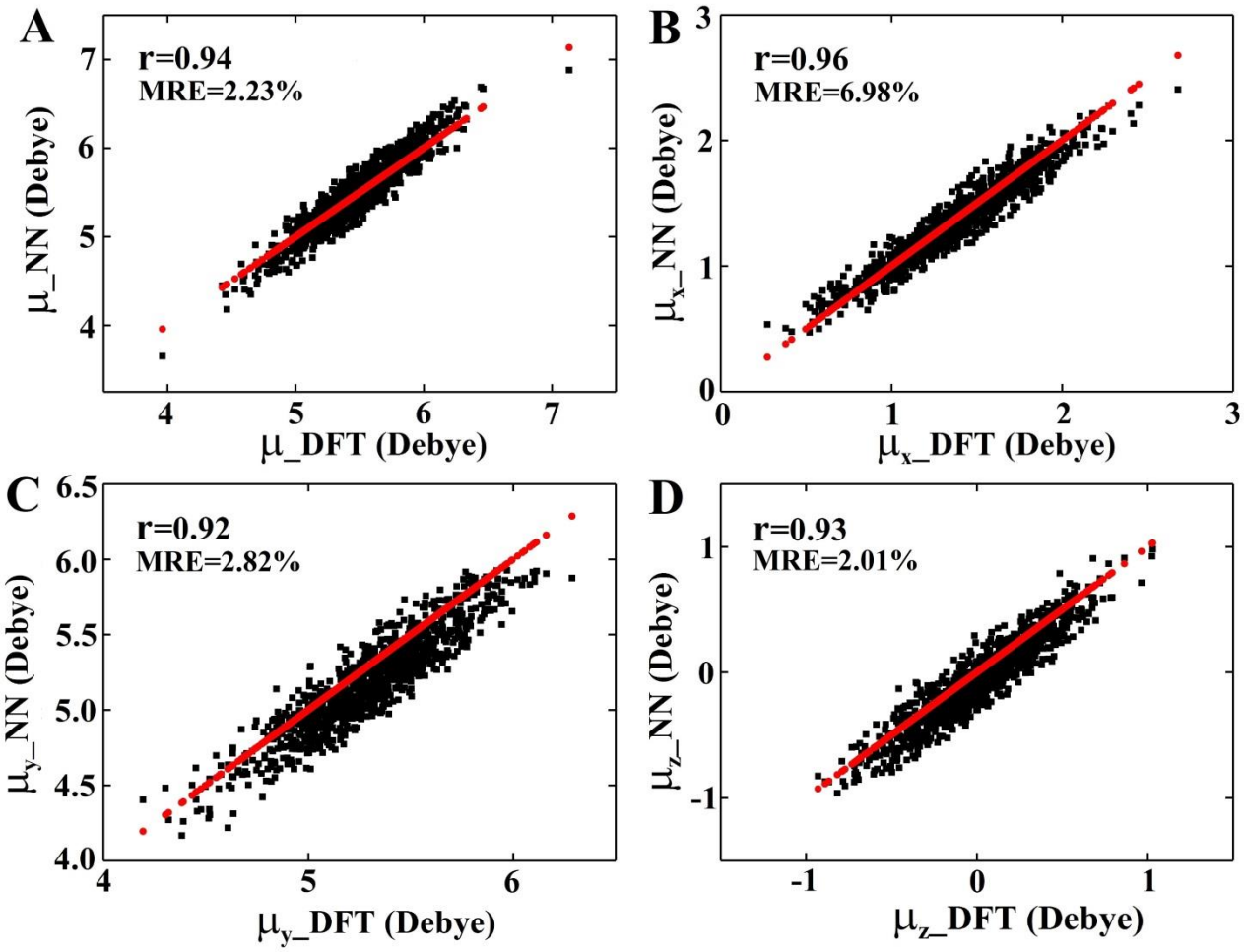


Fig. S6. (A) Correlation of dipole moment by DFT (μ_DFT) and NN (μ_NN). (B) Comparison of dipole moment in the *x* direction by DFT (μ_x_DFT) and NN (μ_x_NN). (C) Comparison of dipole moment in the *y* direction by DFT (μ_y_DFT) and NN (μ_y_NN). (D) Comparison of dipole moment in the z direction by DFT (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.

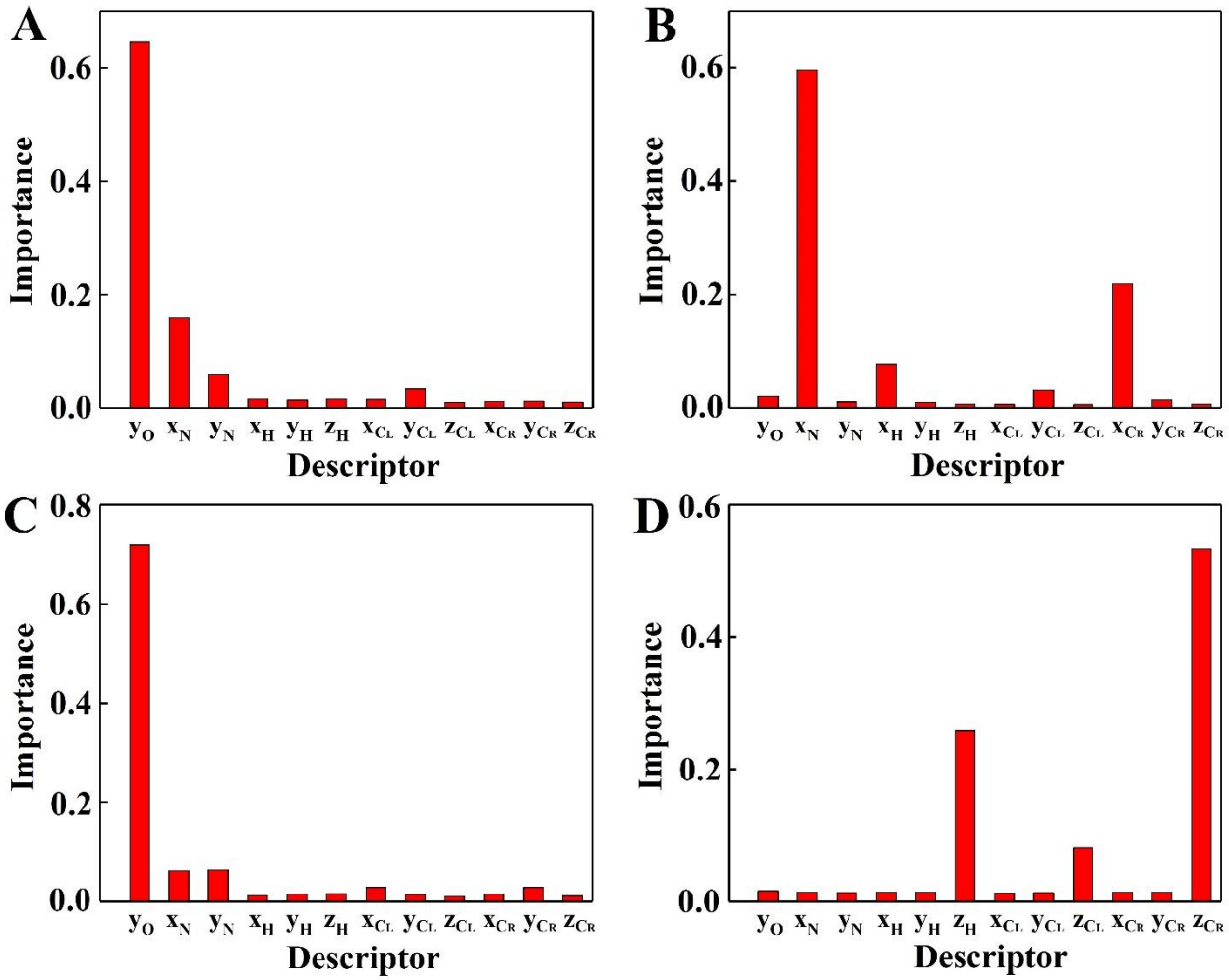


Fig. S7. (A) Descriptor importance analysis of dipole moment. (B) Descriptor importance analysis of dipole moments in *x* direction. (C) Descriptor importance analysis of dipole moments in *y* direction. (D) Descriptor importance analysis of dipole moments in *z* direction.

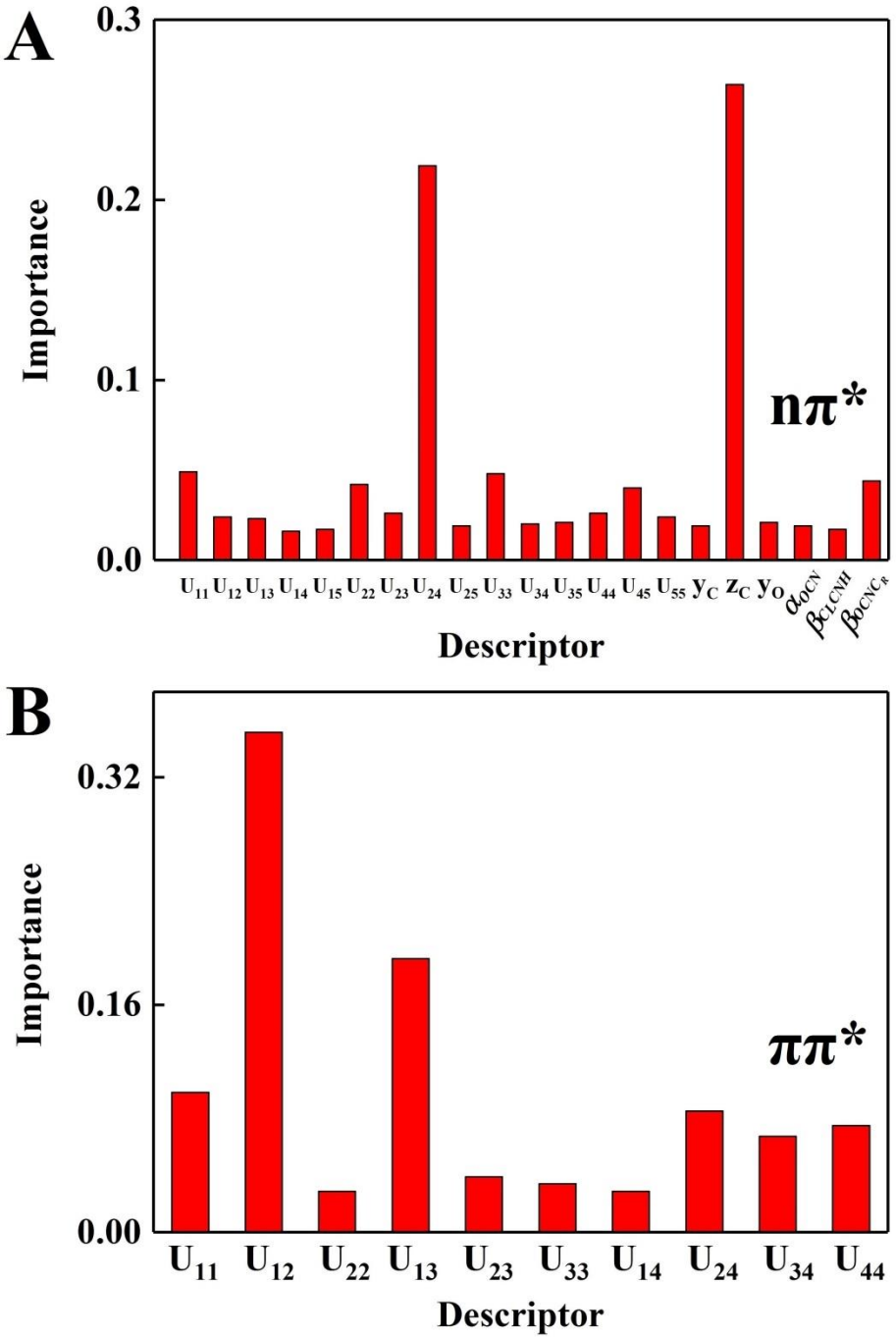


Fig. S8. (A) The importance of transition dipole moment descriptors for nπ*transition. (B) Same as (A) but for the ππ* transition.

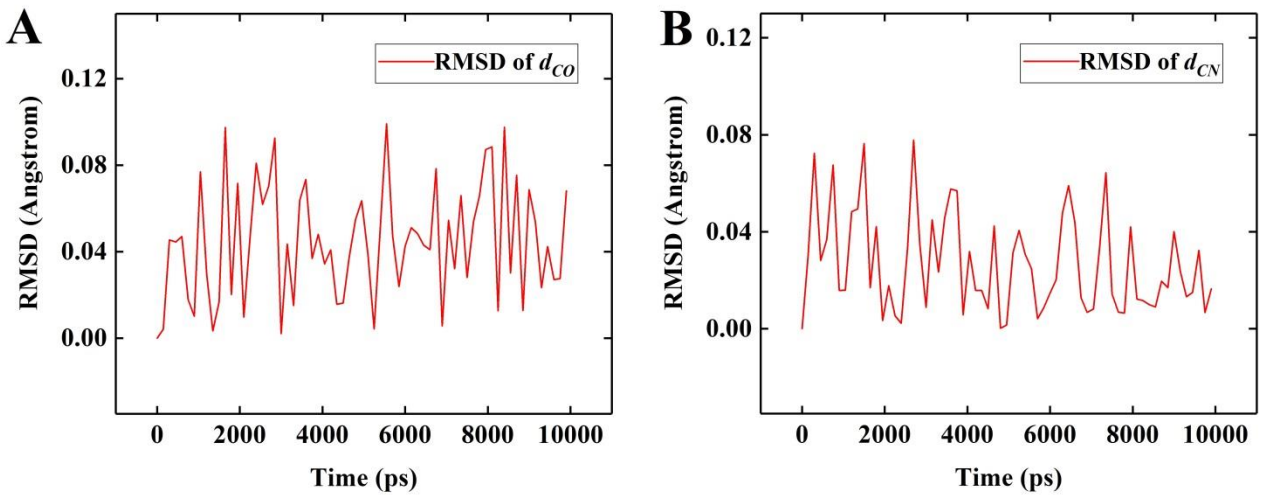


Fig. S9. (A) The root mean square deviation (RMSD) of CO bond. (B) Same as (A) but for the CN bond.

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.