A Machine Learning Approach to Calculate Electronic Couplings between Quasi-Diabatic Molecular Orbitals: The Case of DNA

*Xin Bai, Xin Guo, and Linjun Wang**

Key Laboratory of Excited-State Materials of Zhejiang Province, Department of Chemistry,

Zhejiang University, Hangzhou 310027, China

ABSTRACT: Diabatization of one-electron states in flexible molecular aggregates is a great challenge due to the presence of surface crossings between molecular orbital (MO) levels and the complex interaction between MOs of neighboring molecules. In this work, we present an efficient machine learning approach to calculate electronic couplings between quasi-diabatic MOs without the need of nonadiabatic coupling calculations. Using MOs of rigid molecules as references, the MOs that can be directly regarded to be quasi-diabatic in molecular dynamics are selected out, state tracked, and phase corrected. On the basis of this information, artificial neural networks are trained to characterize the structure-dependent onsite energies of quasi-diabatic MOs and the inter-molecular electronic couplings. A representative sequence of DNA is systematically studied as an illustration. Smooth time evolution of electronic couplings in all base pairs is obtained with quasi-diabatic MOs. Especially, our method can calculate electronic couplings between different quasi-diabatic MOs independently, and thus possesses unique advantages in many applications.

Ab initio electronic structure calculation naturally provides the adiabatic representation of electronic states, which correspond to eigenstates of the electronic Hamiltonian. Because the nonadiabatic couplings change rapidly near state degeneracy in surface crossings,[1,2] it is more appealing to express each adiabatic state as a superposition of diabatic states, whose derivative couplings can be neglected. In such a diabatic representation, the Hamiltonian elements are expected to vary smoothly no matter how the structure rearranges.[3] This property is beneficial especially for nonadiabatic dynamics simulations of charge and exciton dynamics processes.[4,5] By definition, the most straightforward way is to minimize the derivative couplings through unitary transformation of the electronic states.[6] However, this is generally time consuming due to the large amount of nonadiabatic coupling vectors to be calculated, and it has been proved that strict diabatization can only be obtained for electronic states of the same symmetry in diatomic molecules.[7] Thereby, extensive efforts have been devoted to constructing quasi-diabatic representations in the past decades.[8-14]

In conjugated molecular aggregates, molecular orbitals (MOs) of individual molecules are widely adopted as quasi-diabatic states to expand the electronic wavefunction.[15-18] The diabatic couplings between these MOs are usually called electronic couplings or transfer integrals,[19-24] which are key parameters in charge transfer and transport simulations. For instance, the electronic coupling can be calculated based on adiabatic quantities (i.e., vertical excitation energy, transition dipole moment, and change in dipole moment) as in the Mulliken-Hush method.[19] Electronic polarization has been shown to have strong impacts on the site energies, and thus should be taken into account in electronic coupling calculations.[21] In general, one MO per molecule is considered, and the electronic coupling in a molecular pair is assumed to be independent of other molecules. Thereby, the site energies of MOs and the electronic couplings can be combined to construct the one-electron Hamiltonian of the whole system.[25-27]

For flexible molecules, there exist complicated crossings between potential energy surfaces (PESs) of MOs, and thus MOs are not always quasi-diabatic when characterizing the complex intermolecular electronic couplings. To the best of our knowledge, these surface crossings between MO levels have been generally neglected in charge transport simulations. In principle, one needs to consider more MOs per molecule, and thus the difficulty of diabatization is significantly enhanced. With the rapid development of machine learning, developing more efficient diabatization methods using artificial neural networks (ANNs) has attracted growing interest.[28-38] It is demonstrated that ANNs can be used to analytically express the diabatic Hamiltonian elements using *ab initio* data (e.g., adiabatic energies, energy gradients, and nonadiabatic couplings) as the training set.[31] If selected diabatic Hamiltonian elements at some geometries are already known and used as constraints, the ANNs can be also trained with a high accuracy on the basis of only adiabatic energies.[34,35]

Here, we propose an efficient machine learning approach to calculate electronic couplings with the consideration of surface crossings between MO levels. When a large number of MOs can be directly regarded as quasi-diabatic MOs, their information is used to train ANNs for the structure-dependent onsite energies and electronic couplings. Due to the importance of charge transport in DNA, electronic couplings between nucleobases have been extensively studied in the literature.[38-48] In a representative DNA sequence with all types of nucleobases and base pairs, we obtain a reliable description of electronic couplings in all base pairs.

Without loss of generality, we consider a pair of molecules (i.e., A and B) in a certain environment. When a proper force field is constructed, a trajectory of the molecular pair is obtained through molecular dynamics (MD) simulation and a sequence of snapshots are chosen to ensure enough sampling of the geometries. Isolated molecules of A and B are optimized and utilized to best fit each of the corresponding molecular geometries in the MD snapshots with translational and rotational operations, generating an auxiliary series of rigid molecules. In our approach, MOs of these rigid molecules (RMOs) are chosen as references to diabatize MOs of the original molecules (OMOs) and calculate the corresponding electronic couplings between the two molecules. To this end, *ab initio* calculations are carried out to obtain the MOs of the molecular pairs in the MD snapshots and the individual molecules in both the MD and auxiliary sequences of geometries.

Apparently, the same type of rigid molecules in the auxiliary geometries possess identical internal structures but different orientations, and thus their RMOs with the same indexes have exactly the same shapes and energies. The wavefunctions of RMOs, however, still suffer from phase uncertainties associated with the matrix diagonalization in *ab initio* calculations. As shown in Fig. 1A, we may start from RMOs in the first snapshot and make phase correction based on the overlaps between the corresponding RMOs in adjacent snapshots. In detail, a phase factor for the *i*-th RMO in the *m*-th snapshot is calculated as

$$f_{r,i}^{m} = \prod_{j=1}^{m-1} \frac{S_{r,i}^{j,j+1}}{\left| S_{r,i}^{j,j+1} \right|}, \tag{1}$$

where the subscript *r* represents the rigid geometries, and $S_{r,i}^{j,j+1} = \left\langle \phi_{r,i}^{j} \middle| \phi_{r,i}^{j+1} \right\rangle$ is the overlap integral between the *i*-th RMO in snapshot *j*, $\left| \phi_{r,i}^{j} \right\rangle$, and that in snapshot *j*+1, $\left| \phi_{r,i}^{j+1} \right\rangle$. The RMO wavefunction is then corrected as

$$\left| \phi_{r,i}^{\prime m} \right\rangle = f_{r,i}^{m} \left| \phi_{r,i}^{m} \right\rangle, \tag{2}$$

so that the RMO phases of all rigid geometries are kept consistent with those in the first snapshot. Since atomic orbitals (AOs) are widely utilized as basis functions in quantum chemistry codes for molecular systems, the RMO wavefunctions are normally real-valued and all of the phase factors are limited to values of +1 or −1. Note that Akimov has proposed a general phase correction approach to deal with complex-valued MO wavefunctions.[49] When a large time interval is utilized to generate the snapshots, however, the molecular orientations in adjacent snapshots may differ significantly with each other and the sign of $S_{r,i}^{j,j+1}$ may be unreliable. To solve this problem, a critical value, S_{c1}, is introduced as an indicator of smooth change of the RMOs. If $\left| S_{r,i}^{j,j+1} \right| \le S_{c1}$, we interpolate additional rigid geometries to the auxiliary sequence of geometries and perform *ab initio* calculations to get their RMOs until $\left| S_{r,i}^{j,j+1} \right|$ between adjacent geometries are all greater than S_{c1}. Then, the phases of RMOs are similarly corrected by Eqs. (1) and (2). In this process, S_{c1} should be large enough to achieve smooth change of RMOs, but cannot be too large in order to avoid generating too many interpolated geometries. More details are provided in the Supporting Information (SI).

For flexible molecules, PESs of the OMOs may easily cross with each other, and thus OMOs with the same indexes do not always match at different geometries. As shown in Fig. 1A, we use phase-corrected RMOs as references and make state tracking of the OMOs according to their similarities to the corresponding RMOs. For both A and B molecules in any snapshot *m*, we calculate the maximum overlap between RMO *i* and all the OMOs under consideration,

$$S_{i,\max}^{m} = \max_{j}\left\{\left|S_{ij}^{m}\right|\right\}, \tag{3}$$

where $S_{ij}^{m} = \left\langle \phi_{r,i}'^{m} \middle| \phi_{o,j}^{m} \right\rangle$ is the overlap integral between RMO *i* and OMO *j* in the *m*-th snapshot. Suppose the index of OMO with the maximum overlap is a_i^m. If $S_{i,\max}^{m}$ is close to 1, RMO *i* dominates the a_i^m-th OMO, meaning that this OMO can be naturally regarded as the *i*-th quasi-diabatic MO. In practical calculations, we introduce another parameter, S_{c2}. If $S_{i,\max}^{m} > S_{c2}$, the a_i^m-th OMO is considered as a quasi-diabatic MO, and the phase factor is

$$f_{o,a_i^m}^{m} = \frac{S_{i,a_i^m}^{m}}{\left|S_{i,a_i^m}^{m}\right|} f_{r,i}^{m}. \tag{4}$$

Otherwise, the a_i^m-th OMO is supposed to be a superposition of multiple quasi-diabatic MOs. Here, the subscript *o* represents the original geometries, and $\left|\phi_{o,a_i^m}^{m}\right\rangle$ is corrected to

$$\left|\phi_{o,a_i^m}'^{m}\right\rangle = f_{o,a_i^m}^{m}\left|\phi_{o,a_i^m}^{m}\right\rangle. \tag{5}$$

Note that traditional phase correction is usually performed by directly examining the overlap between adiabatic states in adjacent snapshots and relies on the reaction path through surface crossings.[50] In comparison, RMOs are utilized as references in our method, and thus the results are completely path independent. As a result, the MOs of molecules with similar structures are always phase consistent, which is key for machine learning of electronic couplings.

For molecule A or B, if the a_i^m-th OMO is regarded as the *i*-th quasi-diabatic MO in the *m*-th snapshot, its energy, $E_{a_i^m}^{m}$, is recorded as

$$E_i = E_{a_i^m}^{m}, \tag{6}$$

which gives the onsite energy of the *i*-th quasi-diabatic MO. As shown in Fig. 1B, if both the a_i^m -th OMO of molecule A ($\mathrm{OMO}_{a_i^m}^{\mathrm{A}}$) and the a_j^m -th OMO of molecule B ($\mathrm{OMO}_{a_j^m}^{\mathrm{B}}$) are regarded as quasi-diabatic MOs, the electronic coupling between these two OMOs in the molecular pair, V_{ij} , can be calculated with the method proposed by Valeev and co-workers,[21]

$$V_{ij} = \frac{J_{ij} - \frac{1}{2}(e_i + e_j)\mathrm{S}_{ij}}{1 - S_{ij}^2}. \tag{7}$$

Here, $S_{ij} = \left\langle \phi'^{m}_{o,a_i^m} \middle| \phi'^{m}_{o,a_j^m} \right\rangle$, $J_{ij} = \left\langle \phi'^{m}_{o,a_i^m} \middle| \hat{H} \middle| \phi'^{m}_{o,a_j^m} \right\rangle$, and $e_i = \left\langle \phi'^{m}_{o,a_i^m} \middle| \hat{H} \middle| \phi'^{m}_{o,a_i^m} \right\rangle$, where $\left| \phi'^{m}_{o,a_i^m} \right\rangle$ and $\left| \phi'^{m}_{o,a_j^m} \right\rangle$ are phase-corrected OMOs, and $\hat{H}$ is the one-electron Hamiltonian of the molecular pair. Note that other methods to calculate the electronic coupling can be also used here.[19,20,22-24] These onsite energies and electronic couplings are adopted as training and validation sets for further machine learning, and the obtained ANNs give analytical expressions of their structure dependence. In general, an ANN is composed of an input layer, multiple hidden layers, and an output layer (see Fig. S1 in the SI). If interatomic distances are utilized as descriptors of the molecular structure, the number of descriptors equals to the combinatorial number $N(N\text{-}1)/2$ for a system of N atoms. For any input feature x, we first calculate its average value and standard deviation as μ and σ, and standardize the input feature as $x' = (x\text{-}\mu)/\sigma$. The number of hidden layers, neurons in each hidden layer, and activation functions are tuned to achieve the best reliability. The output layer has only one neuron and gives the predicted onsite energy of a quasi-diabatic MO or the intermolecular electronic coupling between quasi-diabatic MOs.

As shown in Fig. 2A, we study a representative double-strand B-form DNA embedded in water solution with K^+ and Cl^- ions following the well-established protocol proposed by the Ascona B-DNA Consortium (ABC).[51] MD simulations are carried out by the AMBER suite using the BSC1 force field.[52] The sequence, 5'-GCAATCCCACGTTAGCTGAGC-3', contains all kinds of inter-strand and intra-strand base pairs (see Fig. 2B). Here, A, C, T, G represent adenine, cytosine, thymine, and guanine, respectively. Oxidative damage to DNA generates holes,[53,54] and simulation of hole transport in DNA usually considers only the highest occupied molecular orbital (HOMO) for each nucleobase in the literature.[38-48] Thereby, we here focus on the electronic couplings between quasi-diabatic HOMOs of neighboring nucleobases. To characterize surface crossings with HOMO levels, we also consider HOMO-1 states.

MD simulation of the DNA sequence is carried out at 300 K for a total time of 50 ns after equilibrium, and a sequence of snapshots of the MD trajectory are picked out. To ensure enough sampling, we choose 30000 snapshots with a time interval of $\Delta t = 2$ fs for the first 60 ps, 9940 snapshots with $\Delta t = 1$ ps between 60 ps and 10 ns, and 20000 snapshots with $\Delta t = 2$ ps for the last 40 ns, resulting in a total number of 59940 snapshots. For each nucleobase, the atomic coordinates are extracted from the DNA strand and the missing hydrogens on the corresponding nitrogen atoms are added. They are then fitted by their optimized geometries to obtain an auxiliary sequence of rigid geometries. *Ab initio* calculations are carried out at the density functional theory (DFT) level by the Gaussian package[55] for individual nucleobases in both the MD and auxiliary sequences of geometries and base pairs only in MD snapshots. The PW91 functional is adopted with the 6-31+G* basis set. More computational details are provided in the SI.

As shown in Figs. 2C and 2D, HOMO and HOMO-1 are significantly different for the same nucleobase in the optimized geometry. Using them as references, we may easily track the characteristics of OMOs. First, phase correction is carried out for HOMO and HOMO-1 of the rigid nucleobases in the auxiliary geometries. As large time intervals have been used to generate the snapshots, we check the role of S_{c1} and find that S_{c1} = 0.80 is generally large enough to give robust results with low computational costs (see the SI). Note that the rigid geometries are arranged according to time here, but they can be also reordered based on geometry similarities to reduce the number of interpolated geometries. State tracking and phase correction are then performed for the OMOs. Because the nonadiabatic coupling between two MOs in a molecule is inverse proportional to their energy difference,[56] we only consider HOMO ~ HOMO-5 to calculate the overlaps with reference HOMO and HOMO-1.

The mean squared error (MSE) is used as the loss function in machine learning, and the number of neurons in each hidden layer i is set as $u_i\sqrt{n_{input}}$, where n_{input} is the dimension of input and u_i is a scale factor. To improve the sampling of the configuration space, the MO phases for the same kind of nucleobases in the first snapshot of the MD trajectory are kept consistent, and the data of the same kind of nucleobases or base pairs in the selected 59940 snapshots is merged together. With a specified S_{c2}, all the data for each nucleobase or base pair is divided into the learning set with geometries whose MOs are considered to be naturally quasi-diabatized and the prediction set with the remaining geometries. The learning set is shuffled and further divided into training and validation sets with a ratio of 8:2.

For each nucleobase, the proportion of learning data in all the data for HOMO and HOMO-1 energies are shown in Fig. S2A and S2B, respectively. In all cases, this proportion drops with S_{c2} due to the presence of surface crossings. Thereby, crossing is the most significant in T, while G suffers the lightest among the four nucleobases. To find out the balance between data quality and diversity, the training sets with different S_{c2} values (i.e., 0.75, 0.80, 0.85, and 0.90) are tested and ANN models with different numbers of hidden layers and neurons are evaluated. In Fig. S2C, we show the mean absolute errors (MAEs) obtained by 12 different ANN models to predict HOMO energies. To ensure better generalization ability, the validation MAE needs to be small, the MAEs of training and validation sets should be close, and the size of ANN should be as small as possible. According to these rules, we find that the ANN model (6) with 5 hidden layers (i.e., $u_1 = 5$, $u_2 = 6$, $u_3 = 7$, $u_4 = 6$, and $u_5 = 4$) using the dataset with $S_{c2} = 0.90$ is optimal. Thereby, we use this model to train the onsite energies for both quasi-diabatic HOMO and HOMO-1 (see Figs. S4 and S5). As shown in Fig. 3A, the MAEs of HOMO and HOMO-1 for all the nucleobases are smaller than 2 meV. Divided by the corresponding average values, the relative errors are all lower than 0.03% (see Fig. S2D), indicating that the obtained ANN models are highly reliable.

In Fig. 4A, we use nucleobase T as an example and show representative PESs of HOMO and HOMO-1 in both quasi-diabatic and adiabatic representations. The quasi-diabatic MOs switch between the adiabatic HOMO and HOMO-1 frequently. The trained ANNs predict reasonable energies of quasi-diabatic MOs in surface crossing regions and give smoother PESs. As shown in Fig. 3B, the adiabatic HOMOs are contributed only by the quasi-diabatic HOMO and HOMO-1 in the learning sets, and thus HOMO and HOMO-1 of all nucleobases could be regarded as two-state systems. As shown in the SI, they should satisfy

$$V_{11} + V_{22} = E_1 + E_2 \,, \tag{8}$$

where V_{11} (V_{22}) and E_1 (E_2) are energies of the first (second) quasi-diabatic and adiabatic states, respectively. This equation can be used to examine the performance of trained ANNs in the prediction sets. To this end, we calculate the average energy difference by

$$\Delta E = \frac{1}{N_{pred}} \sum_{i=1}^{N_{pred}} \left| (E_{i,HOMO}^{adia} + E_{i,HOMO-1}^{adia}) - (E_{i,HOMO}^{dia} + E_{i,HOMO-1}^{dia}) \right|, \tag{9}$$

where $E_{i,HOMO}^{adia}$ and $E_{i,HOMO-1}^{adia}$ are energies of adiabatic HOMO and HOMO-1, $E_{i,HOMO}^{dia}$ and $E_{i,HOMO-1}^{dia}$ are energies of quasi-diabatic HOMO and HOMO-1, and N_{pred} is the number of geometries in the prediction set. The relative error is then calculated as

$$\delta = \frac{\Delta E}{\frac{1}{N_{pred}} \sum_{i=1}^{N_{pred}} (E_{i,HOMO}^{adia} - E_{i,HOMO-1}^{adia})}. \tag{10}$$

As shown in Fig. 3C, δ is lower than 4% for all nucleobases, replying that the independent learning of quasi-diabatic HOMO and HOMO-1 energies catches the essential physics.

Because of the high performance in MO energy predictions, the data set with $S_{c2} = 0.90$ is also used to train ANN models for electronic couplings between quasi-diabatic MOs. ANN models with different numbers of hidden layers and input dimensions are tested on four representative base pairs (i.e., AA, AC, AG, and AT) (see Fig. S3A). As the number of pairwise distances between atoms are large for base pairs, the principal component analysis is adopted to diminish the redundant information and to reduce the size of ANNs (see the SI). Different input dimensions are evaluated, including 100, 200, 3*N*-6 (i.e., the number of vibrational modes for a system with *N* atoms), and *N*(*N*-1)/2 (i.e., the number of all pairwise distances). We find that the ANN model (9) with 5 hidden layers (i.e., $u_1 = 2$, $u_2 = 5$, $u_3 = 7$, $u_4 = 4$, and $u_5 = 2$) and 100 input dimensions can achieve the best balance between minimizing MAE and avoiding overfitting. Thereby, this model is used to predict the electronic couplings of all base pairs (see Figs. S6-S9). As shown in Fig. 3D, MAEs are smaller than 4 meV for all base pairs, corresponding to relative errors all smaller than 10% (see Fig. S3C). In Fig. 4B, the time-dependent electronic couplings are shown using the base pair GT as an example. Different with the quasi-diabatic PESs of nucleobase T shown in Fig. 4A that cross frequently, those of nucleobase G have almost no crossings between HOMO and HOMO-1 PESs (see inset of Fig. 4B). As a result, the time-dependent electronic couplings between the adiabatic HOMO of G and the adiabatic HOMO (HOMO-1) of T are discontinuous in certain regions due to the strong surface crossings in T. Encouragingly, the electronic couplings between quasi-diabatic HOMOs given by our ANNs are still smooth.

In Fig. 5, we show the time-dependent onsite energies of quasi-diabatic HOMO for nucleobase T and electronic couplings between quasi-diabatic HOMOs for base pair GT in the MD trajectory with the trained ANNs. In comparison with onsite energies, the electronic couplings exhibit obvious long-period characteristics, corresponding to conformational transition.[51] The electronic coupling fluctuates in a large range between -0.1 eV and 0.2 eV, implying that strong nonlocal electron-phonon coupling is present. It is important to emphasize that because robust phase correction has been considered, the frequent sign change of electronic coupling is intrinsic and should be explicitly considered for DNA studies. Similar behaviors can be found in other nucleobases and base pairs (see Figs. S10-S17 in the SI).

Fig. 6A shows typical results reported in the literature concerning the absolute values of electronic couplings between HOMOs for different base pairs in DNA. In many studies, the nucleobases are simply placed in the regular DNA structure with a distance of 3.38 Å and a twist angle of 36° between neighboring nucleobases.[39-41,44] Although different *ab initio* methods and basis sets have been used, similar overall trends are obtained. The five largest electronic couplings generally correspond to AT, CT, GC, GT, and TT base pairs. Especially, Troisi and co-workers obtained average electronic couplings based on the most representative geometries extracted from 68 crystallographic structures.[42] The additional structural sampling changes the trend evidently and the five largest electronic couplings now become AC, CA, GA, TA, and TT. This result agrees with the strong fluctuation of electronic couplings in Fig. 5B. Thereby, more extensive sampling is required to characterize the average strength of electronic couplings.

In Fig. 6B, we make a systematic study of the average electronic couplings with our method. Considering a uniform sampling of the geometries with Δt = 2 ps, we calculate the mean absolute electronic coupling

$$|\tilde{V}| = \frac{\sum_{m=1}^{N_{snap}} |V_m|}{N_{snap}}, \tag{11}$$

and the absolute mean electronic coupling

$$|\bar{V}| = \frac{\left|\sum_{m=1}^{N_{snap}} V_m\right|}{N_{snap}}, \tag{12}$$

for each base pair, where V_m is the electronic coupling in the m-th snapshot and N_{snap} is the total number of snapshots. By definition, $|\tilde{V}|$ and $|\bar{V}|$ are identical only when all of the electronic couplings possess the same signs, and thus the difference between $|\tilde{V}|$ and $|\bar{V}|$ indicates the strength of sign change during the MD dynamics. For base pairs like CG and TC,

$|\tilde{V}|$ is much larger than $|\bar{V}|$, and $|\bar{V}|$ is close to zero, implying that positive and negative electronic couplings have similar amounts. For inter-strand base pairs (i.e., iAT and iCG), both $|\tilde{V}|$ and $|\bar{V}|$ are less than 15 meV, which are much smaller than those in most intra-stand base pairs. Compared with ref. 42, the difference between our $|\tilde{V}|$ for different base pairs is much smaller due to two major reasons: sampling in the configuration space is much more sufficient in the present work, and the electronic couplings are calculated between quasi-diabatic HOMOs instead of the traditional adiabatic HOMOs. It is important to note that the mean electronic couplings are almost the same for all base pairs with different choices of S_{c2}. Thereby, the obtained electronic couplings are robust and do not rely on very large amount of data for training of the ANNs.

In Fig. 6B, we also show $|\tilde{V}|$ based on adiabatic HOMOs. It is clear that $|\tilde{V}|$ are close between adiabatic and quasi-diabatic HOMOs only for AA, AG, GA, and GG base pairs. Namely, the major differences come from nucleobases C and T. For C, the energy gap between HOMO and HOMO-1 in the rigid geometry is only 0.04 eV (see Fig. 3E). Thereby, the equilibrium geometry is close to the crossing point between HOMO and HOMO-1 PESs, and these two states are easy to turn over during the MD dynamics. Surprisingly, over 99% of the adiabatic HOMOs of C actually correspond to HOMO-1 in the rigid geometry (see Fig. 3B). For T, the average energy difference between adiabatic HOMO and HOMO-1 is only 0.13 eV (see Fig. 3F), and only about 30% of the adiabatic HOMOs correspond to the quasi-diabatic HOMO (see Fig. 3B) due to the frequent surface crossings. Therefore, C needs to be carefully state tracked, while both state tracking and surface crossing are important for T. For G and A, although only very limited number of adiabatic HOMOs are involved in surface crossings (see Fig. 3B) due to the large energy difference between HOMO and HOMO-1 (see Figs. 3E and 3F), the phase correction based on overlaps between adjacent MOs is also invalid in the long-time limit. Namely, the electronic couplings still suffer from severe sign problems even there exists a small chance of surface crossing. As a result, proper treatments of state tracking, phase correction, and surface crossing are essential for DNA.

In summary, we have proposed a novel machine learning method to calculate electronic couplings between quasi-diabatic MOs. With an auxiliary sequence of rigid geometries, path independent state tracking and phase correction have been realized for the MOs. Without referring to nonadiabatic couplings, we take advantages of MOs that can be directly regarded to be quasi-diabatic and the smooth relation between quasi-diabatic quantities and the molecular structure to obtain the electronic couplings between quasi-adiabatic MOs independently. We have shown that HOMO and HOMO-1 for all nucleobases in DNA can be regarded as two-state systems, and the onsite energies given by trained ANNs satisfy the corresponding physical requirement. Our method has been successfully used to calculate electronic couplings between quasi-diabatic HOMOs in all base pairs. Strongly nonlocal electron-phonon coupling has been observed. With robust description of state tracking, phase correction, and surface crossing, the mean absolute results of electronic couplings between quasi-diabatic HOMOs differ significantly with traditional studies with adiabatic HOMOs. Based on the trained ANNs for onsite energies and electronic couplings, the Hamiltonian with one electronic state per nucleobase can be constructed directly for any sequence of DNA chain and used for further nonadiabatic dynamics simulation of charge transport with computational efficiency significantly higher than *ab initio* calculations. In principle, our method can be also applied to the calculation of exciton couplings in flexible systems, providing that representative electronic states can be adopted as references. Relevant studies are currently under way.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge online.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ljwang@zju.edu.cn

ORCID

Linjun Wang: 0000-0002-6169-7687

Notes

The authors declare no competing financial interests.

ACKNOWLEDGMENT

We thank Professor Zhenggang Lan for helpful discussions. L.W. acknowledges support from the National Natural Science Foundation of China (Grant Nos. 21873080 and 21922305).

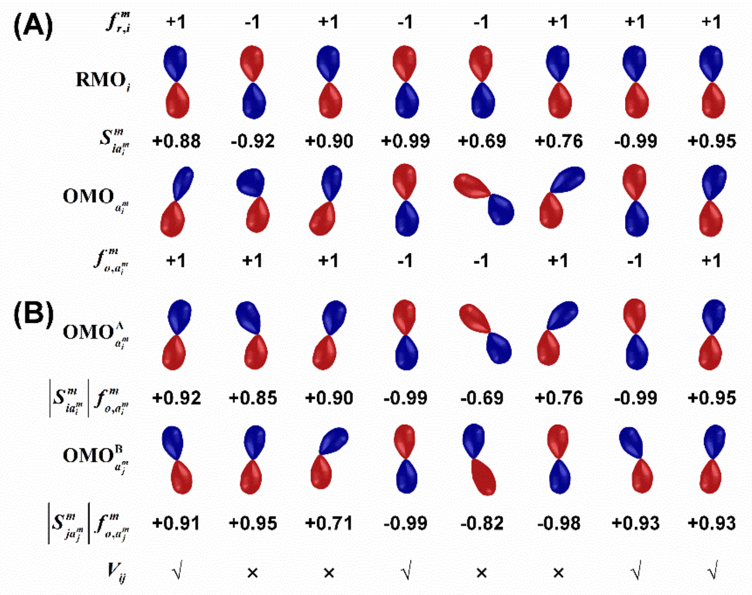

Figure 1. (A) Schematic representation of phase correction for RMOs and the corresponding state-tracked OMOs. a^m_i is the index of OMO with the maximum overlap to the i-th RMO in the m-th snapshot, $S^m_{ia^m_i}$ is the overlap between the i-th RMO and the a^m_i-th OMO, $f^m_{r,i}$ and f^m_{o,a^m_i} are phase factors for the i-th RMO and the a^m_i-th OMO, respectively. (B) Schematic representation of OMO selection for electronic couplings between quasi-diabatic MOs. V_{ij} is the electronic coupling between OMOs of the molecular pair A-B. ' √ ' means both $\mathrm{OMO}^{\mathrm{A}}_{a^m_i}$ and $\mathrm{OMO}^{\mathrm{B}}_{a^m_j}$ are regarded as quasi-diabatic MOs, while ' × ' means at least one of the OMOs is not a proper quasi-diabatic MO for a specified S_{c2} of 0.90.

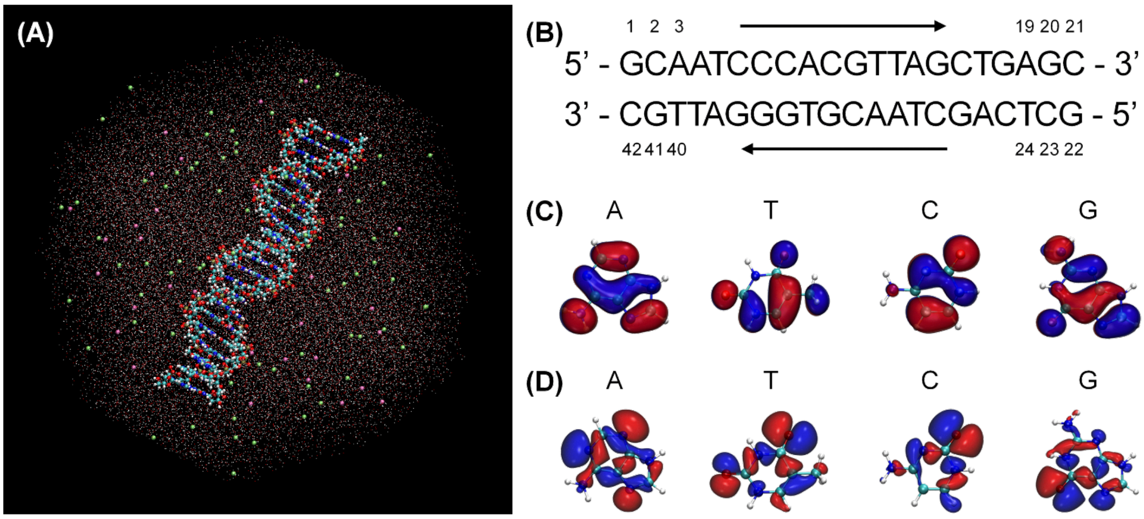
(A)
(B)
1 2 3
19 20 21
5' - GCAATCCCACGTTAGCTGAGC - 3'
3' - CGTTAGGGTGCAATCGACTCG - 5'
42 41 40
24 23 22
(C)
A
T
C
G
(D)
A
T
C
G

Figure 2. (A) System for molecular dynamics simulation and (B) sequence of the double-strand DNA studied in this work. In (A), oxygen and hydrogen atoms in water molecules are shown as red and white points, K^+ and Cl^- ions are shown as green and pink balls, respectively. In (B), nucleobases are numbered from 5' end to 3' end along the arrows. (C) HOMO and (D) HOMO-1 of nucleobases in the optimized geometry.

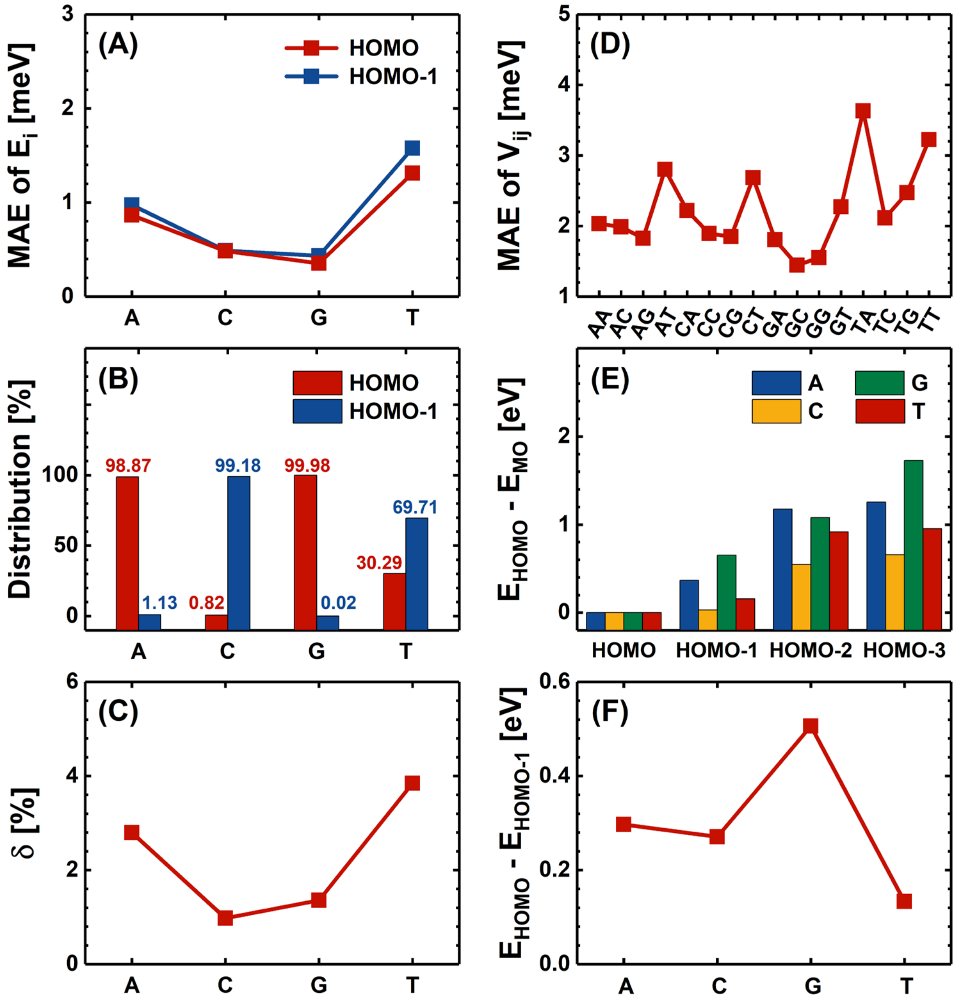

Figure 3. (A) MAEs of HOMO and HOMO-1 energies in the learning set for different nucleobases using the ANN model (6) in Table S1. (B) Proportion of quasi-diabatic HOMO and HOMO-1 that correspond to adiabatic HOMOs in the learning set. (C) Relative errors of the energy differences defined by Eq. (10). (D) MAEs of electronic couplings between HOMOs in the learning set for different base pairs using the ANN model (9) in Table S2 with 100 input dimensions by PCA. (E) Energy differences between HOMO and other MOs in rigid geometries. (F) Energy difference between HOMO and HOMO-1 averaged over the MD geometries. In (A) and (D), $S_{c2} = 0.90$ and tanh activation functions are used, and the number of training iterations and learning rate are 50000 and 2×10^{-5}, respectively.

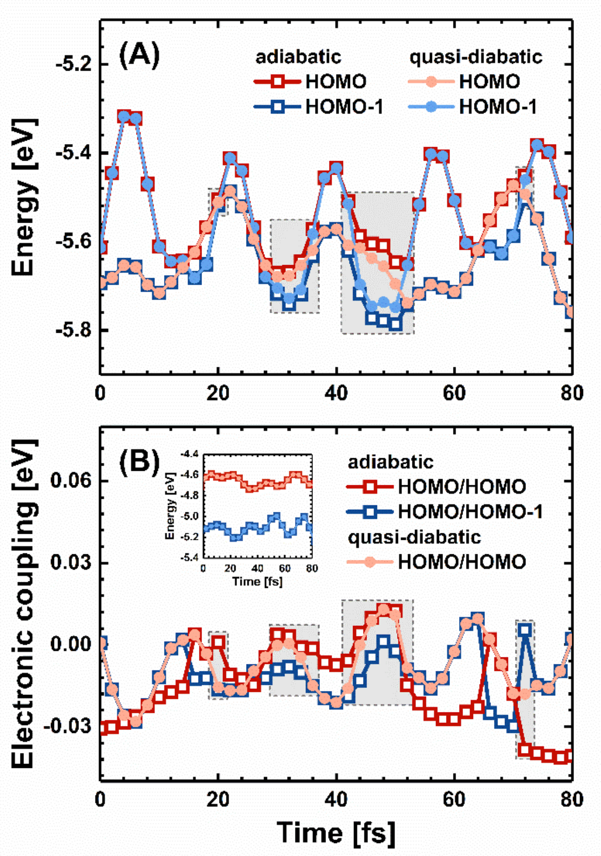

Figure 4. (A) Representative adiabatic and quasi-diabatic PESs of HOMO and HOMO-1 for nucleobase T. (B) Representative time-dependent electronic couplings between the HOMO of G and a specified MO (i.e., HOMO or HOMO-1) of T in adiabatic and quasi-diabatic representations. The corresponding adiabatic and quasi-diabatic PESs of HOMO and HOMO-1 for G are shown in the inset of (B). The ANN-predicted results in surface crossing areas are highlighted by shaded squares.

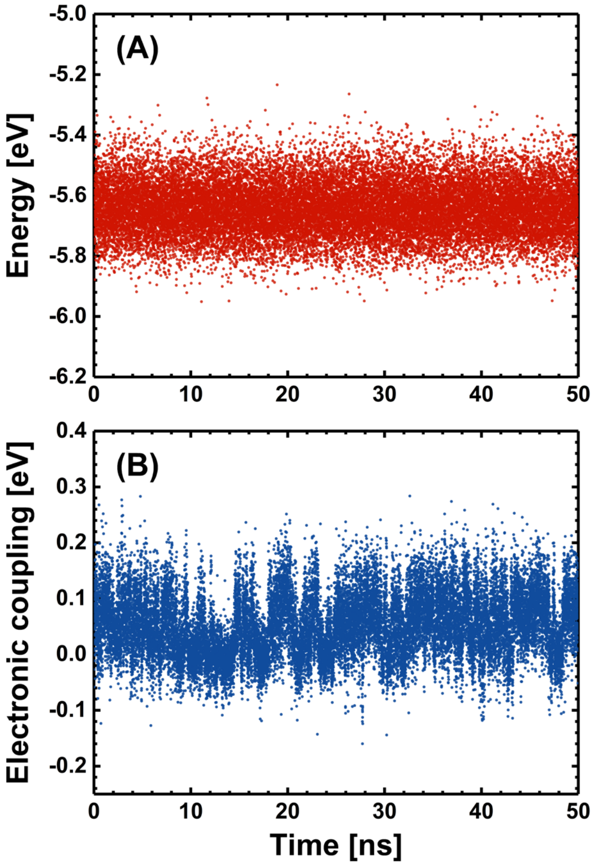

Figure 5. Representative time-dependent (A) onsite energies of quasi-diabatic HOMOs for nucleobase T and (B) electronic couplings between quasi-diabatic HOMOs in base pair GT.

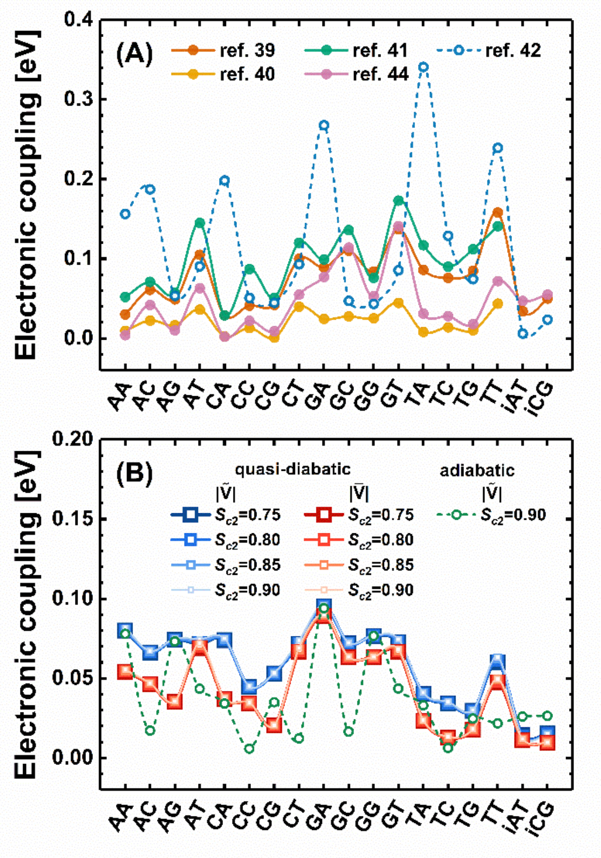

Figure 6. (A) Absolute values of electronic couplings between adiabatic HOMOs for different base pairs in the literature. (B) Average electronic couplings ($|\bar{V}|$ and $|\tilde{V}|$) for all the base pairs between quasi-diabatic and adiabatic HOMOs with specified S_{c2} values in this study. iAT and iCG represent inter-strand base pairs, while the others are all intra-strand base pairs.

REFERENCES

(1) Wang, L. J.; Qiu, J.; Bai, X.; Xu, J. B. Surface Hopping Methods for Nonadiabatic Dynamics in Extended Systems. *WIREs Comput. Mol. Sci.* **2020**, *10*, e1435.

(2) Matsika, S. Electronic Structure Methods for the Description of Nonadiabatic Effects and Conical Intersections. *Chem. Rev.* **2021**, *121*, 9407-9449.

(3) Van Voorhis, T.; Kowalczyk, T.; Kaduk, B.; Wang, L.-P.; Cheng, C.-L.; Wu, Q. The Diabatic Picture of Electron Transfer, Reaction Barriers, and Molecular Dynamics. *Annu. Rev. Phys. Chem.* **2010,** *61*, 149-170.

(4) Wang, L. J.; Prezhdo, O. V.; Beljonne, D. Mixed Quantum-Classical Dynamics for Charge Transport in Organics. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12395-12406.

(5) Prodhan, S.; Giannini, S.; Wang, L. J.; Beljonne, D. Long-Range Interactions Boost Singlet Exciton Diffusion in Nanofibers of *π*-Extended Polymer Chains. *J. Phys. Chem. Lett.* **2021**, *12*, 8188-8193.

(6) Baer, M. Electronic Non-Adiabatic Transitions Derivation of the General Adiabatic-Diabatic Transformation Matrix. *Mol. Phys.* **1980**, *40*, 1011-1013.

(7) Mead, C. A.; Truhlar, D. G. Conditions for the Definition of a Strictly Diabatic Electronic Basis for Molecular Systems. *J. Chem. Phys.* **1982**, *77*, 6090-6098.

(8) Pacher, T.; Cederbaum, L. S.; Köppel, H. Approximately Diabatic States from Block Diagonalization of the Electronic Hamiltonian. *J. Chem. Phys.* **1988**, *89*, 7367-7381.

(9) Atchity, G. J.; Ruedenberg, K. Determination of Diabatic States through Enforcement of Configurational Uniformity. *Theor. Chem. Acc.* **1997**, *97*, 47-58.

(10) Köppel, H.; Gronki, J.; Mahapatra, S. Construction Scheme for Regularized Diabatic States. *J. Chem. Phys.* **2001**, *115*, 2377-2388.

(11) Nakamura, H.; Truhlar, D. G. Direct Diabatization of Electronic States by the Fourfold Way. II. Dynamical Correlation and Rearrangement Processes. *J. Chem. Phys.* **2002**, *117*, 5576-5593.

(12) Subotnik, J. E.; Yeganeh, S.; Cave, R. J.; Ratner, M. A. Constructing Diabatic States from Adiabatic States: Extending Generalized Mulliken-Hush to Multiple Charge Centers with Boys Localization. *J. Chem. Phys.* **2008**, *129*, 244101.

(13) Venghaus, F.; Eisfeld, W. Block-Diagonalization as a Tool for the Robust Diabatization of High-Dimensional Potential Energy Surfaces. *J. Chem. Phys.* **2016**, *144*, 114110.

(14) Zhang, Y.; Wang, W.; Lasorne, B.; Su, P. F.; Wu, W. Diabatization around Conical Intersections with a New Phase-Corrected Valence-Bond-Based Compression Approach. *J. Phys. Chem. Lett.* **2021**, *12*, 1885-1892.

(15) Coropceanu, V.; Cornil, J.; da Silva Filho, D. A.; Olivier, Y.; Silbey, R.; Brédas, J.-L. Charge Transport in Organic Semiconductors. *Chem. Rev.* **2007**, *107*, 926-952.

(16) Wang, L. J.; Li, Q. K.; Shuai, Z. Effects of Pressure and Temperature on the Carrier Transports in Organic Crystal: A First-Principles Study. *J. Chem. Phys.* **2008**, *128*, 194706.

(17) Geng, H.; Zheng, X. Y.; Shuai, Z. G.; Zhu, L. Y.; Yi, Y. P. Understanding the Charge Transport and Polarities in Organic Donor-Acceptor Mixed-Stack Crystals: Molecular Insights from the Super-Exchange Couplings. *Adv. Mater.* **2015**, *27*, 1443-1449.

(18) Li, W. T.; Ren, J. J.; Shuai, Z. G. A General Charge Transport Picture for Organic Semiconductors with Nonlocal Electron-Phonon Couplings. *Nature Commun.* **2021**, *12*, 4260.

(19) Cave, R. J.; Newton, M. D. Generalization of the Mulliken-Hush Treatment for the Calculation of Electron Transfer Matrix Elements. *Chem. Phys. Lett.* **1996**, *249*, 15-19.

(20) Wu, Q.; Van Voorhis, T. Direct Calculation of Electron Transfer Parameters through Constrained Density Functional Theory. *J. Phys. Chem. A* **2006**, *110*, 9212-9218.

(21) Valeev, E. F.; Coropceanu, V.; da Silva Filho, D. A.; Salman, S.; Brédas, J.-L. Effect of Electronic Polarization on Charge-Transport Parameters in Molecular Organic Semiconductors. *J. Am. Chem. Soc.* **2006**, *128*, 9882-9886.

(22) Baumeier, B.; Kirkpatrick, J.; Andrienko, D. Density-Functional Based Determination of Intermolecular Charge Transfer Properties for Large-Scale Morphologies. *Phys. Chem. Chem. Phys.* **2010**, *12*, 11103-11113.

(23) Pavanello, M.; Van Voorhis, T.; Visscher, L.; Neugebauer, J. An Accurate and Linear-Scaling Method for Calculating Charge-Transfer Excitation Energies and Diabatic Couplings. *J. Chem. Phys.* **2013**, *138*, 054101.

(24) Kubas, A.; Hoffmann, F.; Heck, A.; Oberhofer, H.; Elstner, M.; Blumberger, J. Electronic Couplings for Molecular Charge Transfer: Benchmarking CDFT, FODFT, and FODFTB against High-Level *Ab Initio* Calculations. *J. Chem. Phys.* **2014**, *140*, 104105.

(25) Bai, X.; Qiu, J.; Wang, L. J. An Efficient Solution to the Decoherence Enhanced Trivial Crossing Problem in Surface Hopping. *J. Chem. Phys.* **2018**, *148*, 104106.

(26) Qiu, J.; Bai, X.; Wang, L. J. Crossing Classified and Corrected Fewest Switches Surface Hopping. *J. Phys. Chem. Lett.* **2018**, *9*, 4319-4325.

(27) Qiu, J.; Bai, X.; Wang, L. J. Subspace Surface Hopping with Size-Independent Dynamics. *J. Phys. Chem. Lett.* **2019**, *10*, 637-644.

(28) Lenzen, T.; Manthe, U. Neural Network Based Coupled Diabatic Potential Energy Surfaces for Reactive Scattering. *J. Chem. Phys.* **2017**, *147*, 084105.

(29) Williams, D. M. G.; Eisfeld, W. Neural Network Diabatization: A New Ansatz for Accurate High-Dimensional Coupled Potential Energy Surfaces. *J. Chem. Phys.* **2018**, *149*, 204106.

(30) Guan, Y. F.; Guo, H.; Yarkony, D. R. Neural Network Based Quasi-Diabatic Hamiltonians with Symmetry Adaptation and a Correct Description of Conical Intersections. *J. Chem. Phys.* **2019**, *150*, 214101.

(31) Guan, Y. F.; Zhang, D. H.; Guo, H.; Yarkony, D. R. Representation of Coupled Adiabatic Potential Energy Surfaces Using Neural Network Based Quasi-Diabatic Hamiltonians: 1,2 ^{2}A' States of LiFH. *Phys. Chem. Chem. Phys.* **2019**, *21*, 14205-14213.

(32) Yin, Z. X.; Guan, Y. F.; Fu, B.; Zhang, D. H. Two-State Diabatic Potential Energy Surfaces of ClH_2 Based on Nonadiabatic Couplings with Neural Networks. *Phys. Chem. Chem. Phys.* **2019**, *21*, 20372-20383.

(33) Williams, D. M. G.; Eisfeld, W. Complete Nuclear Permutation Inversion Invariant Artificial Neural Network (CNPI-ANN) Diabatization for the Accurate Treatment of Vibronic Coupling Problems. *J. Phys. Chem. A* **2020**, *124*, 7608-7621.

(34) Shu, Y. N.; Truhlar, D. G. Diabatization by Machine Intelligence. *J. Chem. Theory Comput.* **2020**, *16*, 6456-6464.

(35) Shu, Y. N.; Varga, Z.; de Oliveira-Filho, A. G. S.; Truhlar, D. G. Permutationally Restrained Diabatization by Machine Intelligence. *J. Chem. Theory Comput.* **2021**, *17*, 1106-1116.

(36) Yin, Z. X.; Braams, B. J.; Fu, B. N.; Zhang, D. H. Neural Network Representation of Three-State Quasidiabatic Hamiltonians Based on the Transformation Properties from a Valence Bond Model: Three Singlet States of H_3^+. *J. Chem. Theory. Comput.* **2021**, *17*, 1678-1690.

(37) Wang, C.-I.; Joanito, I.; Lan, C.-F.; Hsu, C.-P. Artificial Neural Networks for Predicting Charge Transfer Coupling. *J. Chem. Phys.* **2020**, *153*, 214113.

(38) Bag, S.; Aggarwal, A.; Maiti, P. K. Machine Learning Prediction of Electronic Coupling between the Guanine Bases of DNA. *J. Phys. Chem. A* **2020**, *124*, 7658-7664.

(39) Voityuk, A. A.; Rösch, N.; Bixon, M.; Jortner, J. Electronic Coupling for Charge Transfer and Transport in DNA. *J. Phys. Chem. B* **2000**, *104*, 9740-9745.

(40) Brunaud, G.; Castet, F.; Fritsch, A.; Kreissler, M.; Ducasse, L. Electron Interactions between Nucleoside Pairs in Canonical B-DNA: I. Transfer Integrals. *J. Phys. Chem. B* **2001**, *105*, 12665-12673.

(41) Olofsson, J.; Larsson, S. Electron Hole Transport in DNA. *J. Phys. Chem. B* **2001**, *105*, 10398-10406.

(42) Troisi, A.; Orlandi, G. The Hole Transfer in DNA: Calculation of Electron Coupling between Close Bases. *Chem. Phys. Lett.* **2001**, *344*, 509-518.

(43) Troisi, A.; Orlandi, G. Hole Migration in DNA: A Theoretical Analysis of the Role of Structural Fluctuations. *J. Phys. Chem. B* **2002**, *106*, 2093-2101.

(44) Senthilkumar, K.; Grozema, F. C.; Guerra, C. F.; Bickelhaupt, F. M.; Lewis, F. D.; Berlin, Y. A.; Ratner, M. A.; Siebbeles, L. D. A. Absolute Rates of Hole Transfer in DNA. *J. Am. Chem. Soc.* **2005**, *127*, 14894-14903.

(45) Voityuk, A. A. Fluctuation of the Electronic Coupling in DNA: Multistate *versus* Two-State Model. *Chem. Phys. Lett.* **2007**, *439*, 162-165.

(46) Kubař, T.; Elstner, M. What Governs the Charge Transfer in DNA? The Role of DNA Conformation and Environment. *J. Phys. Chem. B* **2008**, *112*, 8788-8798.

(47) Kubař, T.; Woiczikowski, P. B.; Cuniberti, G.; Elstner, M. Efficient Calculation of Charge-Transfer Matrix Elements for Hole Transfer in DNA. *J. Phys. Chem. B* **2008**, *112*, 7937-7947.

(48) Grozema, F. C.; Tonzani, S.; Berlin, Y. A.; Schatz, G. C.; Siebbeles, L. D. A.; Ratner, M. A. Effect of Structural Dynamics on Charge Transfer in DNA Hairpins. *J. Am. Chem. Soc.* **2008**, *130*, 5157-5166.

(49) Akimov, A. V. A Simple Phase Correction Makes a Big Difference in Nonadiabatic Molecular Dynamics. *J. Phys. Chem. Lett.* **2018**, *9*, 6096-6102.

(50) Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2021**, *121*, 9873-9926.

(51) Pasi, M.; Maddocks, J. H.; Beveridge, D.; Bishop, T. C.; Case, D. A.; Cheatham, T. E., III; Dans, P. D.; Jayaram, B.; Lankas, F.; Laughton, C.; Mitchell, J.; Osman, R.; Orozco, M.; Pérez, A.; Petkevičiūtė, D.; Spackova, N.; Sponer, J.; Zakrzewska, K.; Lavery, R. μABC: A Systematic Microsecond Molecular Dynamics Study of Tetranucleotide Sequence Effects in B-DNA. *Nucleic Acids Res.* **2014**, *42*, 12272-12283.

(52) Ivani, I.; Dans, P. D.; Noy, A.; Pérez, A.; Faustino, I.; Hospital, A.; Walther, J.; Andrio, P.; Goñi, R.; Balaceanu, A.; Portella, G.; Battistini, F.; Lluis Gelpí, J.; González, C.; Vendruscolo, M.; Laughton, C. A.; Harris, S. A.; Case, D. A.; Orozco, M. Parmbsc1: A Refined Force Field for DNA Simulations. *Nat. Methods* **2016**, *13*, 55-58.

(53) Meggers, E.; Michel-Beyerle, M. E.; Giese, B. Sequence Dependent Long Range Hole Transport in DNA. *J. Am. Chem. Soc.* **1998**, *120*, 12950-12955.

(54) Lewis, F. D.; Liu, X. Y.; Liu, J. Q.; Miller, S. E.; Hayes, R. T.; Wasielewski, M. R. Direct Measurement of Hole Transport Dynamics in DNA. *Nature* **2000**, *406*, 51-53.

(55) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.;

Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16, revision C.01. Gaussian Inc., Wallingford, CT, **2019**.

(56) Tommasini, M.; Chernyak, V.; Mukamel, S. Electronic Density-Matrix Algorithm for Nonadiabatic Couplings in Molecular Dynamics Simulations. *Int. J. Quantum Chem.* **2001**, *85*, 225-238.