Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2023, **2**, 1688

Received 14th July 2023
Accepted 13th October 2023

DOI: 10.1039/d3dd00130j

rsc.li/digitaldiscovery

# Atomic fragment approximation from a tensor network†

Haoxiang Lin and Xi Zhu *

We propose atomic-fragment approximation (AFA), which uses the tensor network (TN) as a platform to estimate the molecular properties through "adding up" fragment properties. The AFA framework employs graph neural networks to predict the matrix product states (MPSs) for atoms and matrix product operators (MPOs) for bonds, which are then contracted to obtain the full TN for the full molecule. Subsequent neural network layers then predict molecular properties based on the TN contraction outcome. AFA addresses the limitation of density functional approximation (DFA) by reusing previously calculated results and maintaining constant complexity in fragment contraction regardless of the fragment size. We further show that AFA can overcome error accumulation by optimizing the intermediate fragments. AFA demonstrates the ability to predict the reaction intermediates by calculating and comparing the bond-breaking energies. The experiment also showcases excellent accuracy in reaction intermediate prediction and reaction energy prediction.



## Introduction

Determining the structure–property relationship is essential for discovering drugs,[1] proteins,[2] and catalysts.[3] The most commonly used method is the parameter-free first principles approach, including solving Schrodinger's equation, the density functional approximation (DFA),[4] or density functional theory (DFT). However, DFA lacks transferability from fragments to the whole system, thus similar structures require a duplicate calculation in an "*ab initio*" way, as shown in Fig. 1. Chemical reaction discovery is the key to the design of new reactions, typically necessitating DFA calculations and specialized expertise.[5] Notably, the calculation complexity of DFA scales approximately cubically with molecule size,[6,7] making non-reaction regions computationally expensive compared to the reaction region, which is usually much smaller.

Many machine learning/deep learning approaches have been developed to address the size and charge problems in DFA.[8] Generally, machine learning/deep learning approaches predict by learning patterns in training data, adjusting internal parameters during training, and then applying these learned patterns to make predictions on new, unseen data. One class of these approaches, known as DFA-NNs,[9–11] employ deep learning models that utilize the electron-density-related properties as intermediates, primarily for fitting purposes. However, a significant drawback of DFA-NNs is the unavoidable increase in error with system size, typically exhibiting linear growth.[7,9,11] This problem is also known as error accumulation. As shown in ESI S1,† even state-of-the-art models like torchANI[7] and torchANI-2x[12] still encounter this problem of error accumulation. This accumulation makes distinguishing between energy differences caused by different chemical groups and those resulting from calculation errors difficult. To address this issue, one possible solution is to reuse and modify the calculated results for shared fragments. The fragment molecular orbital method (FMO)[13] implements this approach by using molecular orbital fragments. However, the intermediate step of FMO includes self-consistent approaches, which we expect to bypass *via* machine learning.

The tensor network (TN) framework is a geometry of low-order contracted tensors,[14,15] whose calculation process is fragment-by-fragment, close to the "adding" of properties. Traditional TN methods are mainly first-principles-based, meaning they rely on fundamental physical laws and principles to derive their results, with few relying on the experiment data. In ref. 16, the authors use TN methods to solve the Hubbard model starting from its Hamiltonian. These *ab initio* density matrix renormalization group (DMRG) methods are also included in benchmark systems like π-electron systems, main-group and transition metal dimers, and Mn-oxo-salen and Fe-porphine organometallic compounds.[17] Meanwhile, the TN is a powerful tool in the area whose correlation or entanglement entropy satisfies the area law,[18,19] besides quantum physics,[20–22] thus, the TN is expected to be the solution

*School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, China. E-mail: zhuxi@cuhk.edu.cn*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3dd00130j

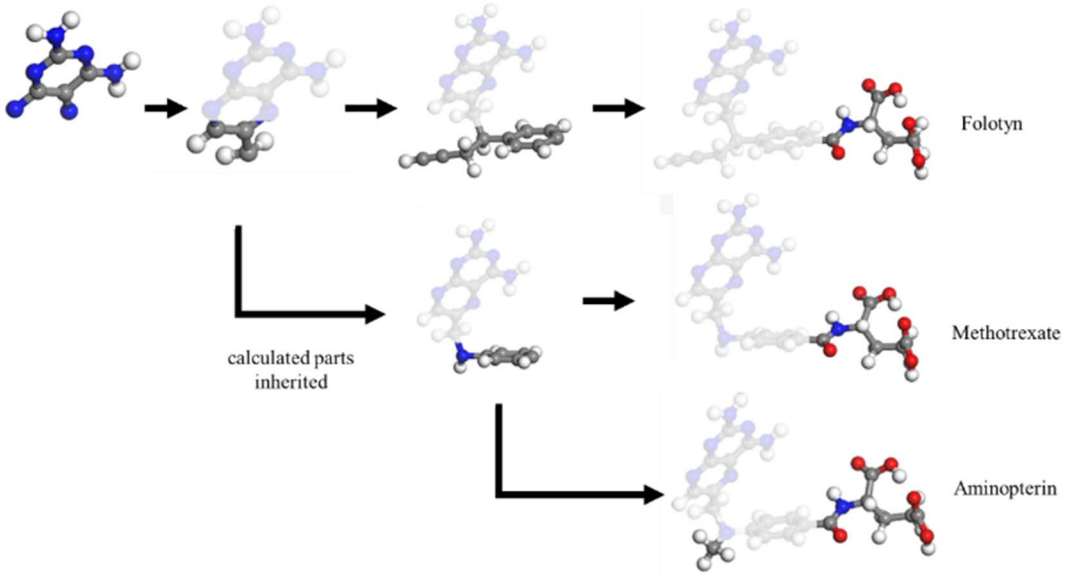


Fig. 1 The calculation process for density-functional approximation (DFA) and atomic-fragment approximation (AFA). While DFA requires separate calculations for each molecule, making it difficult to convey calculated information from left to right, AFA can reuse calculated results for shared fragments, as shown in the shaded parts.

for simulating the "adding" of properties. These traditional tensor network methods greatly inspired us.

In this work, we propose the atomic-fragment approximation (AFA) to provide the structure–property relationship by "adding up" fragments. We first demonstrate the algorithm of AFA. The AFA framework employs a graph neural network to predict the matrix product states (MPSs) for atoms and matrix product operators (MPOs) for bonds, which are then used in a contraction scheme to obtain the TN for the full molecule. A neural network layer is then used to predict the molecular properties based on the TN contraction results. Thanks to the step-by-step contraction scheme of the TN, AFA can realize the "adding up" of fragment properties. AFA is designed to capture the correlation between radicals, which is also applicable to large molecules. This ability makes AFA overcome the limitation of density functional approximation (DFA) by reusing previously calculated results, and we also show that the AFA algorithm can be used for chemical reaction prediction. Through experiments, we show that AFA can reduce the error accumulation in both momenta and real space by optimizing the internal TN states. Error accumulation is a problem that current DFA-simulating NN can never overcome, due to the contaminated correlation of electron density between radicals. We also demonstrate that the AFA algorithm can predict the intermediate with high accuracy and accurately predict the energy barrier for transforming reactants into transient intermediates and then forming the product.

## Theoretical background

### The tensor network as a platform for estimating molecular properties

AFA predicts the structure–property relationship through an "adding" process for each fragment, which has three steps. First, we transform the input molecule into a geometry-enhanced representation (GER),[23] which focuses on the correlation between bonds and the correlation between atoms. After that, we map the wavefunction ansatz of these fragments into their corresponding TN states, including matrix product states[24] (MPSs) for atoms and matrix product operators (MPOs) for bonds. Here atoms and bonds are entirely separated. Finally, these TN states build the tensor network, whose contraction results go through a decoding layer for the target properties. Thanks to the algorithm, AFA can calculate the target properties fragment-by-fragment, which enables the reuse of calculated information.

We first transformed the input molecule structure into the geometry-enhanced representation (GER),[23] which specifically focuses on two parts, the atom-bond graph $G$, representing the correlation between atoms, and the bond-angle graph $D$ representing the correlation between bonds. In this process, we utilize the atom-bond graph for the atom's MPS, while the bond-angle graph is employed for the bond's MPO. Detailed information about the GER is given in ESI S2.†

We begin the second step, mapping into TN states, by defining the high-dimensional space of wave function. Here we work in the Born–Oppenheimer approximation; the wavefunction of the entire molecule depends on the atom positions ($\boldsymbol{r}_1$, $\boldsymbol{r}_2$,…, $\boldsymbol{r}_n$), where $n$ is the total number of atoms. The wavefunction ansatz $\Psi$(molecule) is represented by local map multiplications of MPSs $A = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n\}$:

$$\Psi(\text{molecule}) = \phi_1(\boldsymbol{r}_1) \otimes \phi_2(\boldsymbol{r}_2) \otimes \ldots \otimes \phi_n(\boldsymbol{r}_n) \sum_{\{\alpha_1\},\ldots,\{\alpha_n\}} \boldsymbol{A}_1 \boldsymbol{A}_2 \ldots \boldsymbol{A}_n \quad (1)$$

Here $\{\alpha_n\}$ is the connecting edge for MPS $\boldsymbol{A}_n$, $\otimes$ is the Kronecker product, and $\phi_i(\boldsymbol{r}_i)$ is the orthonormal basis, which is eliminated

during further calculation. Each MPS represents a high-dimensional tensor, whose number of edges represents its dimension. Here $\phi_i(\boldsymbol{r}_i)$ includes information about orbital shapes such as the s orbital, and it also includes orbital spins like spin-up and spin-down. In Nesbet's theorem,[25] the correlation energy can be written exactly as a sum of contributions from occupied pairs of spin orbitals, while in this step, all necessary information for exchange energy calculation is encoded in the tensors. Details on TN states' representation are given in ESI S3.†

The MPS of atoms $\boldsymbol{A}_1, \boldsymbol{A}_2,\ldots,\boldsymbol{A}_n$ (like the grey rounded rectangles for a carbon atom and the red rounded rectangles for an oxygen atom, as shown in Fig. 2, step from correlation to TN states) is estimated through a GNN.

$$\boldsymbol{A}_i = \boldsymbol{f}^{\mathrm{GNN}}(r_i, b_G(r_i)) \tag{2}$$

$f^{\mathrm{GNN}}$ corresponds to the GNN calculation. $b_G(r_i)$ is the graph that represents the atom's nearest atoms, generated from the atom-bond graph $G$ of the GER of the molecule. The nodes of $b_G(r_i)$ are the nearest atoms of the atom with position $r_i$, while the edges of $b_G(r_i)$ correspond to the bond length. One may treat this step as a kernel trick, which maps these features into a high-dimensional space.

The MPOs of bond $O_{i,j} = \{O_{i,j}^{(0)}, \boldsymbol{O}_{i,j}^{(1)},\ldots,O_{i,j}^{(K)}\}$ for atom $i$ and $j$ are estimated through a GNN, whose input is this bond's nearest bond obtained from the bond-angle graph $D$ of the GER:

$$\boldsymbol{O}_{i,j}^{(k)} = f^{\mathrm{GNN},(k)}(r_i, b_D(i,j)) \tag{3}$$

Here $k$ is the index of the bond MPO, while $K$ is a hyper-parameter that determines the number of required MPOs. $b_D(i,j)$ is the graph that represents the bond's nearest bonds, generated from the bond-angle graph $D$. The nodes of $b_D(i,j)$ are the nearest bonds, while the edges of $b_D(i,j)$ correspond to the bond angle. Fig. 2 shows an example using the carbon–oxygen radical in ethanol. The grey/red blocks represent the MPS of atom carbon/oxygen, while the grey-red blocks refer to the MPO of the carbon–oxygen bond. We define the radical TN states of this carbon-oxygen radical as the combination of two atom MPSs and the bond MPO.

The target property is calculated from the decoding of the obtained TN states. We assume that pairs of atoms have no impact for the target properties unless a bond exists between them. As shown in Fig. 2, step from TN states to target properties, the final prediction $P$ for the desired property comes from the TN contraction results $\boldsymbol{T} = \sum_{i,j}\langle \boldsymbol{A_i}|\boldsymbol{O_{ij}}|\boldsymbol{A_j}\rangle$:

$$P = f^{\mathrm{decode}}\left(\left[T^{(1)}, T^{(2)}, \ldots, T^{(n)}\right]\right) = f^{\mathrm{decode}}\left(\left[\sum_{i,j}\left\langle \boldsymbol{A}_i\middle|\boldsymbol{O}_{i,j}^{(1)}\middle|\boldsymbol{A}_j\right\rangle, \ldots, \sum_{i,j}\langle \boldsymbol{A}_i|\boldsymbol{O}_{i,j}{}^{(K)}|\boldsymbol{A}_j\rangle\right]\right) \tag{4}$$

Here the decoding neural network $f^{\mathrm{decode}}$ is the multi-layer perceptron (MLP), whose input is the contraction results of all-atom MPSs and bond MPOs. One may treat this step as bypassing the self-consistent field approach using necessary parameters calculated by tensor network contraction. The pseudocodes are given in ESI S4.†

### "Add" the fragments

Thanks to the tensor contraction in eqn (4), AFA can "add" the fragment properties to obtain the properties of their combination. These steps evaluate the correlation between various fragment radicals. We use radicals instead of closed-shell

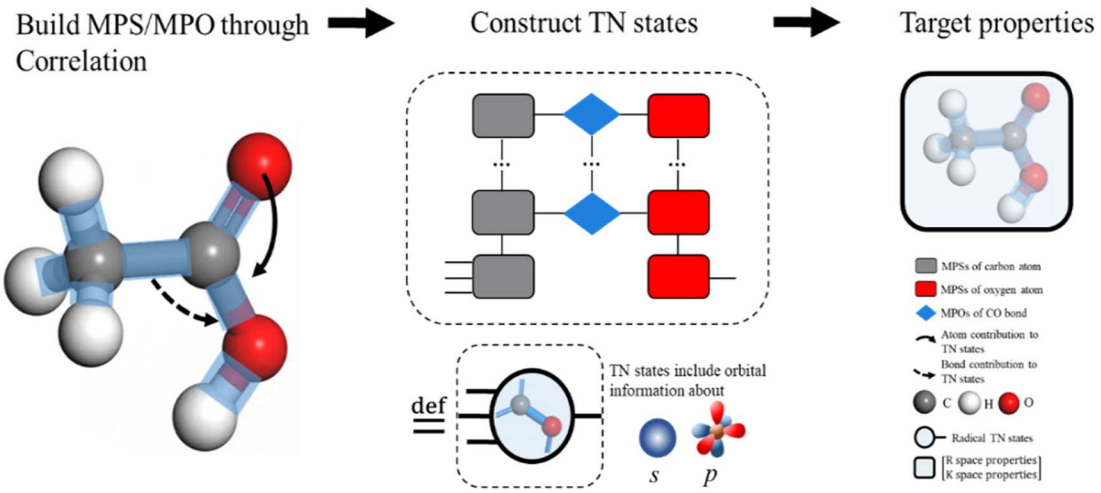


Fig. 2 Illustration of the AFA algorithm using a carbon–oxygen radical in ethanol as an example. The input molecular spatial structure is first converted into the geometry-enhanced representation (GER), which captures the correlation between the nearest bonds and atoms. Subsequently, TN states are calculated based on these correlations, where bonds and atoms are entirely separated. All the nearest atoms of the target carbon/oxygen atoms go through a graph neural network (GNN) for matrix product states (MPSs; grey/red block for carbon/oxygen in the middle graph) of their wavefunction. The nearest bonds of the carbon–oxygen bond undergo a separate GNN to compute matrix product operators (MPOs; blue blocks). The contraction of MPS and MPO results in the tensor network state of this carbon–oxygen bond. This TN state contains necessary information for the final prediction of target properties $P$, which can be decoded through additional neural network layers.

molecules because many large molecules are formulated from radicals, and the TN states of radicals encompass vital information for forming bonds with other radicals. As shown in Fig. 2, the contraction of MPSs (grey blocks for the carbon atom and red blocks for the oxygen atom) and MPOs (the blue blocks) gives a radical TN state with three dangling edges. These edges refer to the bonds to be connected. This radical TN state can be used for all molecules that share such fragments. Polarization and charge transfer commonly exist between functional parts of large molecules. When dealing with unknown molecules, we leverage information from molecules with similar radicals to account for these effects. During the fragment "adding" process, the 3D structure can be modified by changing the bond angle and bond length in $b_G(r_i)$. Therefore, as shown in Fig. 3a, AFA first maps the structure summation into the TN state summation. Mathematically, AFA is trying to add the interaction terms between multiple dependent features:

$$P = p^0 + \sum_{ij} p(f_i, f_j) + \sum_{ijk} p(f_i, f_j, f_k) + \ldots \quad (5)$$

Here $f_i$ corresponds to the features of fragments and $p(\cdot)$ calculates the contribution from interactions between multiple fragments, respectively. The detailed interpretation of the atom approach is attached in ESI S5.† Physically, AFA simulates the perturbation theory. The information for performing perturbation theory calculations is stored in the states of each radical TN state. The "adding" process of AFA is conducting the perturbation theory calculation. The detailed interpretation is attached in ESI S6.† It is important to note that AFA does not target any high-level density functional approximation (DFA) theories. We used the PBE/6-31G level of theory to label our training data and benchmark our method, and AFA is trained to predict results without any information about the level of theory. However, we believe that the AFA algorithm itself can simulate the effects of perturbation theory and correlation energy, which is why we mentioned them.

Our methodology draws inspiration from the fragment molecular orbital (FMO) methods, where interactions between multiple fragments are primarily captured by higher-order terms in the FMO expansion.[26] The tensor network framework we employ has the capability to account for these fragment interactions, effectively describing the higher-order terms in FMO.

One advantage of AFA is the constant complexity for the contraction of pairs of radical TN states. The computation complexity of AFA must not exceed a constant once all essential molecule component information is provided. The total time complexity of one radical contraction must not exceed the constant $C_{\text{total}}$:

$$C_{\text{total}} = n \times C_{\text{M}} + C_{\text{NN}} + C_{\text{b}} \quad (6)$$

Here, the complexity of revising the connected atom MPS has an upper limit $C_{\text{M}}$ and $n$ is the number of atoms that lie near the formulated bond, typically two. The complexity of this bond

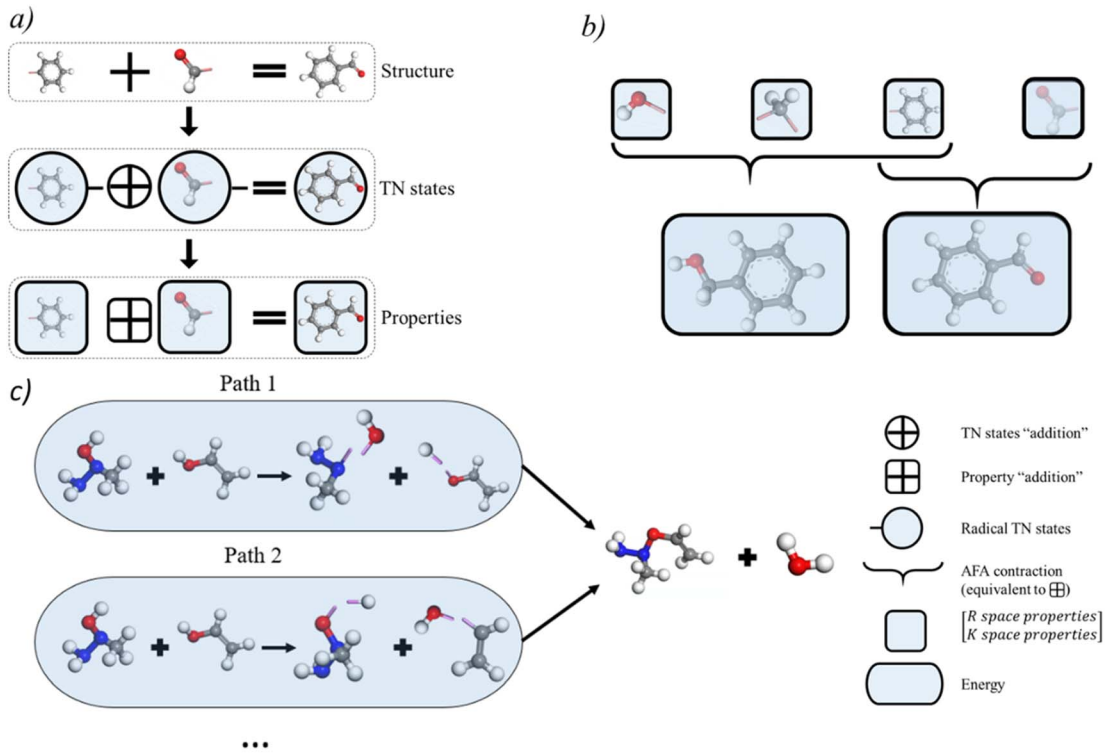


Fig. 3 Application of AFA. (a) Working scheme of "adding". AFA first maps the structure summation into the tensor network (TN) state summation, which contains all necessary information for target properties. Consequently, the TN states summation leads to the property summation. Here the unconnected pink stick represents bonds to be connected. (b) When estimating the TN states of different but similar molecules, like benzyl alcohol and benzaldehyde, AFA can directly use the calculated TN states of benzene, eliminating the need for re-calculation. (c) AFA's workflow for chemical reaction prediction, which involves identifying the most feasible intermediates and computing the transition energy.

must be a constant $C_b$. The complexity of the decoding layer is always a constant $C_{NN}$. More detailed descriptions of contraction complexity are given in ESI S7.†

The basic premise for chemical reaction prediction is the prediction of the intermediate and the energy requirement for transforming reactants into transient intermediates and then forming the product. Here we show that AFA has the potential for reaction pathway prediction using one-step reactions. The one-step reactions are shown to achieve a balance of computational cost and reaction coverage in exploring reaction networks. We developed a refined method for chemical reaction prediction using AFA, as illustrated in Fig. 3c, drawing inspiration from YARP's two-step process.[27] This process involves identifying the reaction centre by analyzing the structural changes in reactants and subsequently predicting the reaction outcome by considering the identified centre, reactants, and reagents to generate the most probable products.[27] Initially, we computed the energies of reactants and products, retaining intermediate fragment results. Subsequently, we identified all bonds within reactants and products, calculating bond-breaking energies while excluding those involved in ring breaking or formation, for which we re-calculated the energy. We proceeded to determine all viable combinations of bond-breaking, approximating bond-breaking energy for each combination and selecting alternatives. The selection criteria encompass the exclusion of high-energy alternatives and the preference for intermediates present in both the lists of reactant and product intermediates. Once we established the resulting molecules for both reactants and products, we compared alternative intermediate lists, opting for intermediates present in both reactant and product lists. Lastly, we assessed the energy of all chosen intermediate molecules, including those with ring breaking or formation, selecting the intermediate with the lowest transition energy as the predicted intermediate. A detailed description of the algorithm is given in ESI S8.†

## Methods

To test the accuracy of molecular property prediction, we created the million molecule dataset obtained from QM9,[28] bindingDB,[29] Chembl,[30] and BDE.[31] We combine the molecular topology of all these datasets as input molecules. We removed some unstable structures and the number of samples is exactly one million. Properties are calculated through Gaussian 09 (ref. 32) at the PBE level of theory with basis 6-31G, including momentum ($K$) space properties and real ($R$) space properties. The unit is converted from Hartree to eV. The properties contain SMILES, the atomic position, the element type, the energy and the orbital energies. Here the orbital energies range from the fifth highest occupied molecular orbital-5 (HOMO-5) to the fifth lowest occupied molecular orbital (LUMO+5). Details on the dataset and unit conversion are given in ESI S9.†

The reaction graph depth (RGD1) dataset[33] is implemented to test the performance of chemical reaction prediction. It contains 176 992 organic reactions with validated transition states, activation energy, heat of reaction, reactant and product geometries, frequencies, and atom mapping. The reactions cover C, H, O, and N-containing molecules with up to 10 heavy atoms. The data are supplied at the GFN2-xTB and B3LYP-D3/TZVP levels of theory. We randomly selected 20% of the datasets for training and used the remaining 80% for testing. This is because the dataset implements different basis and our model requires fine-tuning.

## Results

### The reduction of error accumulation

AFA has two major advantages, the reduction of error accumulation and the avoidance of redundant calculation. We will first discuss the advantage of accuracy, while the advantage of calculation will be discussed later. Fig. 4a–d show the $R$ and $K$ space errors for AFA and DFA-simulating NN. In this case, the atomization energy serves as an example of the $R$ space property, while the orbital energy represents the $K$ space property. We select quantum deep field (QDF)[9] as an example of DFA-simulating NN due to its transferability and its underlying physics. QDF exhibits excellent transferability from small to large molecules, indicating its ability to reduce error accumulation. Meanwhile, QDF directly uses the electron density as the intermediate. In contrast, most other DFA-NNs typically use properties related to electron density, such as orbitals. Other state-of-the-art models like torchANI[7] and ANI-2x[12] also exhibit a similar tendency of error accumulation as shown in ESI.† Here the $R$ space error corresponds to the mean absolute error (MAE) for the total energy,

$$R \text{ space error} = |E_{\text{predict}} - E_{\text{DFA}}| \tag{7}$$

where $E_{\text{predict}}$ refers to the model output from AFA or DFA-simulating NN. The $K$ space error corresponds to the MAE of the HOMO and LUMO,

$$K \text{ space error} = \frac{1}{2} \sum_{i \in \{\text{HOMO, LUMO}\}} \left| E_{i,\text{predict}} - E_{i,\text{DFA}} \right| \tag{8}$$

AFA and DFT-NNs exhibit notably different tendencies: with enlarged molecule size, the error for DFA-NNs increases, while the error for AFA stays relatively flat. Here we use the DFT calculation results as the benchmark. Theoretically, DFT-NNs show a similar tendency to DFT since they share the same computation processes. The circle with a plus symbol inside shows the results for selected drugs, whose names are given in ESI S10.† One may notice a peak around the number of atoms 10–20 due to the first appearance of ring molecules. This peak information on this structure cannot be obtained through AFA method. The calculation details for QDF in Fig. 4 are given in ESI S11.† More detailed experiment results and the experiment parameters for AFA are given in ESI S12.†

Fig. 4 shows the reduction of error accumulation in $R$ space. This difference in tendency between AFA and DFT-NNs is due to the transferability difference, poor transferability accumulates error, especially with molecular size. DFA focuses on the pseudo-potential of atoms, but AFA focuses on the correlation. The error accumulation is addressed through the modification of intermediate TN states in AFA, as shown in Fig. 4b. For an

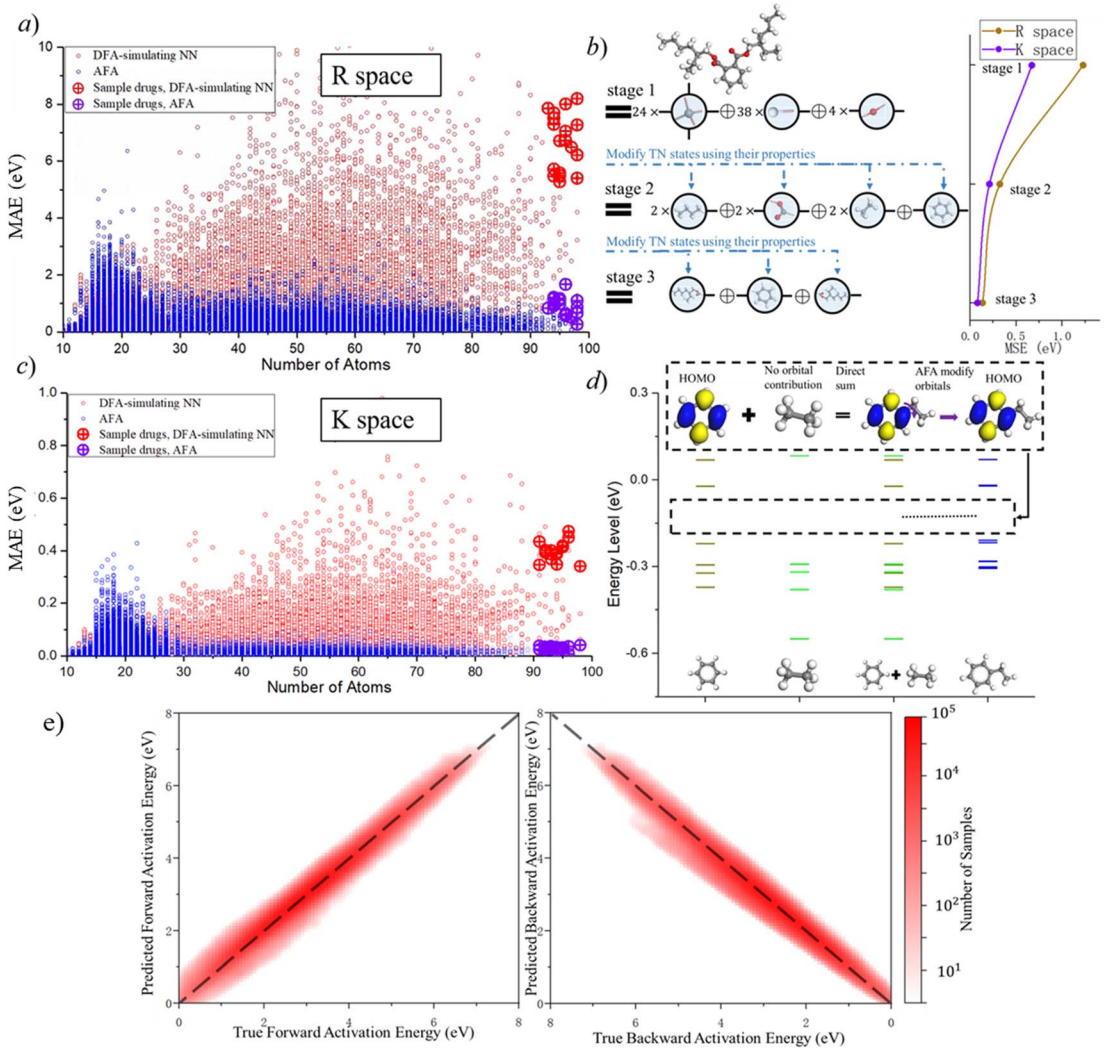


Fig. 4 Performance of AFA in predicting the molecular properties and chemical reaction. (a, c) Blue and red dots show the AFA and DFT-NN performance, respectively, with results in real (*R*) and momentum (*K*) space. Circles with plus symbols indicate results for selected drugs. (b) AFA accumulates error for few bonds compared to DFA-NNs, with progressive error reduction in different calculation stages. The di(2-ethylhexyl) phthalate (DEHP) molecule is used here as an example. (d) AFA accumulates error for a few orbitals compared to DFA-NNs. Here we use the HOMO of ethyl benzene as an example. The yellow, blue, and green lines represent the energy level of benzene, ethyl, and benzene ethyl, respectively. The third column is the direct merging of the first two columns. AFA merges and modifies orbitals to predict the HOMO of ethyl benzene. (e) The mean absolute error of the energy barrier for forward activation energy and backward activation energy. The color bar indicates the number of samples, while the dashed line represents the $y = x$ line, serving as a reference for comparison.

unknown large molecule like the di(2-ethylhexyl)phthalate (DEHP) molecule, AFA first decomposes this molecule into atomic TN states (the blue circle with edges), like atomic TN states of carbon with four unconnected bonds. Then their combination gives the radical TN states, which refer to stage 2 as mentioned in Fig. 4b. Then AFA further modifies them using their properties. The given TN states go through the neural network layer for the target properties, while the intermediate TN states can also be optimized through backpropagation. Then the intermediate TN states are modified using their properties iteratively until these intermediate radicals cannot merge into a known radical. The merging of radicals refers to the stage in Fig. 4b. This modification process greatly reduces error accumulation.

The reduction of error accumulation also holds in *K* space. The *K* space property mainly focuses on the orbital energies. Here we use ethyl benzene as an example, which is made by the contraction of a radical phenyl and the ethyl group. From the training set, AFA obtained the necessary information about these radicals from benzene and ethane. AFA first directly adds the energy level of these two radicals together and then modifies their orbitals. The yellow, blue, and green lines represent the energy levels of benzene, ethyl, and benzene ethyl, respectively. For the HOMO calculation of ethyl benzene, AFA first directly merges the HOMO of benzene with ethyl. Here ethyl contributes no orbital. By modifying the merged orbitals, AFA can predict the HOMO of ethylbenzene. Fig. 4d shows the example of the HOMO. Such a calculation of orbital energy is similar to that of *R*

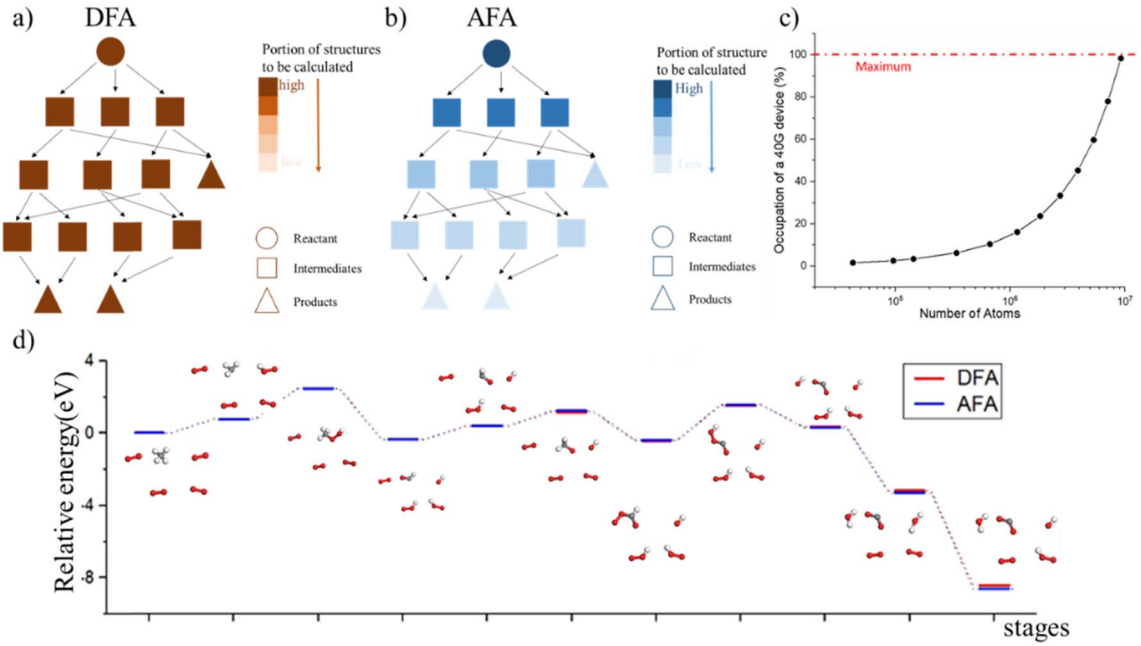


Fig. 5 Chemical reaction calculation. (a) During the reaction process, DFA calculates all structures for the reactant, intermediates, and product. (b) However, AFA reuses the calculated results, progressively reducing the portion of structures to be calculated. (c) The computation consumption. A 40G device quickly reaches capacity as the number of atoms increases. (d) The calculated relative energy for methane combustion, with the energy of the reactant set as zero. AFA demonstrates a MAE of less than 0.01 eV compared with DFA.

space property calculation. Therefore, it is not surprising that the reduction of error accumulation still holds in momentum space.

AFA can be used for suggesting potential reaction pathways, which include two aspects: the prediction of intermediates and the energy requirement for transforming reactants into transient intermediates before forming the product. The intermediate prediction accuracy of AFA is 94.34%, while a recent model has an accuracy of 93.8%.[34] Here the accuracy is measured as top-2 accuracy, where AFA provides the top two most likely intermediate candidates; if the correct intermediate has the same structure as one of them, we consider the prediction to be accurate. Generally speaking, small effects may affect the possible reaction pathways, so further validation and refinement are required to clearly identify the reaction pathway. Fig. 4e shows the accuracy of forward activation energy and backward activation energy, with the colour bar representing the number of samples. AFA demonstrates strong performance in the chemical reaction prediction for both intermediate candidate prediction and the energy difference calculation.

## The avoidance of redundant calculation

AFA's advantage of avoidance of redundant calculation makes it useful in chemical reaction prediction as shown in Fig. 5. Such an advantage also reduces the memory requirement for property prediction as shown in ESI S13.† During the molecular dynamics process, each step is similar to the previous step, with simply one bond breaking or formation. However, if each step is fully calculated, the computation device will be quickly fulfilled, as shown in Fig. 5c. Therefore, the adaption of previously calculated results greatly avoids redundant calculations. However, DFA can hardly re-use the calculated results. Fig. 6d shows the prediction result for methane combustion, AFA has an excellent agreement with DFA results, with an MAE less than 0.01 eV.

## Discussion

### Correlation contamination

AFA captures essential information about radicals from known molecules and transfers this knowledge to unknown large molecules. Here we made a comparison between AFA and DFA to show that current DFA-NN models face challenges when incorporating fragments due to their intermediate, density and related properties. This radical information includes both the radical itself and the correlation between this radical and other radicals. As depicted in Fig. 6a, although di(2-ethylhexyl)phthalate (DEHP) is an unknown large molecule, all its fragments appear in known molecules. We adopted the idea of tight binding, where the nearest radicals contribute most to the target properties. In the training set, besides the radical itself, AFA is trained with the correlation between pairs of radicals. For example, in benzyl alcohol, AFA is trained with radicals like *o*-phenylene and the correlation between radicals, like the correlation between *o*-phenylene and the carbinol group as shown in Fig. 6a.

However, correlation contamination occurs in the electron density calculations. DFA-NN (like QDF,[9] OrbNet,[11] and others[35]) utilizes the electron densities or their related properties for the intermediate step, which impedes the reuse of calculated

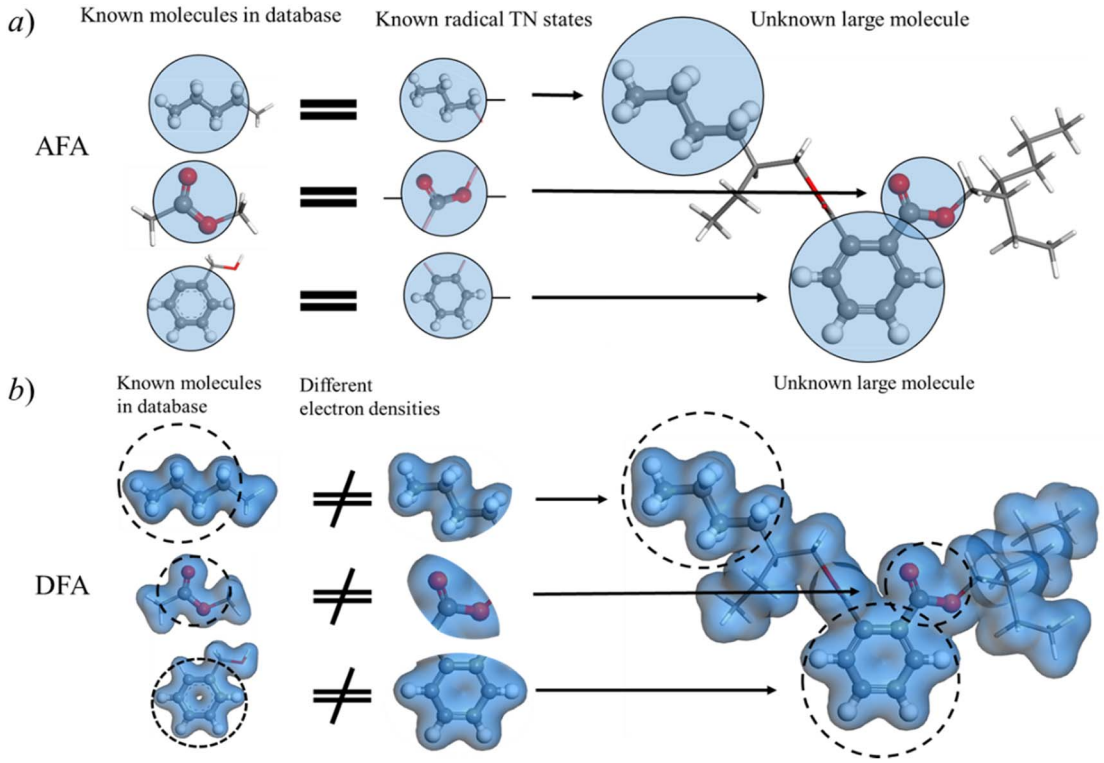


Fig. 6 Correlation contamination. (a) Radical TN states from small molecules can be directly applied to large molecules. Although DEHP is an unknown large molecule, all its radicals have been trained with known small molecules in the database, such as the butyl in pentane, the ester group in methyl acetate, and the *o*-phenylene in benzyl alcohol. (b) However, the electron densities in different molecules, even those sharing the same radical, can vary. Consequently, DFA and DFA-NN cannot reuse the calculated information due to correlation contamination in electron density estimations.

information. For example, as illustrated in Fig. 6, in the connection area of phenyl and carbinol groups of benzene alcohol, it is challenging to tell which radical contributes the electron density. The iteration of the Kohn–Sham equation simply gives the electron density functional, but it has nothing to say about the correlation between radicals.

## Conclusion

In this work, we developed atomic-fragment approximation (AFA), a novel approach for estimating the molecular structure–property relationship by mapping each fragment into its tensor network (TN) states and then contracting them. We show that AFA possesses two key advantages: avoiding redundant calculation and reducing error accumulation. The calculated result for radicals can be re-used for all molecules. The complexity of obtaining a molecule's properties is always constant, independent of the fragment's size, due to sufficient information on the fragments' TN state. Additionally, we have demonstrated that AFA can overcome error accumulation by optimizing intermediate radical TN states. AFA greatly avoids redundant calculation, and exhibits excellent accuracy in chemical reaction prediction. The MPS-based tensor network can be estimated through quantum computing,[20] indicating that a quantum computer can greatly enhance AFA. The AFA framework, combined with artificial intelligence techniques, holds great potential for advancing the field of physics.

## Data availability

The AFA model related code can be found at **https://github.com/hxlin97/AFA**.

## Author contributions

The manuscript was written with contributions from all authors. All authors approved the final version of the manuscript.

## Conflicts of interest

The authors declare no conflict of interest.

## Acknowledgements

Funding from the Shenzhen Fundamental Research Foundation (JCYJ20210324142213036) and China's National Natural Science Foundation (grant no. 22075240) is greatly appreciated.

## References

1 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah and M. Spitzer, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477.
2 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko, *Nature*, 2021, **596**, 583–589.
3 P. Schlexer Lamoureux, K. T. Winther, J. A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen and T. Bligaard, *ChemCatChem*, 2019, **11**, 3581–3601.
4 A. M. Teale, T. Helgaker, A. Savin, C. Adamo, B. Aradi, A. V. Arbuznikov, P. W. Ayers, E. J. Baerends, V. Barone and P. Calaminici, *Phys. Chem. Chem. Phys.*, 2022, **24**, 28700–28781.
5 J. Burés and I. Larrosa, *Nature*, 2023, **613**, 689–695.
6 F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K.-R. Müller, *Nat. Commun.*, 2017, **8**, 1–10.
7 X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, *J. Chem. Inf. Model.*, 2020, **60**, 3408–3415.
8 J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis and A. J. Cohen, *Science*, 2021, **374**, 1385–1389.
9 M. Tsubaki and T. Mizoguchi, *Phys. Rev. Lett.*, 2020, **125**, 206401.
10 H. Lin, S. Ye and X. Zhu, *Carbon*, 2022, **186**, 313–319.
11 Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby and T. F. Miller III, *J. Chem. Phys.*, 2020, **153**, 124111.
12 C. Devereux, J. S. Smith, K. K. Huddleston, K. Barros, R. Zubatyuk, O. Isayev and A. E. Roitberg, *J. Chem. Theory Comput.*, 2020, **16**, 4192–4202.
13 D. G. Fedorov, T. Nagata and K. Kitaura, *Phys. Chem. Chem. Phys.*, 2012, **14**, 7562–7577.
14 R. Orús, *Ann. Phys.*, 2014, **349**, 117–158.
15 R. Orús, *Nat. Rev. Phys.*, 2019, **1**, 538–550.
16 J. P. LeBlanc, A. E. Antipov, F. Becca, I. W. Bulik, G. K.-L. Chan, C.-M. Chung, Y. Deng, M. Ferrero, T. M. Henderson and C. A. Jiménez-Hoyos, *Phys. Rev. X.*, 2015, **5**, 041041.
17 R. Olivares-Amaya, W. Hu, N. Nakatani, S. Sharma, J. Yang and G. K.-L. Chan, *J. Chem. Phys.*, 2015, **142**, 034102.
18 M. M. Wolf, F. Verstraete, M. B. Hastings and J. I. Cirac, *Phys. Rev. Lett.*, 2008, **100**, 070502.
19 J. Eisert, M. Cramer and M. B. Plenio, *Rev. Mod. Phys.*, 2010, **82**, 277.
20 G. Vidal, *Phys. Rev. Lett.*, 2007, **99**, 220405.
21 W. Huggins, P. Patil, B. Mitchell, K. B. Whaley and E. M. Stoudenmire, *Quantum Sci. Technol.*, 2019, **4**, 024001.
22 I. Convy, W. Huggins, H. Liao and K. B. Whaley, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 015017.
23 X. Fang, L. Liu, J. Lei, D. He, S. Zhang, J. Zhou, F. Wang, H. Wu and H. Wang, *Nat. Mach. Intell.*, 2022, 1–8.
24 U. Schollwöck, *Ann. Phys.*, 2011, **326**, 96–192.
25 R. Nesbet, *Phys. Rev.*, 1958, **109**, 1632.
26 S. Tanaka, Fragment Molecular Orbital Method as Cluster Expansion, *Recent Advances of the Fragment Molecular Orbital Method: Enhanced Performance and Applicability*, 2021, pp. 3–14.
27 Q. Zhao and B. M. Savoie, *Nat. Comput. Sci.*, 2021, **1**, 479–490.
28 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
29 T. Liu, Y. Lin, X. Wen, R. N. Jorissen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
30 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich and B. Al-Lazikani, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
31 P. C. St John, Y. Guan, Y. Kim, B. D. Etz, S. Kim and R. S. Paton, Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules, *Sci. Data*, 2020, **7**, 1–6.
32 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, *Gaussian 09, Revision A.02*, Gaussian, Inc., Wallingford CT, 2016.
33 Q. Zhao, S. M. Vaddadi, M. Woulfe, L. A. Ogunfowora, S. S. Garimella, O. Isayev and B. M. Savoie, *Sci. Data*, 2023, **10**, 145.
34 S. Choi, *Nat. Commun.*, 2023, **14**, 1168.
35 More references for DFA-NN are written in Ref_of_DFA_simulating_NN.txt.