PCCP

Neural Network Atomistic Potentials for Global Energy Minima Search in Carbon Clusters

Journal:	*Physical Chemistry Chemical Physics*
Manuscript ID	CP-ART-05-2023-002317.R1
Article Type:	Paper
Date Submitted by the Author:	12-Jul-2023
Complete List of Authors:	Tkachenko, Nikolay; Utah State University, Chemistry and Biochemistry Tkachenko, Anastasiia; Utah State University, Computer Science Department Nebjen, Benjamin; Los Alamos National Laboratory, Center for Integrated Nanotechnologies Tretiak, Sergei; Los Alamos National Laboratory, Theoretical Division Boldyrev, Alexander; Utah State University, Chemistry and Biochemistry

Neural Network Atomistic Potentials for Global Energy Minima Search in Carbon Clusters

Nikolay V. Tkachenko,[a,†,*] Anastasiia A. Tkachenko,[b,†,*] Benjamin Nebgen,[c] Sergei Tretiak,[c] Alexander I. Boldyrev[a,*]

[a] *Department of Chemistry and Biochemistry, Utah State University, Logan, Utah 84322-0300, USA;*
[b] *Department of Computer Science, Utah State University, Logan, Utah 84322-0300, USA;*
[c] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA;*
[†] *These authors contributed equally.*

ABSTRACT

The global energy optimization problem is an acute and important problem in chemistry. It is crucial to know the geometry of the lowest energy isomer (global minimum, GM) of a given compound for the evaluation of its chemical and physical properties. This problem is especially relevant for atomic clusters. Due to the exponential growth of the number of local minima geometries with the increase of the number of atoms in the cluster, it is important to find a computationally efficient and reliable method to navigate the energy landscape and locate a true global minima structure. Newly developed neural network (NN) atomistic potentials offer a numerically efficient and relatively accurate approach for molecular structure optimization. An important question that needs to be answered is: "Can NN potentials, trained on a given set, represent the potential energy surface (PES) of a neighboring domain?". In this work, we tested the applicability of the ANI-1ccx and ANI-nr NN atomistic potentials for the global minima optimization of carbon clusters C_n (n = 3-10). We showed that with the introduction of the cluster connectivity restriction and consequent DFT or *ab initio* calculations, the ANI-1ccx and ANI-nr can be considered as a robust PES pre-sampler that can capture the GM structure even for large clusters such as C_{20}.

INTRODUCTION

The study of atomic clusters attracts attention of many theoreticians and experimentalists. Due to the diversity in structure, physical and chemical properties, clusters can be used for various applications including chemical reactions [1, 2], catalysis [3,4], optical responses [5], energy storage [6], magnetic materials [7], to name a few. This diversity poses a fundamental problem for chemists: how the characteristics of a cluster would evolve into the corresponding bulk system as the number of constituent atoms increases [8, 9]. Importantly, the prediction of size dependent cluster properties involves the non-trivial task of global structure optimization. Most atomistic clusters are considered to be rigid, which is roughly characterized by the few (1–3) isomers in the 3 kcal/mol window with respect to the lowest in energy isomer a.k.a. global minimum (GM) [10]. In this case, the GM structure can properly model the physical properties of the considered cluster at ambient conditions with a given number of atoms.

Procedures for finding a GM include sampling of the potential energy surface (PES), applying optimization methods to initial structures to locate local minima and selecting a point with the lowest energy as a presumed GM. There are several fundamental problems that come with the GM structure search. First, the growth of cluster size leads to the exponential growth of the number of local minima [11-15], thus decreasing drastically the likelihood of finding true GM. An exponential increase of the number of samples or the use of heuristic algorithms may alleviate this problem. The second problem is related to the energy calculation of a given structure on the global PES. Accurate techniques, such as state-of-art *ab initio* approaches or hybrid/double-hybrid DFT functionals, become intractable for large stoichiometries due to their computational cost. In turn, semi-empirical or LDA-DFT methods may not accurately evaluate the PES for the given stoichiometry, which may result in the wrong GM assignment.

Machine Learning Interatomic Potentials (MLIP), such as Neural Network models (NN), are rising as an alternative technique providing an attractive compromise between computational cost and accuracy in predicting the energies of structures within the domain of interest [16-24]. This is due to the vast number of parameters in NN, which enables high flexibility in describing complex PESs providing a properly chosen training dataset. In fact, NN

models trained on relatively small-sized systems (tens of atoms) are extensible and may provide highly accurate estimates for large-sized systems not included in the training set but within its domain [25–27]. However, NNs have their limitations, specifically the preparation of a training dataset – a step considered the most important but time-consuming. The larger the studied system, the more structural configurations it has. Consequently, extended dataset is needed to obtain a good statistical sampling of molecular conformations and thus the PES. This significantly complicates the training of a NN, since each datapoint (i.e., structure-energy correspondence) needs to be distinct and meaningful, whereas the dataset needs to exhaustively span the relevant phase space. However, recent work had successfully developed machine learning approaches toward "smart" reduction of a training dataset for MLIP models and its automated generation [28, 29].

In this work, we explore capabilities of the ANI-1ccx NN potential [30] originally trained on organic molecules containing H, C, N and O atoms and their various configurations by utilizing the correspondent ANI-1ccx dataset [31] and ANI-nr potential that employs active learning (AL) combined with a nanoreactor (NR) sampler for dataset generation [32]. Specifically, ANI-1ccx and ANI-nr potentials are applied to predicting potential energy of an adjacent domain – carbon clusters – without re-training of the NNs. We juxtapose its performance with DFT and semi-empirical methods on carbon cluster of size from three up to ten atoms. We focus on applicability of the NN potentials for tasks related to a GM search in the neighboring domain: Can it predict the relative global minima structure? How do generated local minima structures differ from those obtained with DFT and semi-empirical methods? And how well is the NN potential applicable for PES sampling, particularly larger clusters? The findings of this study shed light on the broad application of MLIP models trained on compounds containing multiple chemical elements beyond the target domain.

METHODS

For each carbon cluster of size *n* (where *n* is a number of carbon atoms in a cluster), random structures (dataset) are prepared using the Coalescence Kick (CK) algorithm [33-35]. The CK algorithm is a stochastic search procedure that is designed to seek most of the minima on the potential energy surface. The size of the dataset is set according to the formula 5×2^n to account for an exponential growth of the number of local minimum structures with cluster size. The initial optimization of structures prepared by the CK algorithm is performed by employing four methods: ANI-1ccx and ANI-nr NN models [30, 32], semi-empirical parametric method PM7 [36], and PBE0 hybrid density functional theory (DFT) method with well-balanced DZ Karlsruhe def2-SVP basis set [37]. The ANI-1ccx network featuring ANI-1 architecture [27], was originally trained to the ANI-1x DFT dataset (~5M data points) then refined on Coupled Cluster data on small organic molecules (~500k data points) containing H, C, N and O atoms using transfer learning. The ANI-1ccx dataset is a small subset of the ANI-1x dataset generated using active learning and recomputed with a correlation and basis set extrapolated CCSD(T) methodology specifically for building networks via transfer learning. The ANI-nr was trained similarly to ANI models on a data generated with NR sampler using reference DFT (B3LYP) data (~500k data points) [32]. The ANI-nr effectively covers all the chemical space covered by ANI-1x and ANI-1ccx. We have been using an ensemble of eight networks for ANI-1ccx or ANI-nr and limit the number of optimization steps to 200.

To compare performance of four methods on the original dataset, we also examine subsets of optimized structures: connected structures and connected structures with distinct energies. Connectivity of a structure is analyzed *via* Fiedler eigenvalue. Each molecule can be represented as a graph, where vertices are atoms and edges represent bonds. We assume that two atoms are not bonded if the distance between them is more than 2Å (being a rough sealing of the largest known carbon-carbon bond length [38]). The graph is connected if the second-smallest Fiedler eigenvalue of a Laplacian matrix $\boldsymbol{L}$ for the graph is greater than zero. Matrix $\boldsymbol{L}$ equals to $\boldsymbol{D} - \boldsymbol{A}$, where $\boldsymbol{D}$ is the degree matrix of the graph, a diagonal matrix with elements representing the number of edges attached to each vertex, and $\boldsymbol{A}$ is the adjacency matrix, elements of which indicates presence of an edge between two vertices. Subset of connected structures are thinned out by their energy to collect structures with energies differing by more than 10^{-4} Hartree, which we call connected structures with distinct energies. Structures with carbon-carbon distances closer than 0.8 Å were considered as artifact geometries and were not included in the consequent optimization described below.

The first ten converged connected structures with distinct energy and shape given by ANI-1ccx, ANI-nr, PM7, and PBE0/def2-SVP are then reoptimized at PBE0/def2-TZVPP level of theory. More accurate single-point coupled-cluster calculations (CCSD(T)/def2-TZVPP) are finally performed using the refined geometries to reliably establish the relative energy ordering. These steps were taken to ensure that we have a reliable understanding of the bottom of the potential energy surface landscape for these stoichiometries. All DFT, semi-empirical, and coupled-cluster calculations mentioned above are performed via Gaussian 16 (Rev B.01) package [39].

In addition, for the cyclic and chain-like structures, the domain-based local pair natural orbital coupled-cluster theory (DLPNO-CCSD(T)) [40] is employed as implemented in ORCA 5.0.3 software [41–43]. The DLPNO-CCSD(T) energies are then extrapolated to the complete basis set (CBS) limit using the three-point extrapolation [44] based on cc-pVDZ, cc-pVTZ, and cc-pVQZ [45] basis sets and corresponding auxiliary basis sets [46,47]. The MP2/def2-TZVPP level of theory is used to reoptimize cyclic and chain-like structures obtained from PBE0/def2-SVP calculations. No significant changes in geometry are observed. Further, the chemical bonding analysis is performed via the AdNDP 2.0 program [48,49].

RESULTS

Performance of the selected methods.

We have sampled eight datasets for each size *n* of carbon clusters *via* the CK algorithm. The sizes of the clusters are chosen from 3 to 10 carbon atoms. The choice of cluster sizes is driven by the idea to balance computational time and representativeness of the data because, according to the previous studies, the global minimum structure for singlet-state carbon clusters alternates from a chain shape into ring geometry for odd and even number of carbon atoms, respectively [50]. Each structure in each dataset is optimized with ANI-1ccx model, semi-empirical method PM7, and hybrid density-functional PBE0. We compare the resulting set of structures in terms of the time to obtain them by each method, proportion of connected structures within the set, diversity of structure shapes and energies.

In Figure 1(a), we depict an average representative time each method took to optimize a cluster of size *n*. As was expected, the ANI-1ccx and ANI-1nr neural network models show the fastest performance with only 1.6 and 2.1 seconds of the average run time (CPU-time) for the largest stoichiometry, while that of the PM7 and of DFT methods are approximately 5 and 60 minutes, respectively. As expected, DFT, in contrast to considered NN potentials and PM7, demonstrates drastic increase in computational cost with the growing cluster size. In that sense, the MLIP and semiempirical methods are computationally much more affordable. This trend can also be confirmed if we compare the average CPU time required for one optimization step for the considered methods (Figure 1(b)). Predictably, the obtained results illustrate that with the growth of the cluster size, the computational cost of ANI-1ccx, ANI-nr and PM7 calculations do not increase significantly, while computational cost of DFT calculations grows polynomially (the polynomial scaling of PM simulations is expected to be observed for much larger molecular sizes). In particular, for 10 carbon atoms, the DFT calculation requires about a minute per optimization step, while MLIPs and PM7 run time stays at level of 0.01 and 5 seconds, respectively. We would like to emphasize that the computational cost plays an important role in the global minima optimization problem.

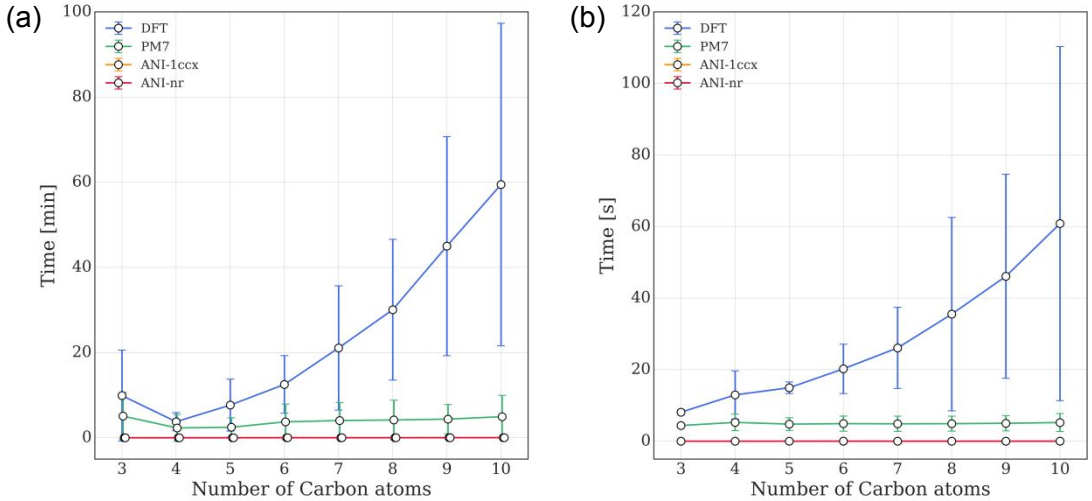

Figure 1. *(a) Average CPU (core) time required to optimize a structure within 200 steps. All calculations are performed using one similar computational node. (b) average CPU time required to compute one step of the optimization.*

Analysis of optimized structures.

The trend of global minima structures of singlet-state carbon clusters has been discussed before [50]. Previous studies found an interesting behavior with the increase of the number of atoms. Thus, for the even number of carbon atoms (4, 6, 8, 10) the global minima structure is cyclic, while for the odd number of atoms (3, 5, 7, 9) the GMs are chain-like structures. The optimization of the CK-generated random structures at PBE0/def2-SVP level of theory is able to capture this trend, showing the correct GM structures (Figure 3). These structures will be used as a ground truth for the comparison of the results obtained with the ANI-1ccx, ANI-nr, and PM7 methods.

The CK algorithm generates completely random atom arrangements, however, ensuring their initial proximity facilitating formation of chemical bonds. Thus, the considered CK structures are significantly off the training domain of both ANI-1ccx and ANI-nr: the method starts many of their optimizations in the unexplored and badly represented regions of PES.

To estimate how ANI-1ccx and ANI-nr behave at various regions of PES, we calculate the potential energy curve of the bond dissociation process of carbon dimer and compared it with other ab-initio methods (Figure 3). Despite the correct representation of the near-equilibrium region (1-1.5 Å), two problematic regions appear on that graph. When the distance between two carbon atoms is less than 0.8 Å, due to lack of training in this region, the MLIP models unphysically decrease the system's total energy, making it favorable to collapse all nuclei of the cluster in one point. Nevertheless, the barrier for that process is relatively high, and the structure will be optimized to this artifact geometry only if the initial positions of carbon

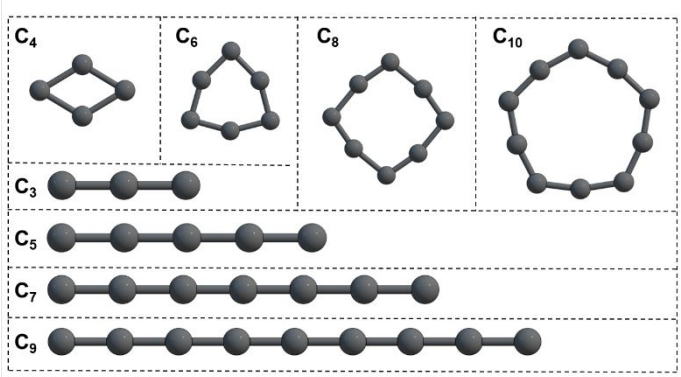

Figure 2. *Global minimum structures of C_n (n = 3-10) clusters obtained using the PBE0 hybrid DFT functional.*

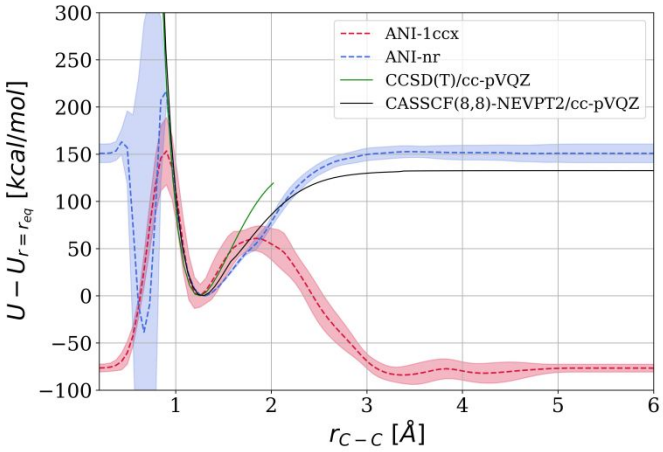

Figure 3. *Potential energy surface of carbon dimer calculated at various levels of theory. Shaded regions given for MLIPs show the ensemble standard deviation.*

atoms are too close to each other. The second and most significant problematic region is the dissociation of two atoms to infinity. We can observe that ANI-1ccx wrongly over-stabilizes the dissociated structure, making it energetically more beneficial than the bonded structure. Moreover, the energy barrier for that process is rather small, so we can expect many disconnected structures to appear in the optimized set. This is attributed to training of ANI-1ccx near-equilibrium structures of molecular systems. Additionally, we would like to emphasize that the performance of ANI-1x is very similar to ANI-1ccx as illustrated in Figure S1. In turn, ANI-nr, as reactive force field by design, correctly represents the energy trend toward the dissociation of two carbon atoms, making this process energetically unfavorable. We would like to emphasize, that issues associated with MLIPs, such as the favorability to produce fragmented or collapsed structures, are not limited to molecules outside the training set. Even for molecules within the training set, ANI-1ccx and ANI-nr exhibit similar bond-breaking curves as exemplified with C-H bond in CH_4 molecule [32].

As a consequence of this, we find many optimized by ANI-1ccx carbon clusters with disconnected geometry. Predictably, ANI-1ccx ascribes those structures very low energy (lower than the connected ones). As an illustration of this phenomenon, we plot the lowest energy geometries proposed by ANI-1ccx method in Figure S2. Here the global minimum can be described as a small chain of carbon atoms surrounded by the isolated atoms. This trend of the GM structure is persistent in all C_n (n = 3-10) series. We note that such structures are not energetically favorable

in reality, and isolated atoms tend to lower their energy by forming chemical bonds with other atoms and agglomerates. In addition, we found several low-lying in energy optimized structures with collapsed carbon atoms. As discussed above, this behavior is expected as illustrated in the Figure 3. Thus, the single MLIP model such as ANI-1ccx method, cannot be considered as a standalone method for global minima optimization of carbon clusters and its blindfolded applications are not recommended. Subsequently, additional restrictions should be introduced to utilize the data obtained by this MLIP model.

The problematic regions of the PES that are sampled by ANI-1ccx and ANI-nr can pose challenges in implementing certain global minimum optimization techniques like the basin-hopping algorithm. As a result, heuristic sampling algorithms will generate fragmented or collapsed geometries (which was observed for the basin-hopping optimization of C_{10} cluster with ANI-1ccx and ANI-nr potentials). In contrast, the CK algorithm, which employs random sampling, allows for the imposition of additional restrictions (such as proximity or connectivity) across the entire PES landscape, enabling the examination of chemically meaningful regions. Furthermore, incorrect energy ordering provided by MLIPs can lead to erroneous GM identification without comprehensive landscape sampling. In turn, CK's extensive sampling of a large number of structures increases the likelihood of identifying the true global minimum geometry, even if the NN provides inaccurate energy ordering of the isomers.

Computing a portion of connected structures within the original dataset suggests that it declines for ANI-1ccx model down to 0.37, whereas the respective DFT and PM7 values are staying around 1 (Figure 4(a)). As expected, ANI-nr produces a greater fraction of connected structures than ANI-1ccx declining down to the value of 0.92 for C_{10} clusters. The main reason of appearance on the disconnected structures during ANI-nr optimization is the over-stabilization of collapsed carbon agglomerates. As the result, global minima geometries produced by ANI-nr also look unphysical (Figure S3). However, it is quite easy to reject unphysical structures by introducing restrictions on the allowed distance between the nuclei, as well as by introducing a check for the structure's connectivity as discussed below.

Despite an increasing fraction of disconnected structures for ANI-1ccx method, on the exponentially increased dataset, all four methods demonstrate an exponential growth of the number of connected structures with distinct energies (n_{cE}): these fit the function $n_{cE} = c \cdot 2^n$ (Figure 4(b)). However, on a fixed-size dataset, we see a plateau at size 7 for the ANI-1ccx model (Figure S4). Thus, without the dataset exponential increase, the ANI-1ccx method eagerly converges structures of carbon clusters to the disconnected ones, which affects its n_{cE}.

In spite of inability of ANI-1ccx to describe structure related to dissociation limits by construction and inability of both ANI-1ccx and ANI-nr describe too close proximity of carbon atoms, by screening out the unphysical structures with unbound or too close atoms, MLIPs may be useful as a pre-optimizers for determining minimum energy structures. As described in the Methods section, the graph

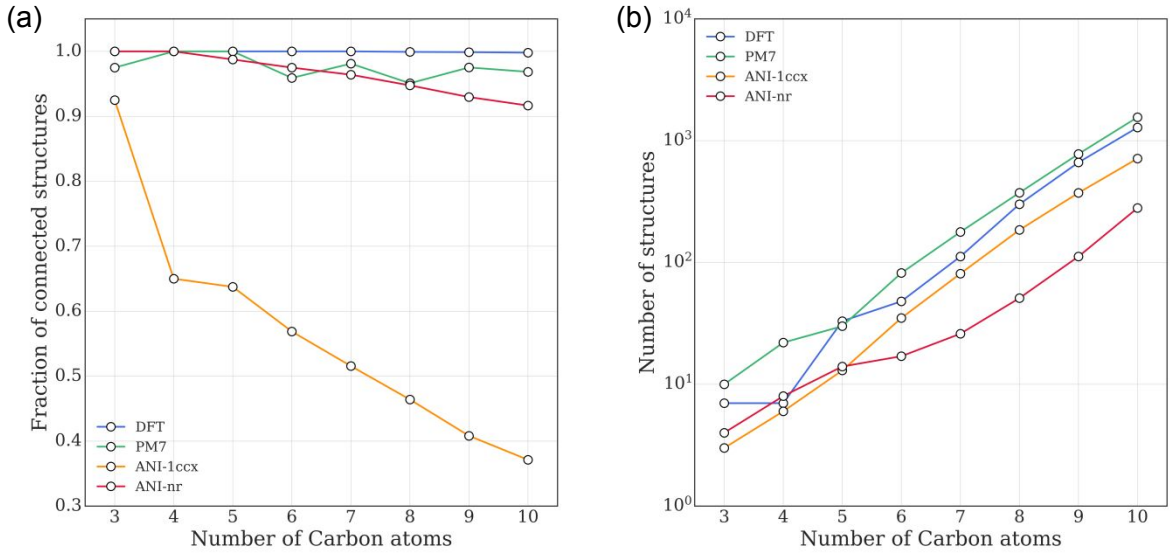

Figure 4. *(a) Fraction of connected structures in the optimized dataset. (b) Number of connected structures with distinct energies for each method. The Y-axis is in log form.*

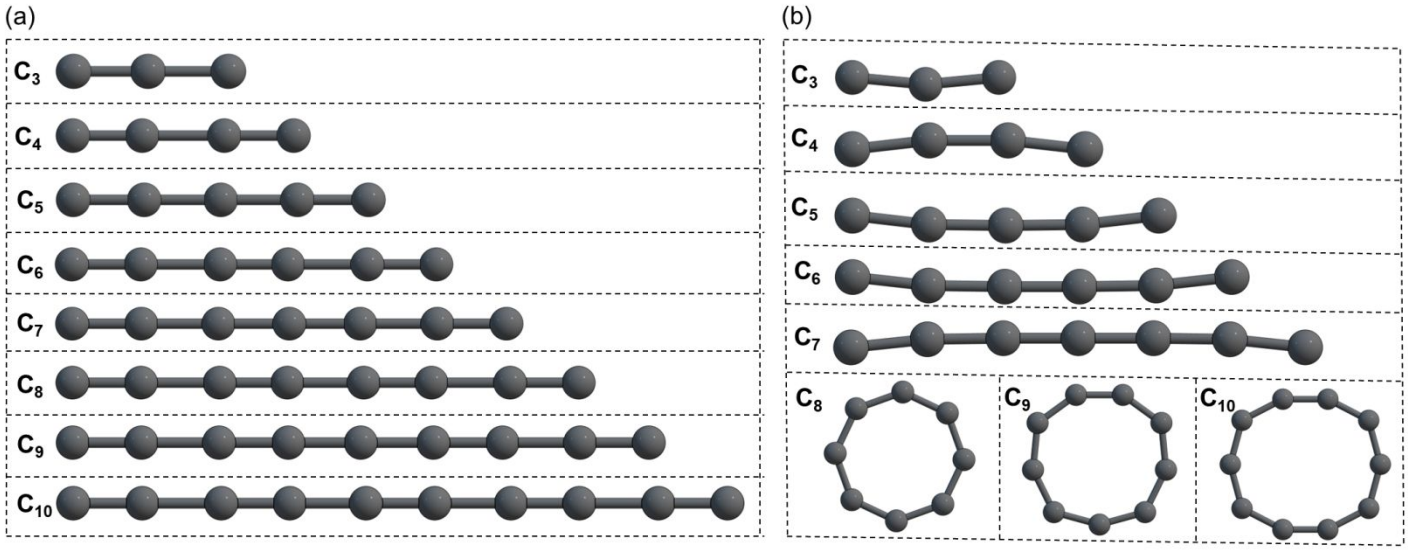

Figure 5. *Global minimum connected structures of C_n (n = 3-10) clusters according to the ANI-1ccx (a) and ANI-nr (b) methods.*

representation of the carbon cluster and the Fiedler eigenvalue were used for that purpose. With this constraint, the connected GM structures generated with the ANI-1ccx and ANI-nr methods are shown in the Figure 5. As we can see, the ANI-1ccx potential is able to capture GMs for every odd stoichiometry (Figure 5(a)). However, no cyclic structures are found to have the lowest energy among the connected structures. Interestingly, the ANI-1ccx training set being composed of saturated organic molecules shows itself in the C_4 and C_6 predictions, where these would be close to correct for fully saturated systems but are wildly wrong for pure carbon systems. Interestingly, the PM7 method is also unable to capture the alternating trend of GM structures for carbon cluster, showing the chain-like structures as a GM for all stoichiometries except for C_{10}, for which the cyclic structure was assigned to be the GM structure (Figure S5). Similarly, ANI-nr assigns the lowest energy to linear structures for clusters up to 7 atoms. Starting from 8 carbon atoms the cyclic structures are more preferable (Figure 5(b)). Interestingly, ANI-1ccx and ANI-nr optimize C-C bond lengths for linear structures to slightly different values. Specifically, ANI-1ccx alternates bonds significantly varying them from ~1.2 to ~1.4 Å within one structure, while ANI-nr bond distances are close to the DFT-obtained structures and do not alternate a lot staying within ~1.3 Å, which is a consequence of the more structurally-rich reactive dataset encompassing out-of-equilibrium conditions.

The inaccurate assignment of the GM for some of the stoichiometries by the ANI-1ccx or ANI-nr does not make the methods useless for the GM optimization problem. The method itself can be used as an efficient pre-sampler of the structures followed by the more accurate DFT and *ab initio* methods that will provide the more accurate energy ordering and geometry of the isomers. For this purpose, it is important that the structural motif of the global minimum be present among the low-energy isomers. Thus, we can reduce the expensive DFT and *ab initio* calculations by applying them only to those few low-energy isomers obtained by fast ANI approach. To this end, we check 10 lowest connected isomers for each stoichiometry for the presence of the cyclic and linear structures. We indeed find them among the 10 lowest connected isomers proposed by ANI-1ccx and ANI-nr methods (Figure S6, S12-S17). This promising result illustrates the usefulness of ANI models as a pre-sampler method for a GM optimization problem of carbon clusters. We note that 10 lowest connected structures constitute a very small percentage of all connected structures obtained by ANI models. Thus, for C_{10} clusters, 10 lowest connected isomers correspond only to ~1.5% of all generated distinct connected structures. Further reoptimization of the lowest 10 isomers (at PBE0/def2-TZVPP level of theory) lead to the correct assignment of the global minima structures showed in Figure 3. The relative energies (at CCSD(T)/def2-TZVPP//PBE0/def2-TZVPP level of theory) and corresponding geometries of all low-lying

isomers for all four methods are given in the Supporting information file (Figure S7-S11). Additionally, single-point energies for the cyclic and chain-like structures are calculated for comparison at the DLPNO-CCSD(T) level of theory with the three-point energy extrapolation to the CBS (Table S1).

Chemical bonding analysis and discussion on the stability of the linear and cyclic isomers.

To understand why the global minimum structure for singlet-state carbon clusters alternates from a chain shape into ring geometry we next perform a chemical bonding analysis using the AdNDP algorithm. The AdNDP is an electron-localization technique that partitions the natural density of the system and reproduces the most occupied spatially localized bonding elements. The approach extends ideas of Weinhold's Natural Bond Orbitals (NBO) analysis [51]. The crucial advantage of the AdDNP approach is that it allows us to represent a chemical bonding pattern not only in terms of localized Lewis bonding elements (lone-pairs and two-center two-electron bonds (2c-2e)) but also in terms of delocalized bonding elements over several atoms (nc–2e bonds, where n>2) related to aromaticity and antiaromaticity concepts.

Based on the AdDNP results, a similar chemical bonding pattern is found for all odd-numbered linear isomers (C_n, n = 3, 5, 7, 9). Thus, the two s-type lone-pairs could be localized on terminal carbon atoms with occupation numbers (ON) 1.97 |e|, indicating that the terminal carbon atoms are unsaturated. The carbon-carbon interactions can be described with the localized 2c-2e σ-bonds. Due to the high $D_{\infty h}$ symmetry, the 2c-2e π-bonds between carbon atoms could not be localized, and the π-bonding interactions are manifested via completely delocalized nc-2e π-bonds with ON = 2.00 |e|. This bonding picture agrees with almost degenerate C-C distance in linear isomers. The chemical bonding of the linear C_3 cluster is shown in Figure 6(a), while the bonding for other chain-like isomers could be found in the Supporting Information file (Figure S18-S20). The question remains: why even-numbered linear isomers (C_n, n = 4, 6, 8, 10) are energetically unstable? The reason for this instability can be explained by examining the molecular orbitals of even-numbered isomers. Unlike the odd-numbered linear isomers, where four electrons occupied two degenerate Π_u or Π_g HOMO orbitals

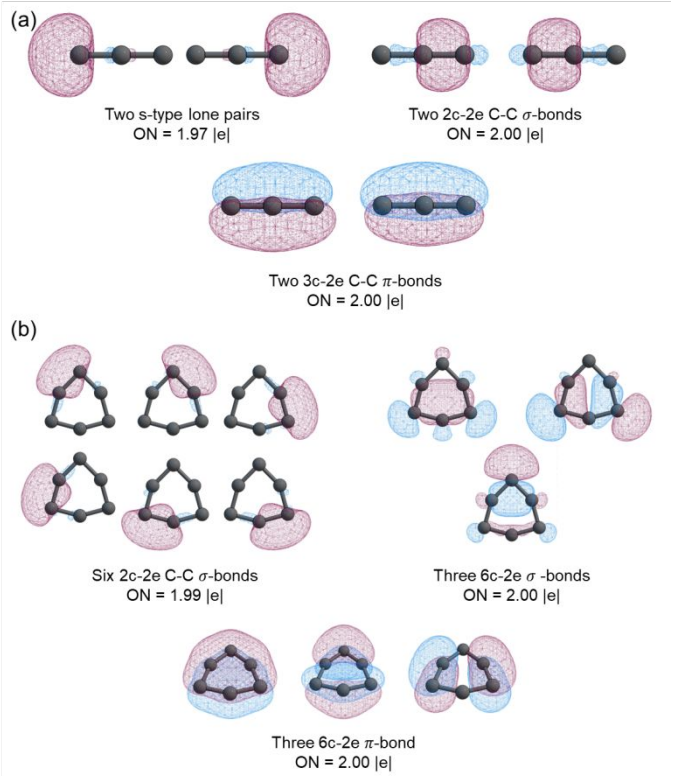

Figure 6. *(a) Chemical bonding pattern of the linear isomer of C_3 cluster obtained from AdNDP analysis; (b) Chemical bonding pattern of the linear isomer of C6 cluster obtained from AdNDP analysis.*

(resulting in $^1\Sigma_g^+$ electronic state), even-numbered linear isomers possess only two electrons on doubly degenerate Π_u or Π_g orbitals. Thus, according to Hund's rule, the $^3\Sigma_g^-$ electronic state will be more energetically preferable for even-numbered isomers than a singlet $^1\Sigma_g^+$ state, which can also be observed from the DFT calculations (Table S2).

Electron delocalization also plays a crucial role in the cyclic isomers of carbon clusters. Thus, the C_6 and C_{10} isomers are found to be doubly σ- and π- aromatic possessing 4n+2 (n = 1, 2) electrons in the delocalized σ- and π- circuits (Figure 6(b) and Figure S23) [52,53]. Although the clusters are doubly-aromatic they are not fully symmetric and belong to D_{3h} and D_{5h} point symmetry groups. The reason of the lower symmetry is the presence of second order Jahn-Teller effects in D_{6h} and D_{10h} structures [54,55].

According to the electron-counting rule, we expect that C_4 and C_8 clusters are doubly anti-aromatic with 4n (n = 1, 2) electrons delocalized in σ- and π- circuits, respectively. However, the C_4 cluster is found to be doubly aromatic instead. Due to the presence of s-type lone-pairs on two

carbon atoms and four 2c-2e σ-bonds, the remaining four electrons form 4c-2e σ-bond and 4c-2e π-bond responsible for doubly-aromatic properties (Figure S21). The C_8 cluster is indeed found to be anti-aromatic (Figure S22), which also can be observed from a significant alternation of carbon-carbon bond lengths in the structure (the shortest bond is 1.25 Å, while the longest bond is 1.38 Å). The question of why odd-numbered cyclic isomers are energetically less preferable[56] can be answered by analyzing the bonding patterns of even-numbered isomers. We can observe that the addition of one carbon atom into the cycle will add one additional 2c-2e σ-bonds, one electron to the sigma-delocalized circuit and one electron to the pi-delocalized circuit. Since the odd number of electrons on degenerate orbitals created Jahn-Teller instabilities, we conclude that planar cyclic odd-numbered isomers will be energetically unfavorable.

As a result, we observe that electron delocalization plays a crucial role in pure carbon clusters and controls the geometry and stability of different isomers.

Structural transition from cycles and chains to fullerenes and planes upon the cluster growth.

To further explore applicability of ANI models to large cluster stoichiometries, we perform a test calculation of C_{20} clusters. The CK generated dataset consists of 10240 random structures and resulted in ~1100 and ~8800 connected optimized structures for ANI-1ccx and ANI-nr, respectively (thus, a fraction of connected structures for ANI-1ccx drops down to ~11%). Interestingly, this number of structures were enough for ANI-nr to capture cyclic structure (one of the lowest isomers for this stoichiometry). However, still this number of random structures is not sufficient to capture the cage and bowl-like isomers, which considered to be one of the lowest isomers of C_{20} stoichiometry [57]. Thus, we further artificially introduce them into the dataset, to see if those structures will appear in a set of low-lying isomers. We indeed found that the artificially introduced structures are among the lowest in energy isomers (Figure 7). Though the relative energy and the ordering of those isomers by ANI models is not accurate compared to the reference [57], the further optimization can refine the data obtained by a cheap MLIP approach.

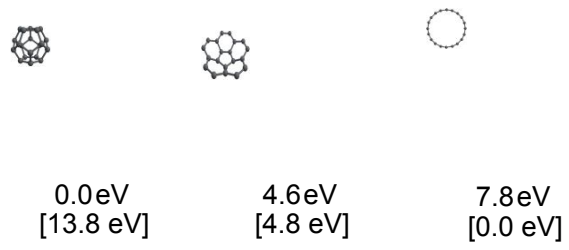

Figure 7. *Low energy structures of C_{20} cluster according to the ANI-1ccx and ANI-nr methods with the relative energies in eV. Energies for ANI-nr are given in square brackets.*

CONCLUSIONS

In summary, we evaluate the ANI-1ccx and ANI-nr neural network atomistic potentials' suitability for the task of the global minima optimization problem of carbon clusters. While ANI-1ccx, trained only on molecular systems, poorly describes systems in the dissociation limit, and both ANI-nr and ANI-1ccx improperly model the structures with short carbon-carbon distance, those MLIPs still can be used to significantly reduce the computational cost *via* combinations of several approaches. Specifically, by culling down stoichiometries of a series of carbon clusters C_n (n = 3-10) with the CK algorithm by imposing the cluster connectivity and short distance restriction criteria, we show that the NN potentials accelerate the low-energy conformer search and reduce the computational cost of the global minima optimization problem. Notably, such clusters of pure carbon atoms were not part of the training set used to build the ANI-1ccx and ANI-nr models. Although the energy ordering of isomers is not accurate for some of the stoichiometries, the correct GM structural motifs are present within the several lowest in energy connected isomers. Using the larger C_{20} cluster stoichiometry, we show, that ANI models capture even nontrivial carbon cluster GM transformations such as transition from cycles and chains to fullerenes and planes upon the cluster growth. We believe that this work provides useful insights to the research community and facilitates future use of ML interatomic potentials in global minima optimization problems. The use of CK to generate systems effectively representing extremely unphysical atomic geometries, may be treated as an adversarial attack on MLIPs [58-61].

Finally, the ability for NN potentials to identify minimum energy carbon structures is of critical importance for a variety of chemical and materials applications. Recent work on large scale MD simulations of carbon systems [62, 63] is largely driven by the interest in understanding carbon cluster formation post combustion. Yet, despite these potentials having passed many large-scale tests, such as the prediction of the carbon phase diagram, perhaps the

addition of low energy carbon clusters would provide an even more stringent test for such potentials.

ACKNOWLEDGMENTS

The support and resources from the Centre for High Performance Computing at the University of Utah are gratefully acknowledged. A.I.B. acknowledges financial support from the R. Gaurth Hansen Professorship fund. The work at Los Alamos National Laboratory (LANL) was supported by the LANL Directed Research and Development Funds (LDRD) and performed in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT), a US Department of Energy (DOE) Office of Science user facility at LANL.

DATA AND SOFTWARE AVAILABILITY

All raw data generated in this work as well as all optimized structures are given in the supplementary files of this manuscript. Examples of input files and example script for MLIP calculations are available in the supplementary files and through the GitHub source (https://github.com/AnaTkachenko/MLIP_for_CarbonClusters).

REFERENCES

(1) Sultan, S.; Tiwari, J. N.; Singh, A. N.; Zhumagali, S.; Ha, M.; Myung, C. W.; Thangavel, P.; Kim, K. S. Single Atoms and Clusters Based Nanomaterials for Hydrogen Evolution, Oxygen Evolution Reactions, and Full Water Splitting. *Adv. Energy Mater.* **2019**, *9* (22), 1900624.

(2) An, S.; Zhang, G.; Wang, T.; Zhang, W; Li, K.; Song, C.; Miller, J. T.; Miao, S.; Wang, J.; Guo, X. High-Density Ultra-small Clusters and Single-Atom Fe Sites Embedded in Graphitic Carbon Nitride (g-C_3N_4) for Highly Efficient Catalytic Advanced Oxidation Processes. *ACS Nano* **2018**, *12* (9), 9441–9450.

(3) Yan, B.; Wu, Q.; Cen, J.; Timoshenko, J; Frenkel, A. I.; Su, D.; Chen, X.; Parise, J. B.; Stach, E.; Orlov, A.; Chen, J. G. Highly active subnanometer Rh clusters derived from Rh-doped $SrTiO_3$ for CO_2 reduction. *Appl. Catal. B* **2018**, *237*, 1003–1011.

(4) Zhou, Y.; Xie, Z.; Jiang, J.; Wang, J.; Song, X.; He, Q.; Ding, W.; Wei, Z. Lattice-confined Ru clusters with high CO tolerance and activity for the hydrogen oxidation reaction. *Nat. Catal.* **2020**, *3*, 454–462.

(5) Wang, Z.; Zhao, G.; Yan, W.; Wu, K.; Wang, F.; Li, Q.; Zhang, J. Tin Metal Cluster Compounds as New Third-Order Nonlinear Optical Materials by Computational Study. *J. Phys. Chem. Lett.* **2021**, *12* (31), 7537–7544.

(6) VanGelder, L. E.; Kosswattaarachchi, A. M.; Forrestel, P. L.; Cook, T. R.; Matson, E. M. Polyoxovanadate-Alkoxide Clusters As Multi-Electron Charge Carriers For Symmetric Non-Aqueous Redox Flow Batteries. *Chem. Sci.* **2018**, *9*, 1692–1699.

(7) Shin, T.-H.; Choi, Y.; Kim, S.; Cheon, J. Recent Advances In Magnetic Nanoparticle-Based Multi-Modal Imaging. *Chem. Soc. Rev.* **2015**, *44*, 4501–4516.

(8) de Heer, W. A. The Physics Of Simple Metal Clusters: Experimental Aspects And Simple Models. *Rev. Mod. Phys.* **1993**, *65*, 611.

(9) Jena, P.; Sun, Q. Super Atomic Clusters: Design Rules And Potential For Building Blocks Of Materials. *Chem. Rev.* **2018**, *118* (11), 5755–5870.

(10) Bursch, M.; Mewes, J.-M.; Hansen, A.; Grimme, S. Best-Practice DFT Protocols for Basic Molecular Computational Chemistry. *Angew. Chem. Int. Ed.* **2022**, *61*, e202205735.

(11) Heiles, S; Johnston, R. L. Global Optimization of Clusters Using Electronic Structure Methods. *Int. J. Quantum Chem.* **2013**, *113*, 2091–2109.

(12) Stillinger, F. H.; Weber, T. A. Hidden structure in liquids. *Phys. Rev. A* **1982**, *25*, 978.

(13) Stillinger, F. H.; Weber, T. A. Packing Structures and Transitions in Liquids and Solids. *Science* **1984**, *225*, 983–989.

(14) Wales, D. J.; Doye, J. P. K. Stationary points and dynamics in high-dimensional systems. *J. Chem. Phys.* **2003**, *119*, 12409–12416.

(15) Doye, J. P. K.; Wales, D. J. Saddle points and dynamics of Lennard-Jones clusters, solids, and supercooled liquids. *J. Chem. Phys.* **2002**, *116*, 3777–3788.

(16) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(17) Han, J.; Zhang, L; Car, R; E, W. Deep Potential: A General Representation of a Many-Body Potential Energy Surface. *Commun. Comput. Phys.* **2018**, *23*, 629–639.

(18) Kobayashi, R.; Giofré, D.; Junge, T.; Ceriotti, M.; Curtin, W.A. Neural network potential for Al-Mg-Si alloys. *Phys. Rev. Materials*. **2017**, *1*, 053604.

(19) Yao, K.; Herr, J.E.; Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017**, *146*, 014106.

(20) Manzhos, S.; Dawes, R.; Carrington, T. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem*. **2015**, *115*, 1012-1020.

(21) Shao, K.; Chen, J.; Zhao, Z.; Zhang, D. H. Communication: Fitting potential energy surfaces with fundamental invariant neural network. *J. Chem. Phys*. **2016**, *145*, 071101.

(22) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv*. **2017**, *3*, e160301.

(23) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyrea, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261-2269.

(24) Thaler, S.; Zavadlav, J. Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. *Nat. Commun*. **2021**, *12*, 6884.

(25) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(26) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178-184.

(27) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(28) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun*. **2018**, *9*, 3887.

(29) Smith, J. S.; Nebgen, B. T.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys*. **2018**, *148*, 241733.

(30) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential through Transfer Learning. *Nat. Commun.* **2019**, *10* (1), 2903.

(31) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.

(32) Zhang, S.; Makos, M. Z.; Jadrich, R. B.; Kraka, E.; Barros, K. M.; Nebgen, B. T.; Tretiak, S.; Isayev, O.; Lubbers, N.; Messerly, R. A.; Smith, J. S. Exploring the frontiers of chemistry with a general reactive machine learning potential. *ChemRxiv*. **2022**, 10.26434/chemrxiv-2022-15ct6-v2.

(33) Saunders, M. Stochastic Search for Isomers on A Quantum Mechanical Surface. *J. Comput. Chem.* **2004**, *25*, 621–626.

(34) Averkiev, B. B. Geometry and Electronic Structure of Doped Clusters via The Coalescence Kick Method. Doctoral Dissertation, Utah State University, Logan, UT, 2009.

(35) Sergeeva, A. P.; Averkiev, B. B.; Zhai, H. J.; Boldyrev, A. I.; Wang, L. S. All-boron analogues of aromatic hydrocarbons: B_{17}^- and B_{18}^-. *J. Chem. Phys*. **2011**, *134*, 224304.

(36) Stewart, J. J. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19* (1), 1–32.

(37) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys*. **2005**, *7*, 3297–3305.

(38) Li, J.; Pang, R.; Li, Z.; Lai, G.; Xiao, X.-Q.; Müller, T. Exceptionally Long C−C Single Bonds in Diamino-o-carborane as Induced by Negative Hyperconjugation. *Angew. Chem. Int .Ed.* **2019**, *58* (5), 1397–1401.

(39) 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; 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 Jr., 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 Rev. B.01, **2016**.

(40) Sparta, M.; Retegan, M.; Pinski, P.; Riplinger, C.; Becker, U.; Neese, F. Multilevel Approaches within the Local Pair Natural Orbital Framework. *J. Chem. Theory Comput.* **2017**, *13*, 3198–3207.

(41) Neese, F. The ORCA Program System. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(42) Neese, F. Software Update: The ORCA Program System, Version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8* (1), e1327.

(43) Neese, F. Software Update: The ORCA Program System—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

(44) Neese, F.; Hansen, A.; Liakos, D. G. Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *J. Chem. Phys.*, **2009**, *131*, 064103.

(45) Dunning Jr., T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.

(46) Weigend, F. Hartree–Fock exchange fitting basis sets for H to Rn. *J. Comput. Chem.* **2008**, *29*, 167–175.

(47) Weigend, F.; Kohn, A.; Hattig, C. Efficient use of the correlation consistent basis sets in resolution of the identity MP2 calculations. *J. Chem. Phys.* **2002**, *116*, 3175.

(48) Zubarev, D. Y.; Boldyrev, A. I. Developing Paradigms of Chemical Bonding: Adaptive Natural Density Partitioning. *Phys. Chem. Chem. Phys.* **2008**, *10*, 5207– 5217.

(49) Tkachenko, N. V.; Boldyrev, A. I. Chemical Bonding Analysis of Excited States Using the Adaptive Natural Density Partitioning Method. *Phys. Chem. Chem. Phys.* **2019**, *21*, 9590– 9596.

(50) Van Orden, A.; Saykally, R. J.; Small Carbon Clusters: Spectroscopy, Structure, and Energetics. *Chem. Rev.* **1998**, *98* (6), 2313–2358.

(51) Weinhold, F.; Landis, C. R. Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective. Cambridge University Press, Cambridge, UK, 2005.

(52) Wodrich, M. D.; Corminboeuf, C.; Park, S. S.; Schleyer, P. v. R. Double Aromaticity in Monocyclic Carbon, Boron, an dBorocarbon RingsBase don Magnetic Criteria. *Chem. Eur. J.* **2007**, *13*, 4582–4593.

(53) Baryshnikov, G. V.; Valiev, R. R.; Nasibullin, R. T.; Sundholm, D.; Kurten, T.; Ågren H. Aromaticity of Even-Number Cyclo[n]carbons (n = 6−100). *J. Phys. Chem. A* **2020**, *124*, 10849−10855.

(54) Saito, M.; Okamoto, Y. Second-order Jahn-Teller effect on carbon 4N+2 member ring clusters. *Phys. Rev. B* **1999** *60*, 8939−8942.

(55) Hong, I.; Ahn, J.; Shin, H.; Bae, H.; Lee, H.; Benali, A.; Kwon, Y. Competition between Hückel's Rule and Jahn–Teller Distortion in Small Carbon Rings: A Quantum Monte Carlo Study. *J. Phys. Chem. A* **2020**, *124*, 3636–3640.

(56) Hoffmann, R. A chemical and theoretical way to look at bonding on surfaces. *Rev. Mod. Phys.* 1988, *60*, 601.

(57) Jin, Y.; Perera, A.; Lotrich, V. F.; Bartlett, R. J. Coupled Cluster Geometries and Energies of C_{20}

Carbon Cluster Isomers – a New Benchmark Study. *Chem. Phys. Lett.* **2015**, *629*, 76–80.

(58) Schwalbe-Koda, D.; Tan, A. R.; Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun*. **2021**, *12*, 5104.

(59) Goodfellow, I. J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint*. **2014**, arXiv:1412.6572.

(60) Szegedy, C.; Zaremba, W.; Sutskever, I., Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv preprint*. **2013**, arXiv:1312.6199.

(61) Biggio, B.; Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit*. **2018**, *84*, 317-331.

(62) Lindsey, R. K.; Fried, L. E.; Goldman, N. ChIMES: A Force Matched Potential with Explicit Three-Body Interactions for Molten Carbon. *J. Chem. Theory Comput.* **2017**, *13*, 6222−6229.

(63) Willman, J. T.; Nguyen-Cong, K.; Williams, A. S.; Belonoshko, A. B.; Moore, S. G.; Thompson, A. P.; Wood, M. A.; Oleynik, I. I. Machine learning interatomic potential for simulations of carbon at extreme conditions. *Phys. Rev. B.* **2022**, *106*, L180101.