PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d4cp03238a

Deep learning-driven prediction of chemical addition patterns for carboncones and fullerenes†

Zhengda Li, Xuyang Chen and Yang Wang *

Carboncones and fullerenes are exemplary π-conjugated carbon nanomaterials with unsaturated, positively curved surfaces, enabling the attachment of atoms or functional groups to enhance their physicochemical properties. However, predicting and understanding the addition patterns in functionalized carboncones and fullerenes are extremely challenging due to the formidable complexity of the regioselectivity exhibited in the adducts. Existing predictive models fall short in systems where the carbon molecular framework undergoes severe distortion upon high degrees of addition. Here, we propose an incremental deep learning approach to predict regioselectivity in the hydrogenation of carboncones and chlorination of fullerenes. Utilizing exclusively graph-based features, our deep neural network (DNN) models rely solely on atomic connectivity, without requiring 3D molecular coordinates as input or their iterative optimization. This advantage inherently avoids the risk of obtaining chemically unreasonable optimized structures, enabling the handling of highly distorted adducts. The DNN models allow us to study regioselectivity in hydrogenated carboncones of $C_{70}H_{20}$ and $C_{62}H_{16}$, accommodating up to at least 40 and 30 additional H atoms, respectively. Our approach also correctly predicts experimental addition patterns in $C_{50}Cl_{10}$ and $C_{76}Cl_n$ (n = 18, 24, and 28), whereas in the latter cases all other known methods have been proven unsuccessful. Compared to our previously developed topology-based models, the DNN's superior predictive power and generalization ability make it a promising tool for investigating complex addition patterns in similar chemical systems.

Received 17th August 2024,
Accepted 13th December 2024

DOI: 10.1039/d4cp03238a

rsc.li/pccp

1 Introduction

Carboncones[1–3] and fullerenes[4–6] represent two typical examples of positively curved π conjugated carbon nanomaterials.[7–9] While fullerenes have long existed in outer space[10,11] and have been constantly produced in the lab for almost four decades,[12] carboncones (see Fig. 1a and b) have only recently been synthesized with atomic precision.[2,3] Owing to their unique structures and properties, these pure carbon or carbon-rich molecules may find versatile potential applications across a wide range of fields, including materials science[13–15] and biomedicine.[16,17] More intriguingly, the vast area of their molecular surface with delocalized π electrons allows chemists to attach additional atoms or groups to carboncones and fullerenes to

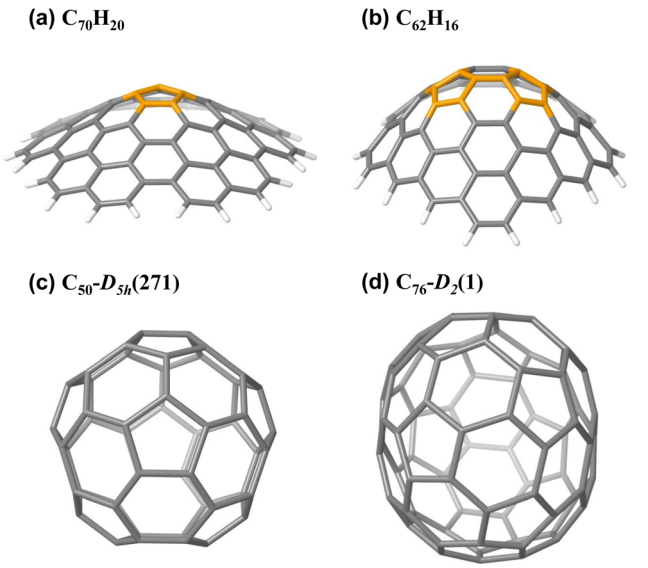


Fig. 1 Molecular structures of carboncones (a) $C_{70}H_{20}$ and (b) $C_{62}H_{16}$, and fullerenes (c) C_{50}-D_{5h}(271) and (d) C_{76}-D_2(1). Pentagonal rings in carboncones are highlighted in orange.

School of Chemistry and Chemical Engineering, Yangzhou University, Yangzhou, Jiangsu 225002, China. E-mail: yangwang@yzu.edu.cn

† Electronic supplementary information (ESI) available: Cutoff energies of RE_{xTB} and RE_{DNN}; PCA feature dimensionality reduction; distortion of the carbon framework upon addition; performance of DNN models on the test set; comparison of performance between DNN and other models; lowest-energy addition patterns of hydrogenated carboncones; and optimized Cartesian coordinates and absolute energies for lowest-energy structures. See DOI: **https://doi.org/10.1039/d4cp03238a**

afford functionalized derivatives that can have improved properties, such as better solubility in water,[2,18,19] improved biocompatibility,[20,21] and supramolecular assembly ability.[22–24]

Among the great variety of possible exohedral functionalizations of fullerenes and carboncones, the simplest and most prototypical one is the radical addition of a number of similar atoms. As the most extensively studied example, a multitude of chlorinated fullerenes have been synthesized and identified experimentally with diverse forms of carbon cages and with a broad range of numbers of Cl atoms.[25–31] On the other hand, carboncones can be regarded as a special kind of polycyclic aromatic hydrocarbon (PAH), whose hydrogenated species are not only interesting in materials science[32] but have recognized astrophysical significance.[33,34] Additionally, hydrogenation of carboncones or fullerenes has relevance to the pursuit for promising carbon-based materials for hydrogen storage applications.[35–37] A fundamental problem about these addition reactions of fullerenes or carboncones is the great complexity shown in addition patterns. For instance, even for the medium-sized, highly-symmetric buckminsterfullerene C_{60} attached with only 8 Cl atoms, there are over 20 million possible ways to distribute the addend atoms on the carbon cage.[38] However, only a few regioisomers of $C_{60}Cl_8$ have actually been synthesized in the lab.[39–41] Moreover, these addition patterns observed in experiments are usually hard to interpret using known organic chemistry rules for traditional addition reactions.[38] To make things even worse, the total number of adduct regioisomers scales factorially with the number of addends.[25] Therefore, the formidable number of possible regioisomers and the erratic addition patterns usually seen in final adducts make it impossible to brute-force compute (using, *e.g.*, semiempirical or DFT calculations) all enumerated structures to find the most stable ones in order to understand or predict the experimentally observed regioselectivity.

To tackle this problem, one of us (ref. 25 and 38) came up with a simple theoretical model, named the exohedral fullerene stabilization index (XSI), that can correctly predict many experimentally synthesized fullerene adducts with all kinds of cage forms and different numbers of hydrogen/halogen atoms or CF_3 groups. The XSI model incorporates the three essential factors governing the relative stability of adducts: the π delocalization, the strain induced by adjacent pentagonal rings in the cage framework, and the steric hindrance between addends situated at adjacent sites. The stabilization or destabilization energies caused by these effects are all quantified on a topological basis, *i.e.*, requiring solely the information of the connectivity between atoms. For instance, the π energy is evaluated using the simple Hückel molecular orbital (HMO) theory,[42–45] in which one only needs to diagonalize the adjacency matrix between carbon atoms. Hence, neither sophisticated quantum chemical calculations nor iterative geometry optimization is demanded to evaluate the XSI values for a measure of relative energies of regioisomers. To efficiently reduce the number of isomers to explore, the model combines a stepwise addition algorithm[46–54] based on the assumption that stable regioisomers of a higher degree of addition are derived from the stable ones of a lower degree of addition, which has been supported by experimental evidence.[55] In that experiment, multistage mass spectrometry confirmed the existence of the pristine cage of C_{74} in the gas phase, suggesting that the final $C_{74}Cl_{10}$ adduct was formed *via* progressive addition of Cl atoms.[55] The observed pristine C_{74} cage is not a stable isomer that can exist under normal conditions, but it was indeed produced in graphite arc-discharge.[55] Hence, the extreme synthetic conditions play a foundational role in determining the experimentally observed isomers, which usually correspond to the thermodynamically most favorable structures. The XSI model is also successful in predicting the structures of hydrogenated PAHs,[56,57] as well as the stability of Diels–Alder adducts of fullerenes. Later on, the XSI model was extended to the ext-XSI model[58] so as to predict hydrogenation patterns of carboncone $C_{70}H_{20}$ (Fig. 1a) with up to 12 additional H atoms. Additional topology-based terms are added to the ingredients to account for the steric repulsion between the added H atoms and the H atoms originally from the carboncone molecule. A detailed explanation of the XSI and ext-XSI models is presented in Section 2.6. However, these models are no longer applicable to adducts with a high addend-to-carbon ratio that causes substantial deformation of the carbon framework, since the simple HMO method becomes invalid if the π conjugated system undergoes severe distortion.[38]

With the recent burgeoning advances in artificial intelligence,[59,60] machine learning[61,62] offers an alternative approach to the prediction of addition patterns for fullerenes. Rooted in the SchNet[63] deep neural network, Liu *et al.*[64] recently devised an automated tool for predicting low-energy adduct structures of fullerenes. In combination with the stepwise addition procedure, the 3D structure of each candidate regioisomer is optimized by a neural network potential (NNP) that was trained on the datasets obtained from large-scale semiempirical or DFT calculations. This method was shown to be successful for radical adducts of different fullerene cages with various kinds of addends. However, as each of the candidate structures needs to be optimized starting from initial quasi-equilibrium geometries, there are chances that the NNP may lead to chemically unreasonable structures with damage or rearrangement in the carbon framework, migration of addends, *etc.*[64] Since such invalid structures are detected by a checker module and thus ruled out from the candidate list, it can be problematic when the fullerene cage is highly distorted upon addition and too many invalid structures may be generated for the algorithm to proceed. Consequently, the prediction procedure would terminate early before reaching the desired number of addends. Therefore, this NNP-based approach failed to handle[64] a series of experimentally observed fullerene chlorides with relatively high chlorination degrees, $C_{76}Cl_n$ (n = 24, 28, and 30).[65,66]

In this article, we propose a deep learning approach to study the addition patterns in both hydrogenated carboncones and chlorinated fullerenes. Instead of performing geometry optimizations as in the previous NNP-based method,[64] our deep neural network (DNN) models deal only with the

topology-based information of adducts, thus requiring solely the connectivity between atoms, as in the XSI and ext-XSI models.[25,38,58] Starting from the adducts with 2 and 4 addends, the DNN models are progressively trained on the dataset that expands steadily with the increasing number of addends guided by the stepwise procedure. Such an incremental learning strategy[67,68] has allowed us to predict with satisfactory precision the relative stability of regioisomers with rather high degrees of addition. Specifically, we have managed to investigate the regioselectivity in heavily hydrogenated carboncones, $C_{70}H_{20}$ and $C_{62}H_{16}$ (see Fig. 1b), with up to 30–40 added H atoms. To further demonstrate the generalizability and predictive power of our approach, we have applied it to the chlorination of two distinct fullerene cages, C_{50}-$D_{5h}(271)$ (Fig. 1c) and C_{76}-$D_2(1)$ (Fig. 1d). The prediction has not only correctly determined the experimental structure of $C_{50}Cl_{10}$,[69,70] but also been successful for the long-standing challenge posed by the above-mentioned series of chlorinated C_{76}.[25,38,64] We have made correct predictions of the experimentally synthesized structures[65,66] of $C_{76}Cl_n$ with n = 18, 24, and 28. Moreover, we have discovered a never reported regioisomer of $C_{76}Cl_{28}$, which is energetically even more stable than the experimental isomer.[66]

2 Methodology and computational methods

2.1 Workflow of a DNN model for predicting addition patterns

We employed a deep learning approach combined with quantum chemistry calculations to predict the regioselectivity in hydrogenation of carboncones $C_{70}H_{20}$ and $C_{62}H_{16}$, as well as that in the chlorination of fullerenes C_{50}-$D_{5h}(271)$ and C_{76}-$D_2(1)$. For convenience, we utilize the chemical formula, $C_{70}H_{20+n}$, to denote the $C_{70}H_{20}$ carboncone attached with n additional H atoms.[56–58] Similar notations apply to the adducts of hydrogenated $C_{62}H_{16}$. In essence, the lowest-energy regioisomers of a varying degree of addition of H or Cl atoms are progressively predicted by means of incremental learning.[67,68] In other words, each time when the number of addends increases, the DNN model is trained with the new knowledge provided by the grown dataset. To outline the DNN training procedure (see Fig. 2), in the following we take the prediction of addition patterns in $C_{70}H_{20+n}$ as an illustrative example. We note that for all carboncones and fullerenes considered in this study, n is always an even number to ensure that the resultant adducts have a stable, closed-shell electronic configuration.

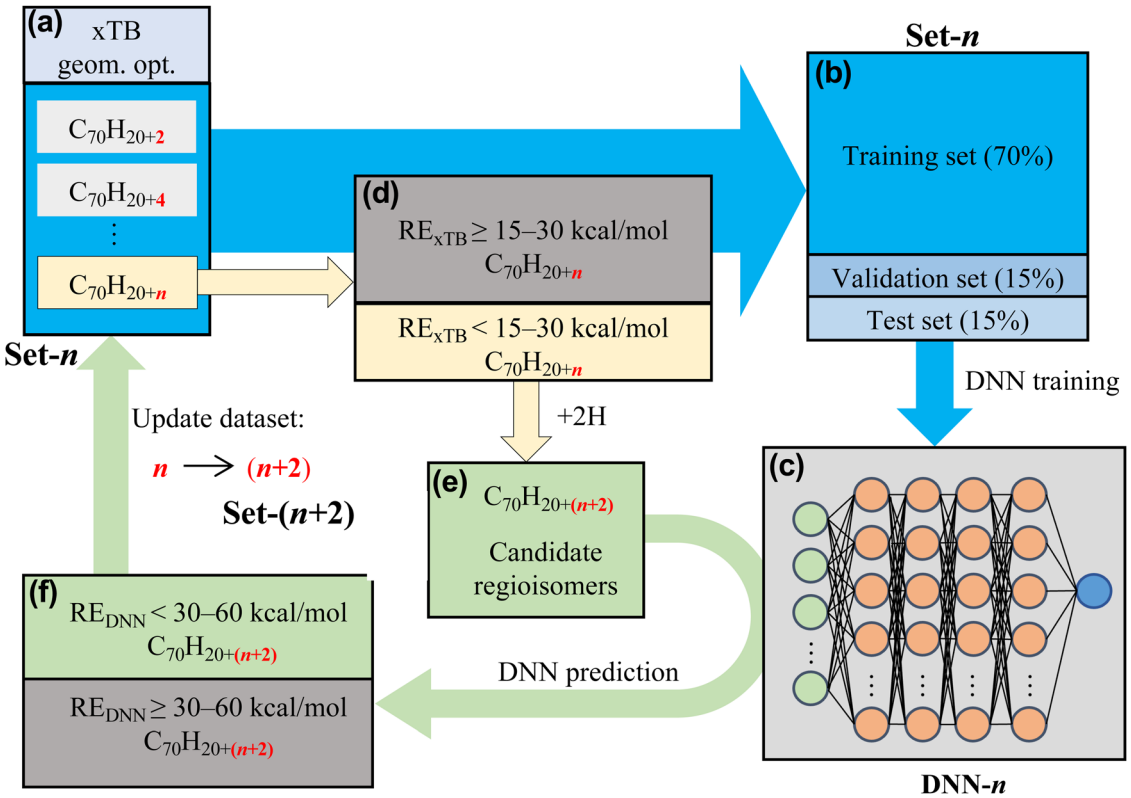


Fig. 2 Workflow of the incremental DNN model for predicting the lowest-energy regioisomers of $C_{70}H_{20+n}$ (n = 2, 4, 6,..., 40). (a) Dataset Set-n including all xTB-optimized regioisomers from $C_{70}H_{20+2}$ to $C_{70}H_{20+n}$. (b) Division of Set-n into the training, validation and test sets. (c) Training of the DNN model, DNN-n, using Set-n. (d) Screening of lower-energy $C_{70}H_{20+n}$ regioisomers using a selection criterion in terms of their relative xTB energies, RE_{xTB}. (e) Generation of $C_{70}H_{20+(n+2)}$ regioisomers by adding two H atoms to all possible sites on each of the $C_{70}H_{20+n}$ regioisomers selected in (d). (f) Prediction of relative energies by model DNN-n (RE_{DNN}) for all regioisomers of $C_{70}H_{20+(n+2)}$ generated in (e). On the basis of RE_{DNN}, lower-energy $C_{70}H_{20+(n+2)}$ regioisomers are selected and added to Set-n after xTB geometry optimization, resulting in the expanded dataset Set-(n + 2).

We start with the initial dataset, named Set-4, including all possible regioisomers of $C_{70}H_{20+2}$ and $C_{70}H_{20+4}$, which are obtained by graph-based addition pattern enumeration,[25,38,58] followed by full geometry optimization at the GFN2-xTB (hereafter xTB for short) level.[71,72] To prevent overfitting and ensure model generalization, Set-4 is randomly divided into training (70%), validation (15%) and test (15%) sets, as shown in Fig. 2b. The DNN model (called DNN-4) is then trained and evaluated on Set-4 (see Fig. 2c). On the other hand, we adopted the stepwise addition algorithm[25,38] to generate all candidate regioisomers of a higher degree of hydrogenation. To be more specific, we select the lower-energy regioisomers of $C_{70}H_{20+4}$ (with a relative xTB energy, RE_{xTB}, less than 30 kcal mol^{-1}) as seed structures (from Fig. 2a to d) and subsequently add two H atoms to each of the seed structures to construct all nonequivalent $C_{70}H_{20+6}$ regioisomers (from Fig. 2d to e). Applying model DNN-4 to all these generated $C_{70}H_{20+6}$ structures, we obtain their predicted relative isomer energies, RE_{DNN}, based on which we select the lower-energy candidates for $C_{70}H_{20+6}$ (Fig. 2f) for the following geometry optimizations at the xTB level. All xTB-computed $C_{70}H_{20+6}$ regioisomers are added to Set-4, thereby expanding the dataset to Set-6, as illustrated in Fig. 2 where f returns to a. The above procedure is repeated multiple times as n increases progressively from 6 to the desired number of addends (n_{max}) (*e.g.*, n_{max} = 40 in this particular case of $C_{70}H_{20}$). Eventually, we attain the final dataset Set-n_{max} and the final model DNN-(n_{max} − 2), which is capable of predicting the relative energies for all of the $C_{70}H_{20+n}$ regioisomers with n ranging from 2 all the way up to n_{max}.

2.2 Dataset construction for DNN learning

We still take the hydrogenation of $C_{70}H_{20}$ as an example to detail the construction of datasets. As defined in the preceding subsection, Set-n consists of all generated regioisomers of $C_{70}H_{20+k}$ with k = 2, 4, 6,..., n. While for Set-4 we enumerate all possible regioisomers, for larger datasets, Set-n ($n \geq 6$), we only include the lower-energy candidate regioisomers from all generated $C_{70}H_{20+n}$ structures using the stepwise addition algorithm,[25,38] as instructed in the preceding subsection. In the latter case ($n \geq 6$), two cutoff energy criteria are employed in the stepwise addition process to generate candidate regioisomers of $C_{70}H_{20+n}$. First, as shown in Fig. 2d, we only consider all $C_{70}H_{20+(n-2)}$ seed regioisomers with a relative xTB energy (RE_{xTB}) lower than a cutoff energy to generate eligible $C_{70}H_{20+n}$ structures by adding two more H atoms to each of the seed molecules. We use different values for this cutoff energy depending on different degrees of addition, n, so as to ensure that there are sufficient numbers (at least *ca.* 100 000) of the generated nonequivalent regioisomers of $C_{70}H_{20+n}$. These values are specified in Table S1 in the ESI.† The second cutoff criterion is based on the relative isomer energy predicted by model DNN-(n − 2), RE_{DNN}, to select lower-energy candidates of $C_{70}H_{20+n}$ for the subsequent xTB calculations, yielding the expanded dataset Set-n (see Fig. 2f). Different values of cutoff RE_{DNN} are adopted for different values of n for $C_{70}H_{20+n}$ (see Table S1 in the ESI†).

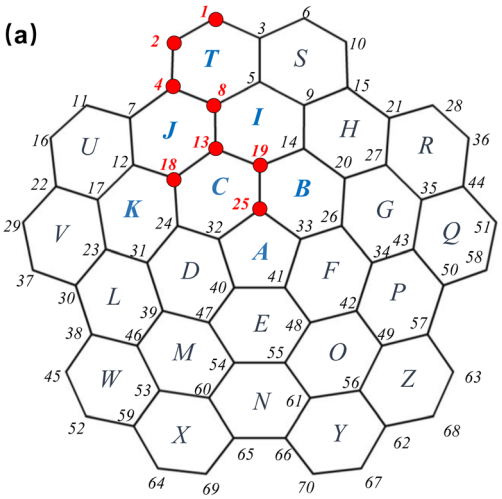
The datasets are constructed in a similar way for hydrogenated carboncone $C_{62}H_{16+n}$ and chlorinated fullerenes $C_{50}Cl_n$ and $C_{76}Cl_n$, except that the cutoff values for selection of adduct regioisomers are somewhat different. The cutoff values for RE_{xTB} to screen the seed regioisomers of $C_{62}H_{16+(n-2)}$ are listed in Table S2 in the ESI.† A universal cutoff RE_{xTB} of 10 kcal mol^{-1} is used for $C_{50}Cl_n$ and $C_{76}Cl_n$. The only exception is that we have increased the cutoff RE_{xTB} to 15 kcal mol^{-1} for generating $C_{76}Cl_{20}$, $C_{76}Cl_{32}$, and $C_{76}Cl_{34}$ structures so as to have enough (over 70) seed structures; otherwise, there are, *e.g.*, only 8 seed isomers of $C_{76}Cl_{18}$ left after screening with a cutoff RE_{xTB} of 10 kcal mol^{-1}. As for the screening of DNN-predicted regioisomers, instead of using a cutoff RE_{DNN} to select candidate structures for further xTB calculations and expansion of the dataset to Set-n, we simply choose the top 10 000 lowest-RE_{DNN} regioisomers for $C_{62}H_{16+n}$ and chlorinated fullerenes.

It is necessary to mention that all datasets have ruled out all open-shell addition patterns.[25,38,58] In order to reduce computational efforts, in the selection of seed structures and in the xTB calculations of newly generated isomers, we exclude all symmetrically equivalent regioisomers (by using a topology-based algorithm exploiting breadth-first search-based[73] canonical labeling[25]). Once the xTB calculations are completed, we generate all other symmetrically equivalent regioisomers as new instances from each of the unique, nonequivalent structures, by running over all equivalent canonical paths.[25] All these generated symmetrically equivalent instances are labeled with the same xTB relative energy as that of the unique regioisomer. The final dataset for hydrogenated $C_{70}H_{20}$ and $C_{62}H_{16}$ is Set-40 and Set-30, respectively, containing *ca.* 3 400 000 and 560 000 regioisomers, respectively. The final dataset for chlorinated C_{50} and C_{76} is Set-10 and Set-34, respectively, corresponding to the maximum numbers of Cl atoms in the experimentally synthesized chlorides, resulting in a total of *ca.* 810 000 and 680 000 regioisomers, respectively.

2.3 Feature engineering

All features designed in the current study are graph-based, *i.e.*, relying solely on the connectivity between carbon atoms and the indication of whether or not a carbon site is occupied by an addend. As a result, there is no need to perform any geometry optimization to extract the features, thus allowing our DNN models to make predictions without being fed with any optimized molecular structures. In contrast, previous machine learning models[63,64] for predicting addition patterns for fullerenes requires as an input a quasi-equilibrium geometry for each molecule, followed by iterative optimization of atomic positions. Evidently, the topology-based features in our approach are advantageous from the viewpoints of data storage and data processing efficiency. It also circumvents generating chemically nonsensical molecular structures in case the geometry optimizations are problematic. As usual, we will take the example of hydrogenated $C_{70}H_{20}$ to explain in detail the feature engineering process.

The first 70 features are one-hot encoded, depending on whether or not the hydrogenation takes place in the



Feature	1	2	3	4	…	8	…	70
Atom index	1	2	3	4	…	8	…	70
Value	1	1	0	1	…	1	…	0

Feature	71	72	73	74	…	80	…	96
Ring index	A	B	C	D	…	I	…	Z
Value	1	2	4	0	…	3	…	0

Fig. 3 (a) Labeling of C atoms and rings in carboncone $C_{70}H_{20}$. Red circles indicate the added H atoms, exemplifying an addition pattern of $C_{70}H_{20+8}$. (b) Atom-based features 1 to 70, defined by whether the addition takes place at the corresponding C atom. (c) Ring-based features 71 to 96, defined as the number of added H atoms on the corresponding ring. The last row in (b) and (c) provides the values of these features for the particular addition pattern showcased in (a).

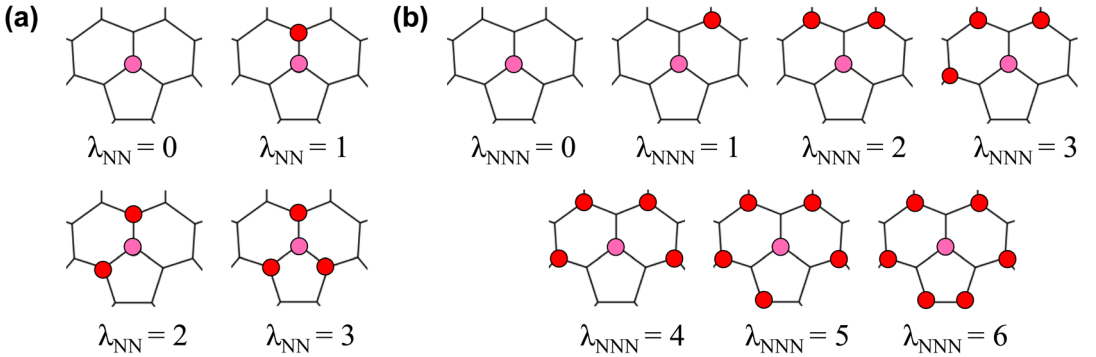
corresponding carbon site. The ith feature takes a value of 1 if the ith C atom (i = 1, 2,. . ., 70, see the atom labeling in Fig. 3a) is attached to an added H atom (red dots in Fig. 3a), and a value of 0 if not. The next 26 features are characterized by the distribution of addends on each of the 26 rings in $C_{70}H_{20}$. The value of the (70 + i)th feature is the number of added H atoms in the ith ring (i = 1, 2,. . ., 26); all rings are labeled alphabetically in Fig. 3a. Note that addends can be multiply counted across different rings. Fig. 3b and c provide the values of features 1–96 for the particular addition pattern showcased in Fig. 3a.

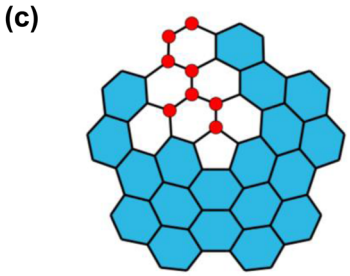
The following features are based on the numbers of addends at nearest neighbor (NN) positions and those at next nearest neighbor (NNN) positions with respect to each of the addition sites (*i.e.*, the C atoms attached with additional H atoms). For any given addition site (pink circle in Fig. 4a), let us define its λ_{NN} as the number of addends present on its NN sites. Accordingly, there are four possible values for λ_{NN}, *i.e.*, 0, 1, 2, and 3 (see the total number of red circles in each of the graphs in Fig. 4a), since there are at most three NN atoms for any carbon atom in carboncones and fullerenes. We can likewise introduce λ_{NNN} for a given addition site, defined as the number of addends located on its NNN positions. Fig. 4b shows all seven possible values for λ_{NNN} (0, 1, 2,. . ., 6) and the corresponding illustrations of addition situations. On the basis of λ_{NN} for each of the addition sites, feature 97 + i (i = 0–3) is defined as the number of addition sites that have a λ_{NN} value of i (see Fig. 4d). Similarly, we define feature 101 + i (i = 0–6) as the number of addition sites with λ_{NNN} = i (Fig. 4e). One can easily visually verify the specific values of features 97–107 in Fig. 4d and e for the particular addition pattern of $C_{70}H_{20+8}$ shown in Fig. 4c.

The last two features correspond to the number of intact rings (NR) that remain upon hydrogenation and the number of C atoms in the intact rings (NA).[58] Taking the addition pattern in Fig. 4c as an example, the addition of 8 H atoms (red circles) destroy 7 aromatic rings (shown in white) where these addends reside, leaving the remaining 19 rings (highlighted in blue) complete and fully π conjugated. Previous study[57,58] has revealed that NR and NA are strongly related to the relative stability of regioisomers. An addition pattern with a larger NR or NA is in general energetically more favorable. The values of features 108 and 109 are listed in Fig. 4f for the example regioisomer shown in Fig. 4c, in which one can count a total of 19 intact rings containing 62 C atoms.

In the case of chlorinated fullerenes, we employed modified definitions of the NR and NA features: only hexagonal rings rather than pentagonal rings are considered to count NR and NA. Such a modification brings significant improvement in predictive performance. For instance, when applying the modified definitions of NR and NA, the squared correlation coefficient (R^2) between the DNN-predicted and xTB-computed relative energies of $C_{50}Cl_8$ increases from 0.25 to 0.90, while the corresponding root mean square deviation (RMSD) drops from 10.7 kcal mol^{-1} to 4.0 kcal mol^{-1}. The reason why the modified NR and NA features work more appropriately for fullerene systems is probably due to the fact that a fullerene structure has a large number (12) of pentagonal rings, which have a nonaromatic nature and thus make no or even negative contribution to stability.

To summarize, we have designed all graph-based features that can be extracted solely from the connectivity between atoms in the adducts. The total number of features is size-dependent for the number of the first two sets of attributes equals, respectively, the number of C atoms and the number of rings in the whole molecule. All feature values are normalized using min–max scaling. In order to enhance the generalization ability of the model, we apply principal components analysis (PCA) to reduce the dimensionality of the features (see Section 2 of the ESI,† for details). As shown in Fig. S1–S4 in the ESI,† the number of features for training the final model for $C_{70}H_{20+n}$,





(d)

Feature	97	98	99	100
λ_{NN}	*0*	*1*	*2*	*3*
Value	**0**	**3**	**4**	**1**

(e)

Feature	101	102	103	104	105	106	107
λ_{NNN}	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Value	**0**	**3**	**4**	**1**	**0**	**0**	**0**

(f)

Feature	108	109
Meaning	NR	NA
Value	**19**	**62**

Fig. 4 (a) All four possible values of λ_{NN} for a given addition site (pink circle in the center). H atoms added at the NN sites are indicated by red circles. (b) Definition of λ_{NNN} with all seven possible values. The pink and red circles represent, respectively, the given addition site and the H atoms added at its NNN sites. (c) An addition pattern of $C_{70}H_{20+8}$. Intact rings upon addition are highlighted in blue. (d) Features 97–100 and (e) 101–107 are defined based on λ_{NN} and λ_{NNN}, respectively. (f) Features 108 and 109 are given by parameters NR and NA, respectively, which are defined by counting the numbers of intact rings and of the C atoms involved in them. The exemplified values of these features are listed for the particular addition pattern shown in (c).

$C_{62}H_{16+n}$, $C_{50}Cl_n$, and $C_{76}Cl_n$, is reduced from 109 to 79, from 99 to 70, from 90 to 59 and from 129 to 88, respectively.

2.4 Construction and training of DNN models

The construction of the neural network was realized using the Python programming language and the open-source TensorFlow[74–76] framework, which is well-suited and fine-tuned for large-scale machine learning. We primarily utilized the Keras module within TensorFlow for building and training DNN models. The DNN is composed of an input layer with the same number of neurons as the number of features, four hidden layers with consecutively 1024, 512, 256, and 256 neurons, and a one-neuron output layer.

The DNN training process is overviewed in Fig. 5. The original graph-based features are extracted from the structures in Set-*n* and then reduced to considerably fewer features through PCA. The xTB relative energy, RE_{xTB}, for each of the instances in Set-*n* serves as the target (label) for the DNN models. In the training of

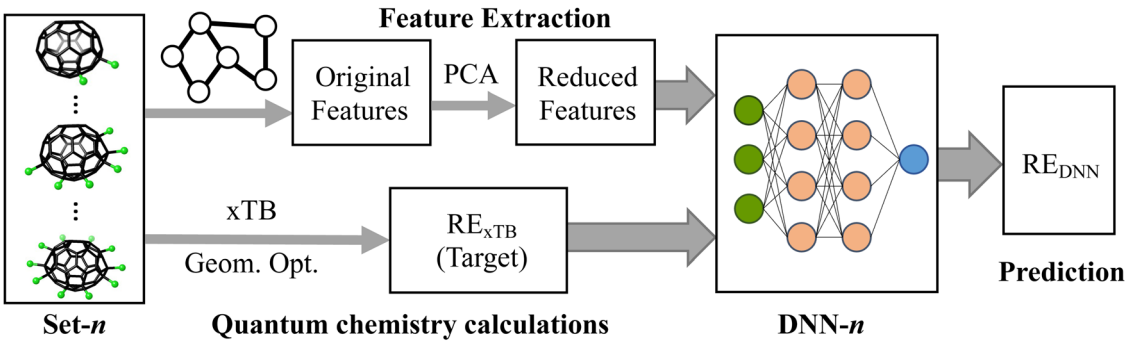


Fig. 5 Workflow of training the model DNN-*n* on dataset Set-*n* for the addition of *n* X addends to the parent carboncone or fullerene molecule.

model DNN-n, a rectified linear unit (ReLU) activation function[77] is used in the hidden layers due to its faster computation compared to many other activation functions and its ability to prevent gradient descent from getting stuck on plateaus.[76] For gradient descent to minimize the loss function (quadratic mean square error), we employed the commonly-used Adam[78] algorithm. To ensure sufficient learning on the training set, the number of epochs is set to 2000. The loss function is evaluated on the validation set after each epoch of training. Meanwhile, we utilize the EarlyStopping function in TensorFlow to prevent the overfitting resulting from excessive model training. If there is no improvement in the loss function over multiple epochs on the validation set, the model will terminate training. We eventually evaluate the performance of model DNN-n on the test set, as measured by R^2 and RMSD. Once the trained model DNN-n achieves a satisfactory performance, it will be deployed to predict the relative energies, RE_{DNN}, for the new adduct structures of a higher degree of addition.

2.5 Details of quantum chemical calculations

We performed full geometry optimizations for all generated regioisomers of the adducts using the extended semiempirical tight-binding GFN2-xTB method,[71,72] as implemented in the free software xtb (version 6.3.3).[71,79,80] Previous assessment[58] showed that the relative energies of hydrogenated $C_{70}H_{20}$ regioisomers predicted by the xTB method are in reasonable agreement with those obtained from the DFT calculations. As for chlorinated fullerenes, the xTB relative energies also correlate well with the DFT results for representative regioisomers of $C_{50}Cl_{10}$ and $C_{76}Cl_{28}$, as shown in Fig. S5 in the ESI.† These results suggest that the xTB method is a reasonable choice for prescreening the relatively stable regioisomers in this study. For each degree of hydrogenation or chlorination, we selected the top 20 lowest-energy regioisomers of adducts and refined the geometry optimization at the DFT level. By using the Gaussian 16 software,[81] we carried out M06-2X[82]/def2-SVP[83,84] calculations with the DFT total energy corrected by Grimme's DFT-D3 approach. This level of theory was previously used to study the hydrogenation of PAHs[85–87] and carboncone $C_{70}H_{20}$.[58] In all DFT calculations, vibrational frequency analyses were performed to verify the true minima on the potential energy surface and to compute zero-point energies (ZPEs).

2.6 XSI and ext-XSI prediction models

To demonstrate the advantageous performance of our DNN models over other known models in predicting the addition patterns, we also make predictions of relative isomer energies for hydrogenated carboncones and chlorinated fullerenes using, respectively, the ext-XSI and XSI models proposed in our previous studies.[25,38,56–58,88] We present below a brief description of the XSI and ext-XSI models.

On the basis of the simple HMO theory and the connectivity between atoms, the XSI model[25,38] evaluates the relative energy of a given regioisomer j of fullerene adducts C_mX_n as the following quantity:

$$\mathrm{XSI}_j = X_j + 0.2\cdot\mathrm{NAPP}_j + \gamma_\mathrm{X}\cdot\mathrm{NAX}_j. \quad (1)$$

By convention, XSI_j and its constituting terms are all expressed in units of -2β, where β is the resonance integral in the simple HMO theory. The first term, X_j, corresponds to the π energy and is given by

$$X_j = \sum_{k=1}^{m/2} \chi_k - \sum_{k=1}^{(m-n)/2} \chi_k^j, \quad (2)$$

where the first summation runs over the largest $m/2$ eigenvalues, $\{\chi_k\}$, of the adjacency matrix for the pristine fullerene cage, while the second summation is over the largest $(m-n)/2$ eigenvalues, $\{\chi_k^j\}$, of the adjacency matrix for the adduct regioisomer j. The second term in eqn (1) quantifies the cage strain effect following the pentagon adjacency penalty rule,[89,90] with NAPP_j representing the effective number of adjacent pentagon pairs (APPs) in the given regioisomer j and 0.2 being an empirical parameter for the energy penalty per APP. By running over all APPs, we count NAPP_j for each APP as $1 - n_\mathrm{X}$, where n_X is half of the number of addends X located at the interpentagonal C–C bond between each APP.[25,38] The last term in eqn (1) accounts for the energy penalty caused by steric hindrance; NAX_j is the number of adjacent addends in regioisomer j and the empirical parameter γ_X depends on the nature of the addends X. The γ_X values determined in the previous study[25] are 0.2230 and 0.2827 (in units of -2β) for H and Cl atoms, respectively.

When it comes to the prediction of addition patterns for hydrogenated carboncones, a more appropriate model is the ext-XSI model, which also takes into account the steric repulsions between the added H atoms and the H atoms that are originally possessed by the carboncone molecule.[58] Additionally, the ext-XSI model includes the stabilizing effect arising from the H addition to the pentagons in the carboncone framework. On the other hand, as there are no APPs in the carboncones considered in the present study, the penalty term introduced by APPs is excluded in the ext-XSI formalism. Therefore, for a given regioisomer j of hydrogenated carboncones, its relative energy can be predicted by the following expression

$$\text{ext-XSI}_j = X_j + \gamma_\mathrm{H}\mathrm{NAH}_j + \rho\mathrm{NRH}_j + \delta\mathrm{NARH}_j + \kappa\mathrm{NP}_j \quad (3)$$

where NAH_j is the number of H pairs added to adjacent sites, but we do not consider the H pairs added to adjacent rim sites of the carboncone.[58] NRH_j is the number of H atoms added to the rim positions, no matter if they are located at adjacent sites or not. NARH_j denotes the number of H pairs added to adjacent rim sites. One can find illustrative examples for counting these numbers in ref. 58. NP_j represents the number of H atoms added to the pentagon sites. The coefficients γ_H, ρ, δ, and κ in eqn (3) are empirical parameters that can be determined by a least-squares fit of ext-XSI to the DFT relative isomer energies.[58]

In this study, we apply two variants of the ext-XSI model, the 'fixed' model (named ext-XSI0) with the original parameterization (as in ref. 58) and the 'dynamically adapted' model (named ext-XSI-n) in which all parameters in eqn (3) are refitted using the data in Set-n (with a progressively increasing n).

3 Results and discussion

3.1 Hydrogenation of carboncones

3.1.1 Model performance. We first present the results for the hydrogenation of carboncone $C_{70}H_{20}$. Fig. 6 shows clearly that model DNN-n markedly outperforms model ext-XSI-n for all training regioisomers of $C_{70}H_{20+n}$ (n = 2–38). The performance of both models is measured by R^2 and RMSD for predicting the relative energies of regioisomers of $C_{70}H_{20+n}$ on the same test set. The R^2 (left y-axis in Fig. 6) for model DNN-n (solid blue squares) is very close to 1 (ranging from 0.98 to 0.99) at all values of n, while model ext-XSI-n (empty blue squares) shows a notably lower R^2 (from 0.66 to 0.92). Moreover, the R^2 for model ext-XSI-n generally drops when adding more H atoms; it remains above 0.8 for $n \leq 26$ whereas it falls to 0.66 for n = 38. An even larger difference is seen in RMSD (right y-axis in Fig. 6) between the two models. Model DNN-n has an RMSD less than 4 kcal mol^{-1} and typically around 2 kcal mol^{-1} (see solid red diamonds). In comparison, model ext-XSI-n exhibits a substantially larger RMSD value (empty red diamonds) that generally increases from 7.2 kcal mol^{-1} to 17.5 kcal mol^{-1} as the number of added H atoms grows from 4 to 38. The main reason is probably that the simple HMO theory on which the ext-XSI model is based is less appropriate to describe the electronic structure of carboncones with a larger number of added H atoms, leading to a severely deformed carbon framework that compromises the ideal of π conjugation (see, *e.g.*, Fig. S6 in the ESI†).

Next, let us assess the generalization abilities of both models by looking at the predictions for the totally unknown isomers of $C_{70}H_{20+n}$ by the models trained on Set-(n − 2) that includes isomers from $C_{70}H_{20+2}$ to $C_{70}H_{20+(n-2)}$. Fig. 7 compares the predictions of relative energies of $C_{70}H_{20+18}$ regioisomers by the DNN-16, ext-XSI-16, and ext-XSI0 models. The results for other degrees of hydrogenation are provided in Fig. S9–S14 in the ESI.† In each of the plots in Fig. 7, xTB-computed relative isomer energies, RE_{xTB}, are presented against those predicted by each of the models. The data points are colorized according to the NR value for better visualization. As we can see in Fig. 7a, the relative energies predicted by model DNN-16, RE_{DNN-16}, agree well with RE_{xTB} (R^2 = 0.91 and RMSD = 5.0 kcal mol^{-1}). More importantly, larger deviations are almost exclusively distributed in the high-energy region while in the low-energy region (*e.g.*, within the energy window of RE_{xTB} < 30 kcal mol^{-1}) excellent correlation is observed between RE_{DNN-16} and RE_{xTB}. As a result, model DNN-16 predicts the lowest-energy regioisomer that is precisely the same as the one determined by xTB and DFT calculations (see the encircled data point). In comparison, the prediction results in Fig. 7b indicate the considerably poorer performance of model ext-XSI-16. Despite a satisfactory overall correlation between the model-predicted and the xTB relative energies (R^2 = 0.84 and RMSD = 5.7 kcal mol^{-1}), significantly large deviations with data points dispersed from the overall correlation are seen in the low-energy region. For instance, model ext-XSI-16 correctly predicts that the $C_{70}H_{20+18}$ regioisomer with the lowest RE_{xTB} has the lowest ext-XSI value, but the regioisomer with the second lowest ext-XSI value is the 790th isomer in ascending order of RE_{xTB}. The predictive performance of model ext-XSI0 is even worse than model ext-XSI-16, as shown in Fig. 7c. This is understandable since the dynamic ext-XSI-n model has been trained with constantly updated knowledge whereas the original model ext-XSI0 was only trained once on all 259 regioisomers of $C_{70}H_{20+2}$ (using DFT energies).[58] Although the overall correlation still seems fairly acceptable (R^2 = 0.74 and RMSD = 7.2 kcal mol^{-1}), no correlation is evident between the predicted and the xTB relative energies in the low-energy region. Nevertheless, the lowest-energy regioisomer is still found among those with the lowest ext-XSI values (ranked the 5th lowest). As we can conclude, the problem for these ext-XSI models is that one needs to include a large number of candidates to guarantee the search for the lowest-energy isomers, making the study of regioselectivity much less efficient and likely less reliable.

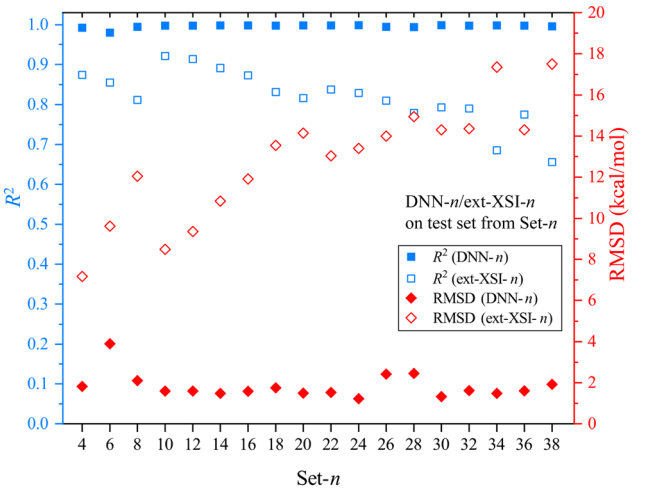


Fig. 6 Comparison of the predictive performance between the DNN-n (solid symbols) and the ext-XSI-n (empty symbols) models, as evaluated on the same test set from Set-n for $C_{70}H_{20+n}$ regioisomers. Red diamonds (right y-axis) denote the RMSD of the model-predicted relative isomer energies from the xTB computed ones, while blue squares (left y-axis) represent the corresponding squared correlation coefficient, R^2.

We have also applied the DNN models to predict the relative stability of hydrogenated carboncone $C_{62}H_{16}$ and compared the performances with those for model ext-XSI0. As shown in Fig. S8a in the ESI,† model DNN-n exhibits good performance on the test set from Set-n for all considered degrees of hydrogenation (n = 4–28), with R^2 ranging from 0.96 to 0.99 and RMSD from 1.4 to 2.8 kcal mol^{-1}. Furthermore, the DNN-n model predicts satisfactorily the relative regioisomer energies of $C_{62}H_{16+(n+2)}$ (n + 2 = 6–30), as evidenced in Fig. S15–S19 in the ESI.† In a direct comparison, model ext-XSI0 makes very poor predictions and even the overall R^2 between the predicted and the xTB relative energies is mostly only 0.1–0.3.

3.1.2 Energetically favorable addition patterns. The DNN predictions combined with the subsequent xTB and DFT refinement calculations have allowed us to identify the lowest-energy regioisomers of hydrogenated carboncones, from which we

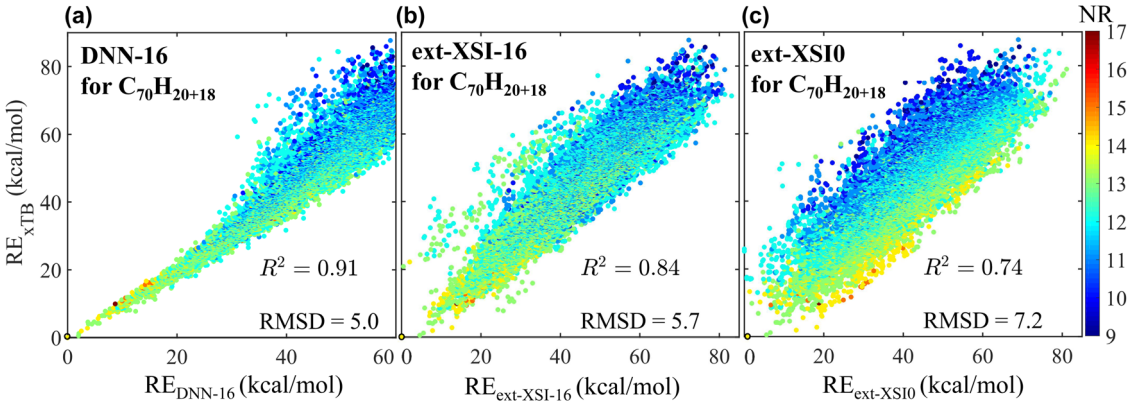


Fig. 7 Performances of the DNN and ext-XSI models in predicting the xTB relative energies, RE_{xTB}, of $C_{70}H_{20+18}$ regioisomers. RE_{xTB} is compared with the relative energies predicted by (a) the DNN-16, (b) the ext-XSI-16, and (c) the original ext-XSI0 models. R^2 and RMSD (in kcal mol^{-1}) between the xTB and the model-predicted values are indicated in each plot. The data points are colorized according to the NR values for the corresponding addition patterns.

may extract useful information about the favorable addition patterns. Fig. 8 presents the DFT determined top 5 lowest-energy addition patterns of $C_{70}H_{20+n}$ for n = 2, 4, 10, 20, 30, and 40 (see Fig. S26–S28 in the ESI,† for all values of n). As mentioned in the previous study of $C_{70}H_{20+n}$ with n up to 12,[58] the addition of H atoms preferably takes place on the rim sites of the carboncone when a relatively smaller number H atoms are added ($n \leq 12$), as can be seen in the energetically low-lying regioisomers depicted in Fig. 8 and Fig. S26 in the ESI.† The driving force for such addition patterns is primarily the π stabilization effect,[57,58] which is supported by the observation that the NR and NA have a general positive correlation with the stability of adducts.[58] Addition patterns with larger values of NR and NA maintain a larger area of π conjugated system and are therefore more stable.

When adding more H atoms, however, the central pentagon becomes competitive addition sites. One of the top 5 isomers of $C_{70}H_{20+14}$ and 4 of the top 5 isomers of $C_{70}H_{20+16}$ have 1 to 3 H atoms added to the central pentagon (see Fig. S27 in the ESI†). For $n \geq 18$, all top 5 regioisomers of adducts have H atoms added to the pentagon sites, as shown in Fig. 8 and Fig. S27 (ESI†). This implies that π delocalization is not the, or at least not the only, determining factor for the stability of adducts for higher degrees of hydrogenation. Indeed, the lowest-energy regioisomers are no longer those that maximize NR when adding a large number of H atoms. For instance, as shown in Fig. 8, all top 5 lowest-energy regioisomers of $C_{70}H_{20+20}$ have an NR value of 13 or 14, considerably less than the maximum possible value of 18 (as indicated in the square brackets in Fig. 8). Conversely, all top 5 regioisomers of $C_{70}H_{20+10}$ possess the maximum possible NR of 21, as shown in Fig. 8. The preference of the central pentagon sites over other sites for adding a larger number of H atoms is most likely driven by the strain release upon addition to the pentagon sites (*cf.* the κNP_j term in eqn (3)).

Interestingly, as the number of added H atoms further increases, the pentagon sites are preferably not fully occupied by the addends (see, *e.g.*, **iso30-1** and **iso40-1** in Fig. 8). Instead, the additional H atoms tend to aggregate to form a chain-like structure that divides the carboncone molecule into two roughly even parts, as can be seen in **iso30-1**, **iso30-2**, and **iso40-1–iso40-5** in Fig. 8. Consequently, in the resultant hydrogenated molecule the original conical carbon framework is deformed to a gable-roof-like structure, where the aforementioned two parts split by the hydrogen chain maintain a more or less planar surface (see the 3D molecular structure of **iso40-1** in Fig. S5a, ESI†). Thereby, the two quasi-planar areas containing aromatic intact rings acquire more efficient π conjugation and therefore achieve optimal aromatization stabilization.

Regarding the hydrogenation of carboncone $C_{62}H_{16}$, the top 5 lowest-energy regioisomers are presented in Fig. 9 for representative numbers of added H atoms. When adding 2 or 4 H atoms, the addends prefer to go to the peripheral sites as in the case of $C_{70}H_{20+2,4}$, but it is not intuitive to see that they reside separately rather than being paired up with each other on the same C–C bond. Adding more H atoms ($n \geq 6$), we see in Fig. 9 and Fig. S29–S31 in the ESI† that the pentagon sites become more competitive. When the number of addends is sufficiently large ($n \geq 10$), the excessive H atoms are prone to exhibit a chain-like arrangement forming a greatly distorted gable-roof-like molecular structure, which is divided by the "roof ridge" into two nearly planar aromatic subsystems (see Fig. S5b, ESI†). These results are generally similar to those for the afore-discussed hydrogenation of $C_{70}H_{20}$.

3.2 Chlorination of fullerenes

Now, we move to chlorination of fullerenes. Since there is an abundance of experimentally synthesized chlorinated fullerenes,[38,91] here we consider only two representative fullerene cages for chlorination: the non-isolated-pentagon-rule[92] (non-IPR) C_{50}-D_{5h}(271) cage and the larger, IPR C_{76}-D_2(1) cage (hereafter referred to simply as C_{50} and C_{76}, respectively). Our aim is to correctly predict the experimentally identified chlorination patterns for both cages, including $C_{50}Cl_{10}$,[69,70] $C_{76}Cl_{18}$,[65] $C_{76}Cl_{24}$,[66] $C_{76}Cl_{28}$,[66] and $C_{76}Cl_{34}$.[66] Especially, the above series of C_{76} chlorides poses a very challenging task, for all those addition patterns have not been successfully predicted by any of

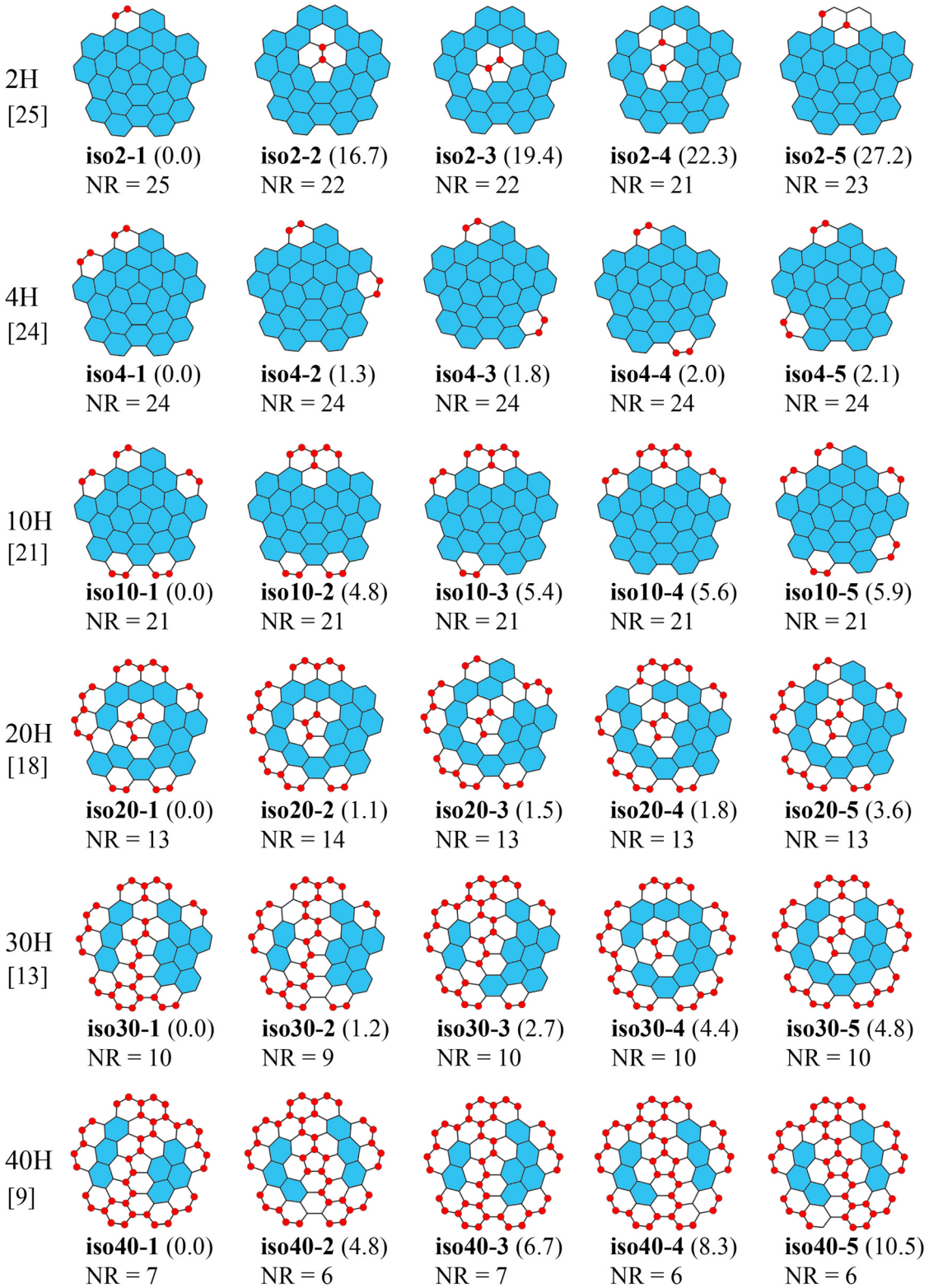


Fig. 8 The five lowest-energy regioisomers of hydrogenated carboncones $C_{70}H_{20+n}$ (n = 2, 4, 10, 20, 30, 40). Below each of the addition pattern illustrations the DFT relative energy including ZPE (in parentheses) and the NR value are provided. The maximum possible NR value for each degree of hydrogenation is given in square brackets in the leftmost column.

the known models,[38,64] due partly to their highly distorted cage framework upon such relatively high degrees of chlorination. Therefore, in the following we will pay special attention to the chlorinated C_{76} cases.

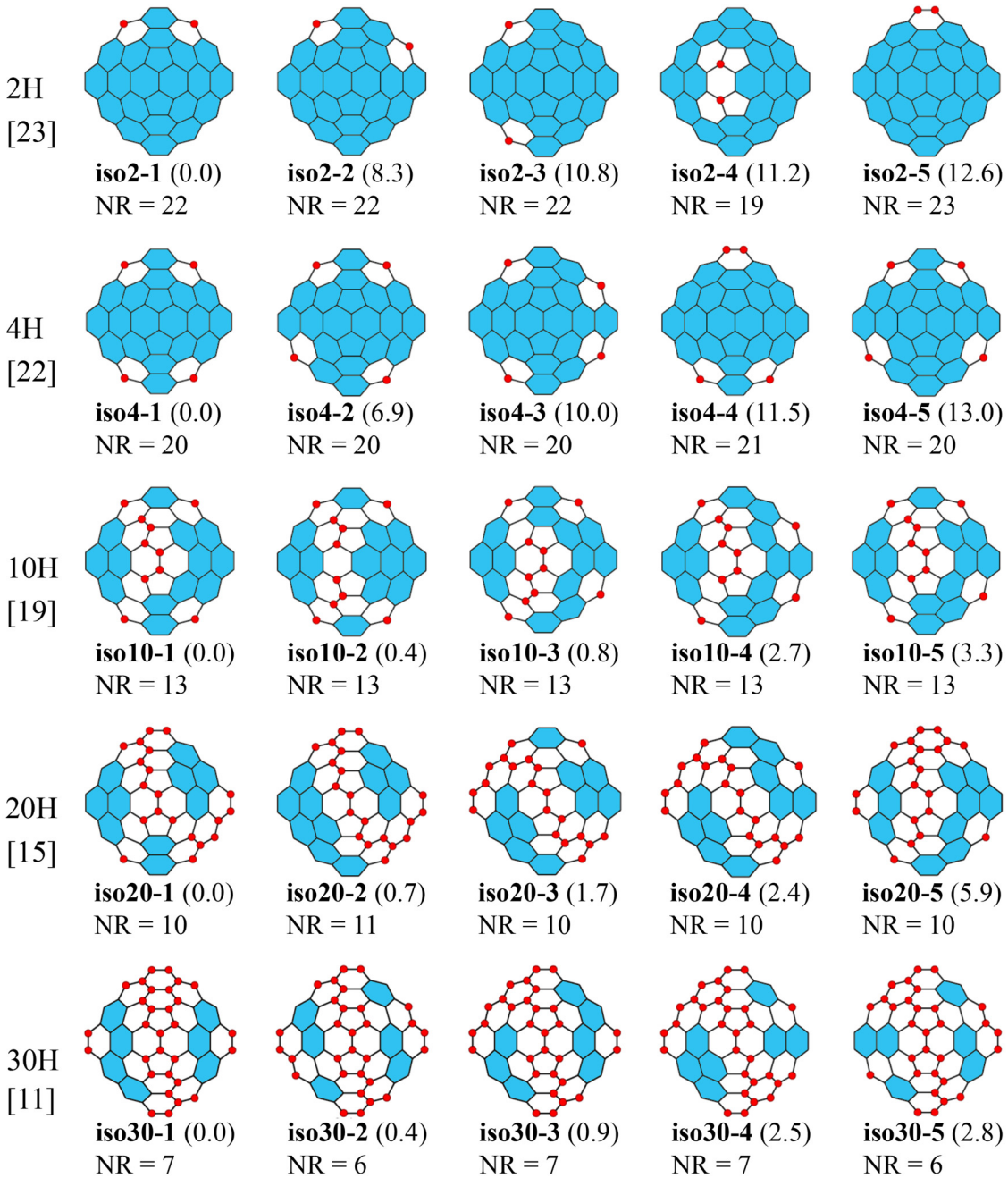


Fig. 9 The same as in Fig. 8 but for regioisomers of $C_{62}H_{16+n}$ (n = 2, 4, 10, 20, 30).

3.2.1 Model performance. We first trained the DNN models to predict the regioselectivity of $C_{50}Cl_n$ (n = 6–10). Model DNN-4 shows a fairly good performance for the test set (R^2 = 0.80 and RMSD = 1.8 kcal mol^{-1}), as shown in Fig. S8b in the ESI.† Its performance in predicting RE_{xTB} of $C_{50}Cl_6$ regioisomers is less impressive, with R^2 = 0.58 and RMSD = 6.9 kcal mol^{-1} (Fig. S20a, ESI†). The predictive power of DNN is boosted with the increasing number of added Cl atoms. Models DNN-6 and DNN-8 behave quite well for both the evaluation on the test set (R^2 about 0.95 and RMSD around 2.6 kcal mol^{-1}, see Fig. S8b, ESI†), and the prediction for $C_{50}Cl_8$ and $C_{50}Cl_{10}$ (R^2 = 0.86 and 0.90, respectively, and RMSD = 4.5 and 3.7 kcal mol^{-1}, respectively, see Fig. S20c and e, ESI†). We have also checked the performance of the XSI model for these systems. As we can compare in Fig. S20 in the ESI,† the DNN model outperforms the XSI model in all cases. Nonetheless, the overall prediction by the XSI model is still satisfactory, especially for $C_{50}Cl_{10}$ regioisomers (R^2 = 0.76, see Fig. 10b). As shown in Fig. 10 and Table 1, both the DNN and XSI models correctly predict that the experimentally determined structure[69,70] (see Fig. S32 in the ESI†) corresponds to the energetically most favorable isomer that stands out prominently from all other regioisomers.

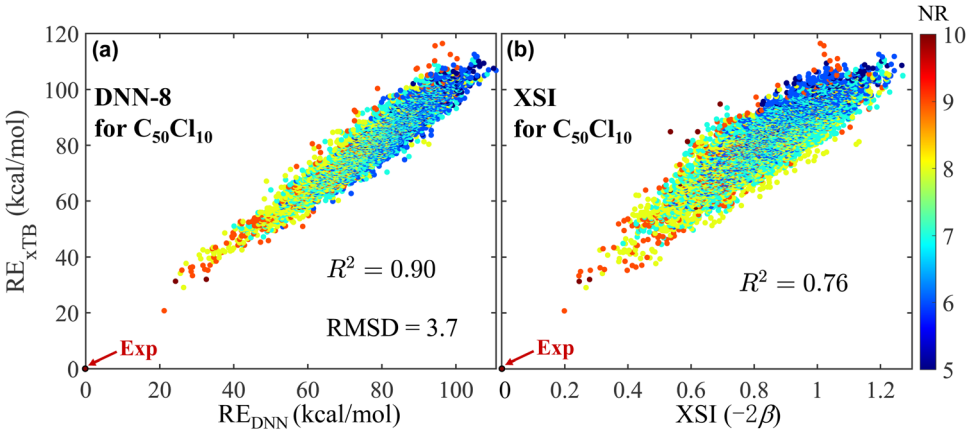


Fig. 10 Performances of the DNN and XSI models in predicting the xTB relative energies, RE_{xTB}, of regioisomers of $C_{50}Cl_{10}$. RE_{xTB} is plotted against (a) the relative energies predicted by the DNN model and (b) the XSI values given by eqn (1). Squared correlation coefficient, R^2, is provided and the RMSD is indicated in (a). The color code of data points scales with the NR value. The experimentally identified regioisomer[69,70] is pointed out.

Now, let us focus on the challenging case of chlorination of C_{76}. All of the trained models, DNN-n (n = 4–32), exhibit a good performance on the test set, with an R^2 of 0.94–0.99 and an RMSD of 1.6–2.4 kcal mol^{-1} (see Fig. S8c in the ESI†). With this confidence, we applied the trained DNN-n models to predict the relative isomer energies of $C_{76}Cl_{n+2}$. Fig. 11 compares RE_{xTB} with the DNN-predicted values, RE_{DNN}, for n = 18, 24, 28, and 34, all four cases where experimental structures[65,66] have been unambiguously observed. Note that here we only present all regioisomers within a relatively low-energy region (RE_{xTB} < 35 kcal mol^{-1}), excluding the unstable and hence less important structures. As we can see, model DNN-16 makes a good prediction of the relative stability for $C_{76}Cl_{18}$ regioisomers (R^2 = 0.92 and RMSD = 1.6 kcal mol^{-1}) and the experimental regioisomer[65] (see Fig. 12a) has the lowest value of both RE_{xTB} and RE_{DNN}. For $C_{76}Cl_{24}$ and $C_{76}Cl_{28}$ with significantly greater degrees of chlorination, the DNN model still maintains satisfactory performance, with an R^2 around 0.8 and an RMSD about 2–3 kcal mol^{-1} (Fig. 11b and c). Although the DNN does not predict the experimental isomer[66] as the lowest-energy structure in either case, the experimental structure ranks very highly according to the ascending order of RE_{DNN}, achieving a ranking number of 3 and 5 for $C_{76}Cl_{24}$ and $C_{76}Cl_{28}$, respectively (see Table 1). Incidentally, the experimental isomer of $C_{76}Cl_{28}$[66] (Fig. 12c) does not have the lowest xTB energy (being the 11th lowest-energy isomer, see Table 1). Subsequent DFT refinement calculations reveal that it is actually the second lowest-energy structure, lying 1.8 kcal mol^{-1} higher in energy than a never reported regioisomer (see Fig. 12f). It is sensible to anticipate the formation of this lowest-energy structure of $C_{76}Cl_{28}$ under thermodynamically-controlled conditions. For the same reason, we may also expect the production of the second lowest-energy isomer of $C_{76}Cl_{24}$ (see Fig. 11e), which lies only 2.7 kcal mol^{-1} above the experimental structure.[66]

However, $C_{76}Cl_{34}$ is proved to be still an intractable case. As shown in Fig. 11d, the DNN-predicted relative energies have a low correlation with the xTB results (R^2 = 0.50). More problematically, the structure observed in experiments[66] was not even present in the generated structures of $C_{76}Cl_{34}$ following the stepwise addition procedure guided by DNN. As a matter of fact, both the xTB and DFT calculations confirm that the experimental structure of $C_{76}Cl_{34}$ corresponds indeed to the lowest-energy isomer. Although DNN has missed the experimental isomer, we have still evaluated its RE_{DNN} value to be 25.0 kcal mol^{-1}, leading to a ranking number of 1782 for the experimental structure among all considered regioisomers (see Table 1).

Table 1 Performance of DNN and XSI models in predicting the experimental regioisomer of $C_{50}Cl_{10}$ and $C_{76}Cl_n$ (n = 18, 24, 28, and 34). We present R^2 and RMSD between RE_{xTB} and RE_{DNN} for all considered regioisomers, relative energy of the experimental isomer, RE(exp) (in kcal mol^{-1}), predicted by DFT, xTB, and DNN methods, and ranking number of the experimental isomer, Rank(exp), determined by DFT, xTB, DNN, and XSI models

			RE(exp)			Rank(exp)			
Compound	R^2	RMSD	DFT	xTB	DNN	DFT	xTB	DNN	XSI
$C_{50}Cl_{10}$	0.90	3.7	0.0	0.0	0.0	1	1	1	1
$C_{76}Cl_{18}$	0.92	1.6	0.0	0.0	0.0	1	1	1	207
$C_{76}Cl_{24}$	0.80	2.3	0.0	0.0	0.7	1	1	3	1993
$C_{76}Cl_{28}$	0.79	2.9	1.8[a]	3.6	1.9	2[a]	11	5	2263
$C_{76}Cl_{34}$	0.50	3.0	0.0	0.0	25.0	1	1	1782[b]	8153[b]

[a] We have discovered an unprecedented $C_{76}Cl_{28}$ regioisomer lower in energy than the experimental structure. [b] Both the DNN-32 and XSI models failed to generate the experimental regioisomer of $C_{76}Cl_{34}$.

On the other hand, the XSI model behaves very poorly in predicting the relative stability of $C_{76}Cl_n$, with an R^2 from 0.09–0.21, as evidenced in Fig. 11e–h. The XSI method failed to even generate any of the experimentally observed regioisomers of $C_{76}Cl_n$ (n = 18, 24, 28, and 34). Based on the XSI values, the experimental isomers are ranked at rather low positions among all generated regioisomers of $C_{76}Cl_n$ (*e.g.*, 2263rd for n = 28), as shown in Table 1. The reason for the inadequacy of XSI is probably because of the significant distortion of carbon cage

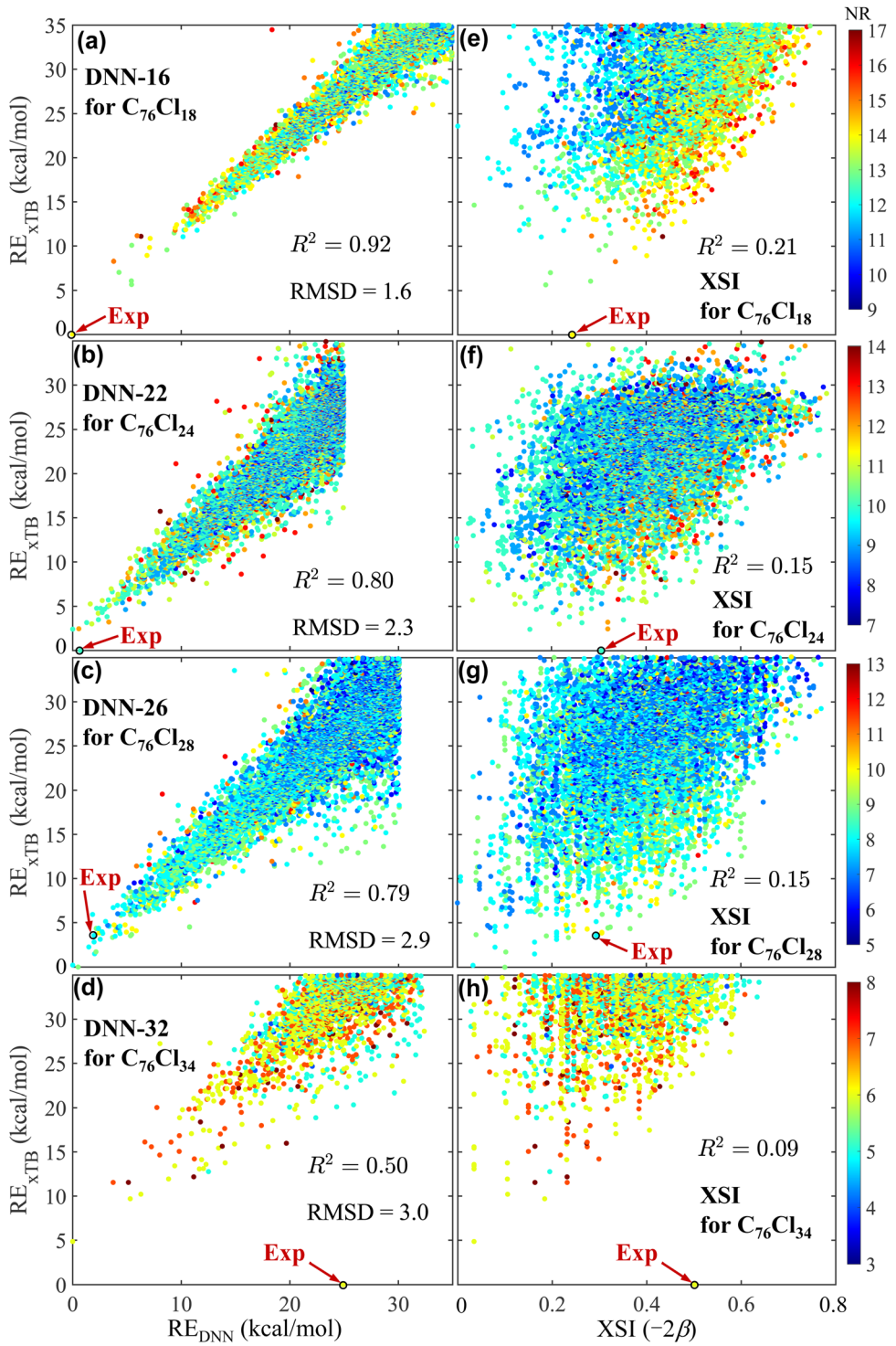


Fig. 11 Performances of the DNN and XSI models in predicting the xTB relative energies, RE_{xTB}, of regioisomers of $C_{76}Cl_n$ (n = 18, 24, 28, and 34). (a)–(d) RE_{xTB} *versus* the relative energies predicted by the DNN model. (e)–(h) RE_{xTB} *versus* the XSI values given by eqn (1). Other descriptions are the same as in Fig. 10.

upon chlorination, similar to the cases of highly hydrogenated carboncone as discussed in Section 3.1.1.

3.2.2 Understanding the failure of the DNN model for $C_{76}Cl_{34}$. The reason for the failure of the DNN model in predicting the experimental structure of $C_{76}Cl_{34}$ is probably that the experimental addition pattern at such a high degree of chlorination (covering nearly 45% area of fullerene surface) is substantially different from the addition patterns at lower degrees of chlorination. As we can see in Fig. 12a and b, the addition pattern of $C_{76}Cl_{24}$ includes the entire addition pattern

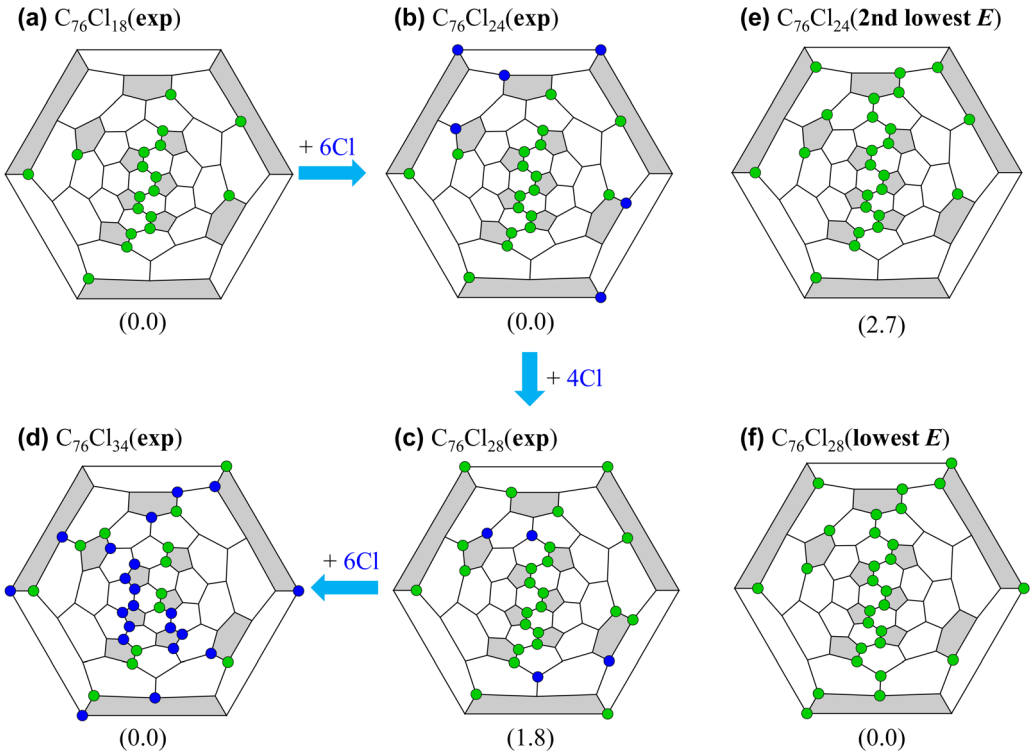


Fig. 12 Schlegel diagrams showing the evolution of the experimentally observed[65,66] addition patterns in (a) $C_{76}Cl_{18}$, (b) $C_{76}Cl_{24}$, (c) $C_{76}Cl_{28}$, and (d) $C_{76}Cl_{34}$. In this sequence, the addition positions that differ from those in the previous structure are highlighted in blue. (e) The second lowest-energy regioisomer of $C_{76}Cl_{24}$. (f) The lowest-energy regioisomer of $C_{76}Cl_{28}$. The DFT relative energy including ZPE in kcal mol^{-1} is given below each structure.

of $C_{76}Cl_{18}$ (green circles), plus 6 additional Cl atoms at new positions on the fullerene cage (blue circles). Likewise, the addition pattern of $C_{76}Cl_{28}$ (Fig. 12c) can be regarded as simply adding 4 more Cl atoms on top of the addition pattern of $C_{76}Cl_{24}$ (Fig. 12b). However, when going from $C_{76}Cl_{28}$ to $C_{76}Cl_{34}$, almost 40% Cl atoms (blue circles in Fig. 12d) are located at different sites with respect to the addition pattern in $C_{76}Cl_{28}$. Consequently, it is probably the drastic change in the addition pattern of $C_{76}Cl_{34}$ that causes the failure of the incremental DNN model, which attempts to predict largely new addition patterns but with the knowledge learned from very dissimilar addition patterns in lower degrees of chlorination.

Moreover, the stepwise addition algorithm[25,38] also seems ineffective in handling the substantial variation in addition patterns, thus explaining why the experimental $C_{76}Cl_{34}$ regioisomer was not generated by this algorithm. Since a previous work[64] suggests that the remarkable distortion of the C_{76} cage framework upon chlorination is primarily responsible for the deficiency of machine learning models, we have inspected the distortion of the fullerene cage in $C_{76}Cl_n$ with an increasing n. As shown in Fig. S7 in the ESI,† the cage distortion upon chlorination, D (see Section 4 of the ESI,† for definition and evaluation), increases generally with the increasing number of added Cl atoms. Fig. S33 in the ESI,† also visually reveals the deformation of these highly chlorinated fullerene cages. Albeit the considerable distortion in cage framework, the DNN models are still able to predict these experimental structures. However, in the failure case of the experimental $C_{76}Cl_{34}$ structure the cage distortion is even slightly less than that in the lowest-energy structure of $C_{76}Cl_{28}$, as indicated in Fig. S7 in the ESI.† Therefore, we believe that the key reason why the DNN fails to deal with $C_{76}Cl_{34}$ lies in the profound change in addition patterns rather than the change in fullerene cage geometry.

3.2.3 Feature importance analysis. To elucidate the roles of the topology-based features in determining the addition patterns, we have evaluated the importance of each individual feature in our DNN models. We quantify a feature's importance as the difference between the baseline performance (*i.e.*, that given by the original model including all features) and the reduced performance after neutralizing the corresponding feature in the model prediction. A large such difference indicates that the feature significantly impacts model performance, while a small difference suggests a minor role. More details on the computation of feature importances can be found in Section 8 of the ESI.†

Fig. 13 presents the feature importances for predicting addition patterns in the chlorinated fullerenes that include the experimentally determined regioisomers. As we can see, the NR feature (red bar) plays a primary role in determining the addition patterns in $C_{50}Cl_{10}$, $C_{76}Cl_{18}$, and $C_{76}Cl_{24}$ (Fig. 13a–c), and is the second most important feature in $C_{76}Cl_{28}$ (Fig. 13d). This observation suggests that aromaticity (or π electronic effect) substantially influences addition regioselectivity.

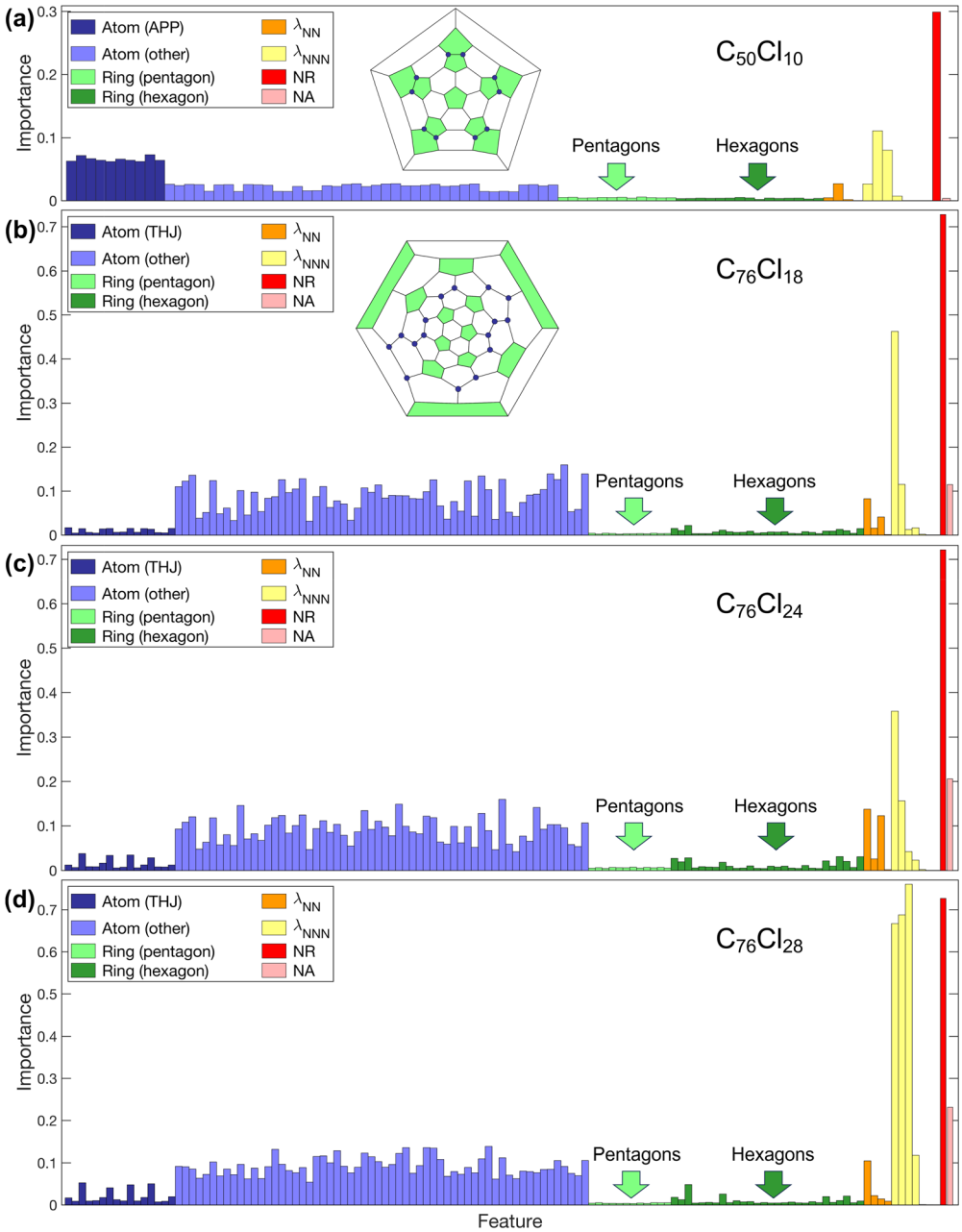


Fig. 13 Importance of individual features for (a) $C_{50}Cl_{10}$, (b) $C_{76}Cl_{18}$, (c) $C_{76}Cl_{24}$, and (d) $C_{76}Cl_{28}$. The bars are colorized according to their feature types. Dark blue circles in the Schlegel diagram in (a) and (b) represent the APP and THJ carbon atoms, respectively.

Notably, the NA feature (pink bar) has little impact in the $C_{50}Cl_{10}$ case and exhibits increasing importance as more Cl atoms are added to the C_{76} cage. The λ_{NNN}-based features (yellow bars) emerge as the next most influential factor, matching the NR feature in importance for $C_{76}Cl_{28}$. In contrast, λ_{NN}-based features (orange bars) are relatively less impactful. This result suggests that the NNN relationship between addends has a stronger influence on improving the model's predictive power. Thus, we may speculate that incorporating an additional layer for addition site information, *i.e.*, the next-next-nearest-neighbor relationship, could further enhance the model accuracy.

The atom-based features also show appreciable importance. For $C_{50}Cl_{10}$ (Fig. 13a), the features linked to the addition sites at adjacent pentagon pairs (APPs, dark blue bar) are notably more important than those at other sites (light blue bars). This is consistent with the preference for chlorine addition at APPs due to ring strain stabilization.[38] For chlorinated C_{76} cases, the features (dark blue bar) associated with the triple-hexagon junction (THJ)[93] sites (see in the Schlegel diagram in Fig. 13b) are much less critical than those at other sites (light blue bars). This result aligns with the fact that THJ sites are less reactive due to the more planar, sp^2-like configuration of C atoms.[94] Lastly, all ring-based features (green bars) generally

play a minor role in predicting addition patterns for the systems considered.

4 Conclusions

This study employs DNN models coupled with the stepwise addition algorithm to predict the lowest-energy addition patterns in hydrogenated carboncones and chlorinated fullerenes. Unlike the previous machine learning approaches relying on NNP geometry optimizations,[63,64] our DNN models utilize purely graph-based features requiring no information of 3D coordinates of atoms in the training or prediction process. Thus, this advantage inherently avoids the problem that invalid candidate structures may emerge after NNP optimizations for higher degrees of chemical additions. Hence, the present approach has enabled us to explore the addition patterns in highly hydrogenated carboncones and highly chlorinated fullerenes. In the carboncone cases, the DNN models perform reasonably well even if the carboncone framework suffers from an extensive structural deformation when adding a large number of H atoms. Our DNN method shows remarkable predictive power for chlorinated fullerenes. In particular, the experimental structures of $C_{76}Cl_n$ (n = 18, 24, and 28),[65,66] which were intractable to previous approaches,[38,64] have been successfully predicted by DNN as one of the lowest-energy regioisomers. We have also found two new isomers of $C_{28}Cl_{24}$ and $C_{76}Cl_{28}$, which are almost equally or even more stable than the experimental isomers and would therefore have a good chance to be observed in future experiments. Compared with the previous XSI and ext-XSI models, the DNN models are shown to be superior in almost all cases, especially for high degrees of addition where the carbon framework is notably deformed.

Admittedly, the performance of DNN deteriorated significantly when the number of Cl atoms reached 34, the maximum chlorination degree for C_{76} reported in experiments.[66] The failure of our approach in this case is not essentially due to the fullerene cage distortion upon chlorination, but the drastic change in addition pattern when transitioning from $C_{76}Cl_{28}$ to $C_{76}Cl_{34}$. Hence, the key to solving this problem would be to find more effective ways, in lieu of the stepwise algorithm, to thoroughly explore the potential energy surface. Unlike the usual cases of molecular structure optimization, the search space of addition patterns is not continuous and less intuitive; moving just a single additional atom from one site to its adjacent site may lead to a dramatic change in energy. In light of this, a stochastic search driven by global optimization algorithms (like genetic optimization,[95] simulated annealing, and Monte Carlo simulation, *etc*), in combination with the stepwise addition scheme, seems a possible strategy and will be our research focus in future work.

Data availability

All datasets, TensorFlow scripts for training models, and the final trained DNN models are at **https://drive.google.com/drive/folders/1fLn3jwzCiCAc7rmQDCwnZ1c8EMlSPB0B**, which are freely available. The XSI calculations combined with the stepwise algorithm were automatically conducted by the XSIopt module[25] in our open-source code, FullFun.[96] The ext-XSI calculations were performed using our homemade[97] scripts, which are freely available *via* the aforementioned Google Drive link. We used MATLAB to produce most of the data plots and used Jmol (version 14.30.2)[98] to draw 3D molecular structures.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (22073080) and the Double Innovation Talent Program of Jiangsu Province (JSSCRC2021542).

References

1 S. Iijima, T. Ichihashi and Y. Ando, Pentagons, heptagons and negative curvature in graphite microtubule growth, *Nature*, 1992, **356**, 776–778.
2 Z.-Z. Zhu, Z.-C. Chen, Y.-R. Yao, C.-H. Cui, S.-H. Li, X.-J. Zhao, Q. Zhang, H.-R. Tian, P.-Y. Xu, F.-F. Xie, X.-M. Xie, Y.-Z. Tan, S.-L. Deng, J. M. Quimby, L. T. Scott, S.-Y. Xie, R.-B. Huang and L.-S. Zheng, Rational synthesis of an atomically precise carboncone under mild conditions, *Sci. Adv.*, 2019, **5**, eaaw0982.
3 K. Shoyama and F. Würthner, Synthesis of a Carbon Nanocone by Cascade Annulation, *J. Am. Chem. Soc.*, 2019, **141**, 13008–13012.
4 Y. Wang, M. Alcamí and F. Martín, in *Handbook of Nanophysics*, ed. K. D. Sattler, *Clusters and Fullerenes*, Taylor & Francis Publisher (CRC Press), London, 2010, vol. 2, ch. 25, pp. 1–23.
5 F. L. De La Puente and J.-F. Nierengarten, *Fullerenes: principles and applications*, Royal Society of Chemistry, Cambridge, 2011.
6 X. Lu, T. Akasaka and Z. Slanina, *Handbook of fullerene science and technology*, Springer Nature, Singapore, 2022.
7 D. J. Klein and A. T. Balaban, The Eight Classes of Positive-Curvature Graphitic Nanocones, *J. Chem. Inf. Model.*, 2006, **46**, 307–320.
8 P. W. Fowler, S. Nikolić, R. De Los Reyes and W. Myrvold, Distributed curvature and stability of fullerenes, *Phys. Chem. Chem. Phys.*, 2015, **17**, 23257–23264.
9 S. H. Pun and Q. Miao, Toward Negatively Curved Carbons, *Acc. Chem. Res.*, 2018, **51**, 1630–1642.
10 P. Woods, The discovery of cosmic fullerenes, *Nat. Astron.*, 2020, **4**, 299–305.
11 Y. Zhang, S. Sadjadi and C.-H. Hsia, Hydrogenated fullerenes (fulleranes) in space, *Astrophys. Space Sci.*, 2020, **365**, 67.

12 H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl and R. E. Smalley, C_{60}: Buckminsterfullerene, *Nature*, 1985, **318**, 162–163.
13 M. Prato, [60]Fullerene chemistry for materials science applications, *J. Mater. Chem.*, 1997, **7**, 1097–1109.
14 F. Wudl, Fullerene materials, *J. Mater. Chem.*, 2002, **12**, 1959–1963.
15 S. Yao, X. Yuan, L. Jiang, T. Xiong and J. Zhang, Recent Progress on Fullerene-Based Materials: Synthesis, Properties, Modifications, and Photocatalytic Applications, *Materials*, 2020, **13**, 2924.
16 S. Goodarzi, T. Da Ros, J. Conde, F. Sefat and M. Mozafari, Fullerene: Biomedical engineers get to revisit an old friend, *Mater. Today*, 2017, **20**, 460–480.
17 C. Zhou, M. Zhen, M. Yu, X. Li, T. Yu, J. Liu, W. Jia, S. Liu, L. Li, J. Li, Z. Sun, Z. Zhao, X. Wang, X. Zhang, C. Wang and C. Bai, Gadofullerene inhibits the degradation of apolipoprotein B100 and boosts triglyceride transport for reversing hepatic steatosis, *Sci. Adv.*, 2020, **6**, eabc1586.
18 E. Nakamura and H. Isobe, Functionalized Fullerenes in Water. The First 10 Years of Their Chemistry, Biology, and Nanoscience, *Acc. Chem. Res.*, 2003, **36**, 807–815.
19 R. Biswas, C. Batista Da Rocha, R. A. Bennick and J. Zhang, Water-Soluble Fullerene Monoderivatives for Biomedical Applications, *ChemMedChem*, 2023, **18**, e202300296.
20 J. Li, M. Zhang, B. Sun, G. Xing, Y. Song, H. Guo, Y. Chang, Y. Ge and Y. Zhao, Separation and purification of fullerenols for improved biocompatibility, *Carbon*, 2012, **50**, 460–469.
21 V. V. Sharoyko, O. S. Shemchuk, A. A. Meshcheriakov, L. V. Vasina, N. R. Iamalova, M. D. Luttsev, D. A. Ivanova, A. V. Petrov, D. N. Maystrenko, O. E. Molchanov and K. N. Semenov, Biocompatibility, antioxidant activity and collagen photoprotection properties of C_{60} fullerene adduct with L-methionine, *Nanomedicine*, 2022, **40**, 102500.
22 S. S. Babu, H. Möhwald and T. Nakanishi, Recent progress in morphology control of supramolecular fullerene assemblies and its applications, *Chem. Soc. Rev.*, 2010, **39**, 4021–4035.
23 A. V. Baskar, M. R. Benzigar, S. N. Talapaneni, G. Singh, A. S. Karakoti, J. Yi, A. H. Al-Muhtaseb, K. Ariga, P. M. Ajayan and A. Vinu, Self-Assembled Fullerene Nanostructures: Synthesis and Applications, *Adv. Funct. Mater.*, 2022, **32**, 2106924.
24 X. Chang, Y. Xu and M. von Delius, Recent advances in supramolecular fullerene chemistry, *Chem. Soc. Rev.*, 2024, **53**, 47–83.
25 Y. Wang, S. Díaz-Tendero, M. Alcamí and F. Martín, Topology-Based Approach to Predict Relative Stabilities of Charged and Functionalized Fullerenes, *J. Chem. Theory Comput.*, 2018, **14**, 1791–1810.
26 S. Yang, I. N. Ioffe and S. I. Troyanov, Chlorination-Promoted Skeletal Transformations of Fullerenes, *Acc. Chem. Res.*, 2019, **52**, 1783–1792.
27 R. Guan, M. Chen, F. Jin and S. Yang, Strain Release of Fused Pentagons in Fullerene Cages by Chemical Functionalization, *Angew. Chem., Int. Ed.*, 2020, **59**, 1048–1073.
28 N. B. Tamm, V. A. Brotsman, V. Y. Markov and S. I. Troyanov, Fused-Pentagon $C_{70}Cl_6$ and $C_{70}Cl_8$ Obtained via Chlorination-Promoted Skeletal Transformation of IPR C_{70}, *Inorg. Chem.*, 2020, **59**, 10400–10403.
29 N. B. Tamm, V. Y. Markov, A. A. Goryunkov and S. I. Troyanov, Intermediate Products of C_{60} High-Temperature Chlorination-$C_{60}Cl_n$ (n = 8, 10, 14, 20, 24), *Eur. J. Org. Chem.*, 2020, 6801–6804.
30 V. A. Brotsman, N. B. Tamm and S. I. Troyanov, Structural Chemistry of Pentagon-Fused C_{82} Fullerene Derivatives $^{\#39173}C_{82}(CF_3)_{14,16,18}$ and $^{\#39173}C_{82}Cl_{28}$, *Inorg. Chem.*, 2023, **62**, 2425–2429.
31 V. A. Brotsman and S. I. Troyanov, Non-classical (NC), heptagon-containing fullerenes obtained via chlorination-promoted cage transformations: C_{76} (NC2a)Cl_{24} and C_{76} (NC2b)Cl_{28}, *Chem. Commun.*, 2024, **60**, 893–896.
32 D. C. Elias, R. R. Nair, T. M. G. Mohiuddin, S. V. Morozov, P. Blake, M. P. Halsall, A. C. Ferrari, D. W. Boukhvalov, M. I. Katsnelson, A. K. Geim and K. S. Novoselov, Control of Graphene's Properties by Reversible Hydrogenation: Evidence for Graphane, *Science*, 2009, **323**, 610–613.
33 B. Klærke, Y. Toker, D. B. Rahbek, L. Hornekær and L. H. Andersen, Formation and stability of hydrogenated PAHs in the gas phase, *Astron. Astrophys.*, 2013, **549**, A84.
34 V. Mennella, L. Hornekaer, J. Thrower and M. Accolla, The Catalytic Role of Coronene for Molecular Hydrogen Formation, *Astrophys. J., Lett.*, 2011, **745**, L2.
35 S. M. Luzan, Y. O. Tsybin and A. V. Talyzin, Reaction of C_{60} with Hydrogen Gas: *In Situ* Monitoring and Pathways, *J. Phys. Chem. C*, 2011, **115**, 11484–11492.
36 Q. Wang and P. Jena, Density Functional Theory Study of the Interaction of Hydrogen with Li_6C_{60}, *J. Phys. Chem. Lett.*, 2012, **3**, 1084–1088.
37 K. Jastrzebski and P. Kula, Emerging Technology for a Green, Sustainable Energy-Promising Materials for Hydrogen Storage, from Nanotubes to Graphene-A Review, *Materials*, 2021, **14**, 2499.
38 Y. Wang, S. Díaz-Tendero, M. Alcamí and F. Martín, Relative Stability of Empty Exohedral Fullerenes: π Delocalization versus Strain and Steric Hindrance, *J. Am. Chem. Soc.*, 2017, **139**, 1609–1617.
39 P. A. Troshin, O. Popkov and R. N. Lyubovskaya, Some New Aspects of Chlorination of Fullerenes, *Fullerenes, Nanotubes Carbon Nanostruct.*, 2003, **11**, 165–185.
40 I. V. Kuvychko, A. V. Streletskii, N. B. Shustova, K. Seppelt, T. Drewello, A. A. Popov, S. H. Strauss and O. V. Boltalina, Soluble Chlorofullerenes $C_{60}Cl_{2,4,6,8,10}$. Synthesis, Purification, Compositional Analysis, Stability, and Experimental/Theoretical Structure Elucidation, Including the X-ray Structure of C_1-$C_{60}Cl_{10}$, *J. Am. Chem. Soc.*, 2010, **132**, 6443–6462.
41 K. Ziegler, K. Y. Amsharov and M. Jansen, Synthesis, Separation and Structure Elucidation of a Missing C_{60} Chloride: C_{2v}-$C_{60}Cl_8$, *Z. Naturforsch., B: J. Chem. Sci.*, 2012, **67**, 1091–1097.
42 E. Hückel, Die freien Radikale der organischen Chemie, *Z. Phys.*, 1933, **83**, 632–668.
43 E. Hückel, Quanstentheoretische Beiträge zum Benzolproblem. II, *Z. Phys.*, 1931, **72**, 310–337.

44 E. Hückel, Quantentheoretische Beiträge zum Benzolproblem. I., *Z. Phys.*, 1931, **70**, 204–286.
45 E. Hückel, Quantentheoretische Beiträge zum Problem der aromatischen und ungesättigten Verbindungen. III, *Z. Phys.*, 1932, **76**, 628–648.
46 K. M. Rogers and P. W. Fowler, A model for pathways of radical addition to fullerenes, *Chem. Commun.*, 1999, 2357–2358.
47 W.-W. Wang, J.-S. Dang, J.-J. Zheng and X. Zhao, Heptagons in C_{68}: Impact on Stabilities, Growth, and Exohedral Derivatization of Fullerenes, *J. Phys. Chem. C*, 2012, **116**, 17288–17293.
48 E. F. Sheka, Stepwise computational synthesis of fullerene C_{60} derivatives. Fluorinated fullerenes $C_{60}F_{2k}$, *J. Exp. Theor. Phys.*, 2010, **111**, 397–414.
49 P. A. Cahill and C. M. Rohlfing, Theoretical studies of derivatized buckyballs and buckytubes, *Tetrahedron*, 1996, **52**, 5247–5256.
50 B. W. Clare and D. L. Kepert, Stereochemical patterns in $C_{60}X_n$, *J. Phys. Chem. Solids*, 1997, **58**, 1815–1821.
51 G. Van Lier, M. Cases, C. P. Ewels, R. Taylor and P. Geerlings, Theoretical Study of the Addition Patterns of C_{60} Fluorination: $C_{60}F_n$ (n = 1–60), *J. Org. Chem.*, 2005, **70**, 1565–1579.
52 C. P. Ewels, G. Van Lier, P. Geerlings and J.-C. Charlier, Meta-Code for Systematic Analysis of Chemical Addition (SACHA): Application to Fluorination of C_{70} and Carbon Nanostructure Growth, *J. Chem. Inf. Model.*, 2007, **47**, 2208–2215.
53 A. Bihlmeier, D. P. Tew and W. Klopper, Low energy hydrogenation products of extended π systems C_nH_{2x}: A density functional theory search strategy, benchmarked against CCSD(T), and applied to C_{60}, *J. Chem. Phys.*, 2008, **129**, 114303.
54 A. Bihlmeier, Derivatives and dimers of C_{50}-D_{5h} and C_{50}-D_3: a comparison of two closely related but quite differently behaving fullerenes, *J. Chem. Phys.*, 2011, **135**, 044310.
55 C.-l Gao, L. Abella, Y.-Z. Tan, X.-Z. Wu, A. Rodrguez-Fortea, J. M. Poblet, S.-Y. Xie, R.-B. Huang and L.-S. Zheng, Capturing the Fused-Pentagon C_{74} by Stepwise Chlorination, *Inorg. Chem.*, 2016, **55**, 6861–6865.
56 P. Pla, Y. Wang, F. Martín and M. Alcamí, Isomers of Hydrogenated Polycyclic Aromatic Hydrocarbons Explain the Presence of Infrared Bands in the 3 μm Region, *Astrophys. J.*, 2020, **899**, 18.
57 P. Pla, Y. Wang, F. Martín and M. Alcamí, Hydrogenated polycyclic aromatic hydrocarbons: isomerism and aromaticity, *Phys. Chem. Chem. Phys.*, 2020, **22**, 21968–21976.
58 X. Chen, Y. Sun and Y. Wang, Stereo- and Regioselectivity of Hydrogenation of a Recently Synthesized Carboncone and Its Predictive Models, *J. Org. Chem.*, 2022, **87**, 10755–10767.
59 Z. J. Baum, X. Yu, P. Y. Ayala, Y. Zhao, S. P. Watkins and Q. Zhou, Artificial Intelligence in Chemistry: Current Trends and Future Directions, *J. Chem. Inf. Model.*, 2021, **61**, 3197–3212.
60 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio and M. Zitnik, Scientific discovery in the age of artificial intelligence, *Nature*, 2023, **620**, 47–60.
61 J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems, *Chem. Rev.*, 2021, **121**, 9816–9872.
62 B. Huang and O. A. von Lilienfeld, Ab Initio Machine Learning in Chemical Compound Space, *Chem. Rev.*, 2021, **121**, 10001–10036.
63 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K. R. Müller, SchNetPack: A Deep Learning Toolbox For Atomistic Systems, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.
64 M. Liu, Y. Han, Y. Cheng, X. Zhao and H. Zheng, Exploring exohedral functionalization of fullerene with automation and Neural Network Potential, *Carbon*, 2023, **213**, 118180.
65 K. S. Simeonov, K. Y. Amsharov and M. Jansen, Connectivity of the Chiral D_2-Symmetric Isomer of C_{76} through a Crystal-Structure Determination of $C_{76}Cl_{18}\cdot TiCl_4$, *Angew. Chem., Int. Ed.*, 2007, **46**, 8419–8421.
66 I. N. Ioffe, O. N. Mazaleva, C. Chen, S. Yang, E. Kemnitz and S. I. Troyanov, C_{76} fullerene chlorides and cage transformations. Structural and theoretical study, *Dalton Trans.*, 2011, **40**, 11005–11011.
67 G. I. Parisi, R. Kemker, J. L. Part, C. Kanan and S. Wermter, Continual lifelong learning with neural networks: a review, *Neural Networks*, 2019, **113**, 54–71.
68 G. M. van de Ven, T. Tuytelaars and A. S. Tolias, Three types of incremental learning, *Nat. Mach. Intell.*, 2022, **4**, 1185–1197.
69 S.-Y. Xie, F. Gao, X. Lu, R.-B. Huang, C.-R. Wang, X. Zhang, M.-L. Liu, S.-L. Deng and L.-S. Zheng, Capturing the Labile Fullerene[50] as $C_{50}Cl_{10}$, *Science*, 2004, **304**, 699.
70 X. Han, S.-J. Zhou, Y.-Z. Tan, X. Wu, F. Gao, Z.-J. Liao, R.-B. Huang, Y.-Q. Feng, X. Lu, S.-Y. Xie and L.-S. Zheng, Crystal Structures of Saturn-Like $C_{50}Cl_{10}$ and Pineapple-Shaped $C_{64}Cl_4$: Geometric Implications of Double- and Triple-Pentagon-Fused Chlorofullerenes, *Angew. Chem., Int. Ed.*, 2008, **47**, 5340–5343.
71 S. Grimme, C. Bannwarth and P. Shushkov, A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86), *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
72 C. Bannwarth, S. Ehlert and S. Grimme, GFN2-xTB--An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
73 T. Cormen, C. Leiserson, R. Rivest and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge MA, 2nd edn, 2001, ch. 22.

74 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *TensorFlow Large-Scale Machine Learning on Heterogeneous Distributed Systems*, 2016.

75 B. Pang, E. Nijkamp and Y. N. Wu, Deep Learning With TensorFlow: A Review, *J. Educ. Behav. Stat.*, 2019, **45**, 227–248.

76 A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, Inc., Beijing, 2nd edn, 2022.

77 V. Nair and G. E. Hinton, Rectified linear units improve restricted Boltzmann machines, *Proceedings of the 27th International Conference on International Conference on Machine Learning*, Madison, WI, USA, 2010, pp. 807–814.

78 D. P. Kingma and J. Ba, Adam: A Method for Stochastic Optimization, *arXiv*, 2017, preprint, arXiv:1412.6980, DOI: **10.48550/arXiv.1412.6980**.

79 Semiempirical Extended Tight-Binding Program Package. Sep 17, 2020; **https://github.com/grimme-lab/xtb/tree/v6.3.3**.

80 C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher and S. Grimme, Extended tight-binding quantum chemistry methods, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2020, **11**, e01493.

81 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. V. 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. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. 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 16 Rev. B.01*, Wallingford, CT, 2016.

82 Y. Zhao and D. G. Truhlar, The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals, *Theor. Chem. Acc.*, 2008, **120**, 215–241.

83 F. Weigend and R. Ahlrichs, 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.

84 F. Weigend, Accurate Coulomb-fitting basis sets for H to Rn, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.

85 P. A. Jensen, M. Leccese, F. D. S. Simonsen, A. W. Skov, M. Bonfanti, J. D. Thrower, R. Martinazzo and L. Hornekær, Identification of stable configurations in the superhydrogenation sequence of polycyclic aromatic hydrocarbon molecules, *Mon. Not. R. Astron. Soc.*, 2019, **486**, 5492–5498.

86 S. D. Wiersma, A. Candian, J. M. Bakker, J. Martens, G. Berden, J. Oomens, W. J. Buma and A. Petrignani, Photolysis-induced scrambling of PAHs as a mechanism for deuterium storage, *Astron. Astrophys.*, 2020, **635**, A9.

87 D. Campisi, F. D. S. Simonsen, J. D. Thrower, R. Jaganathan, L. Hornekær, R. Martinazzo and A. G. G. M. Tielens, Superhydrogenation of pentacene: the reactivity of zigzag-edges, *Phys. Chem. Chem. Phys.*, 2020, **22**, 1557–1565.

88 P. Pla, Y. Wang and M. Alcamí, Simple bond patterns predict the stability of Diels-Alder adducts of empty fullerenes, *Chem. Commun.*, 2018, **54**, 4156–4159.

89 E. Campbell, P. Fowler, D. Mitchell and F. Zerbetto, Increasing cost of pentagon adjacency for larger fullerenes, *Chem. Phys. Lett.*, 1996, **250**, 544–548.

90 E. Albertazzi, C. Domene, P. W. Fowler, T. Heine, G. Seifert, C. Van Alsenoy and F. Zerbetto, Pentagon adjacency as a determinant of fullerene stability, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2913–2918.

91 S. Wang, Q. Chang, G. Zhang, F. Li., X. Wang, S. Yang and S. Troyanov, Structural Studies of Giant Empty and Endohedral Fullerenes, *Front. Chem.*, 2020, **8**, 607712.

92 H. W. Kroto, The stability of the fullerenes C_n, with n = 24, 28, 32, 36, 50, 60 and 70, *Nature*, 1987, **329**, 529–531.

93 O. V. Boltalina, A. A. Popov, I. V. Kuvychko, N. B. Shustova and S. H. Strauss, Perfluoroalkylfullerenes, *Chem. Rev.*, 2015, **115**, 1051–1105.

94 M. Chen, L. Bao, M. Ai, W. Shen and X. Lu, $Sc_3N@I_h$-C_{80} as a novel Lewis acid to trap abnormal N-heterocyclic carbenes: the unprecedented formation of a singly bonded [6,6,6]-adduct, *Chem. Sci.*, 2016, **7**, 2331–2334.

95 M. A. Addicoat, A. J. Page, Z. E. Brain, L. Flack, K. Morokuma and S. Irle, Optimization of a Genetic Algorithm for the Functionalization of Fullerenes, *J. Chem. Theory Comput.*, 2012, **8**, 1841–1851.

96 Y. Wang, The FullFun package, 2017, **https://campusys.qui.uam.es/?page_id=1491**.

97 *MATLAB version: 9.14.0.2206163 (R2023a)*, The MathWorks Inc., 2023, **https://www.mathworks.com**.

98 *Jmol: an open-source Java viewer for chemical structures in 3D (version 14.30.2)*, 2015, **https://www.jmol.org/**.