Supplementary Information for: Transferable Machine Learning Interatomic Potential for Carbon Hydrogen Systems

Somayeh Faraji[1] and Mingjie Liu[1*]

[1]Department of Chemistry, University of Florida, Gainesville, FL 32611, United States.

[*]mingjieliu@ufl.edu

August 6, 2024

1 Impact of training data energy span on potential accuracy

Despite increasing the training dataset size, we observed minimal improvement in the accuracy of the trained potential during the last training cycles. We hypothesized that this could be attributed to the complexity and significant diversity in energy among the data points. Our dataset displayed a wide range of energy differences between structures, ranging from lower to higher energy values, with a magnitude of 4.52 eV/atom. This broad spectrum of energy, coupled with variations in the ratios of C and H atoms within the structures, poses a formidable and complex challenge for the training process. Regarding this complexity, we conducted multiple training sessions at different energy range values (ΔE), including 1.0, 1.5, 2.0, 2.5, and 3.0 eV/atom above the minimum energy in the final dataset. Figure S1 displays the configuration analysis of data across each of these energy ranges. For each ΔE, we evaluated the RMSE of energy and forces in the training and validation sets for the best-selected potentials. As depicted in Figures S2 and S3), the potentials constructed with ΔE set to 1.0, 1.5, and 2.5 eV/atom exhibited overfitting, evidenced by larger errors in the validation than in the training sets. However, for ΔE = 2.0 eV/atom, the RMSE values were lower compared to the previous ranges, with instances of overfitting but also cases without it. For ΔE = 2.5 eV/atom, overfitting was again observed. On the contrary, ΔE = 3.0 eV/atom showed no overfitting (i.e. the error of valid is smaller than that of training), but yielded higher RMSE values. Therefore, we found that the training dataset within the energy range of 2.0 eV/atom produced more accurate potentials, thereby becoming our focus.

Focusing on this energy range revealed outlier structures both in training and validation datasets, resulting large errors in the produced potentials. we visualized the outliers and noticed that they were structures with atoms that had no bonds with each other. By excluding these outliers from

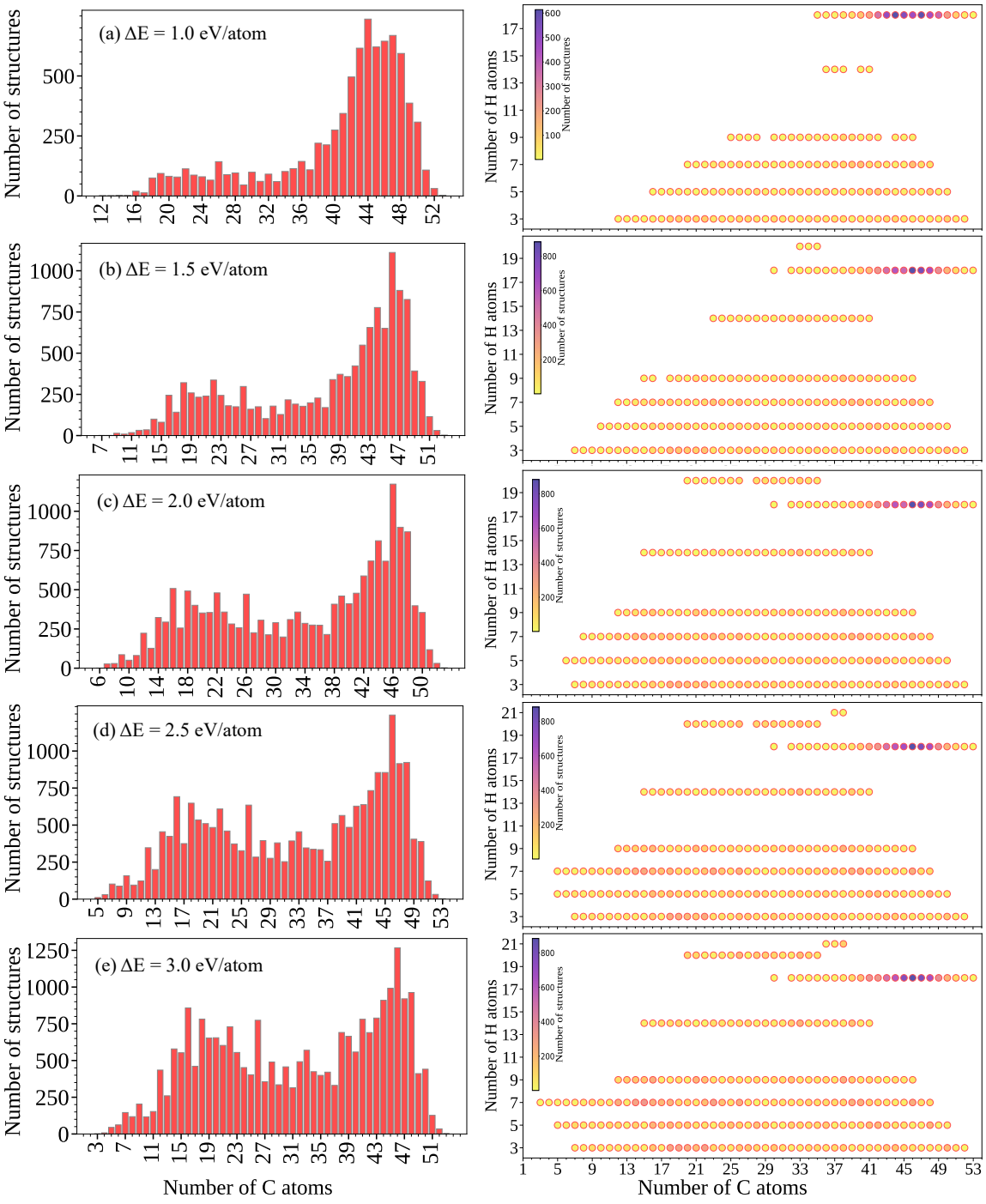

Figure S1: The configuration analysis of data at different energy ranges tested for selecting the energy changes for train. The first column illustrates the distribution of structures based on the count of carbon (C) atoms at different energy ranges (ΔE), ranging from 1.0 to 3.0 eV/atom. The second column displays the scatter plots showcase the frequency of structures for distinct C-H atom ratios, depicted by a colormap (color intensity) representing the count of structures within each ratio. The color bar provides a quantitative representation of the structure count for the observed C-H compositions.

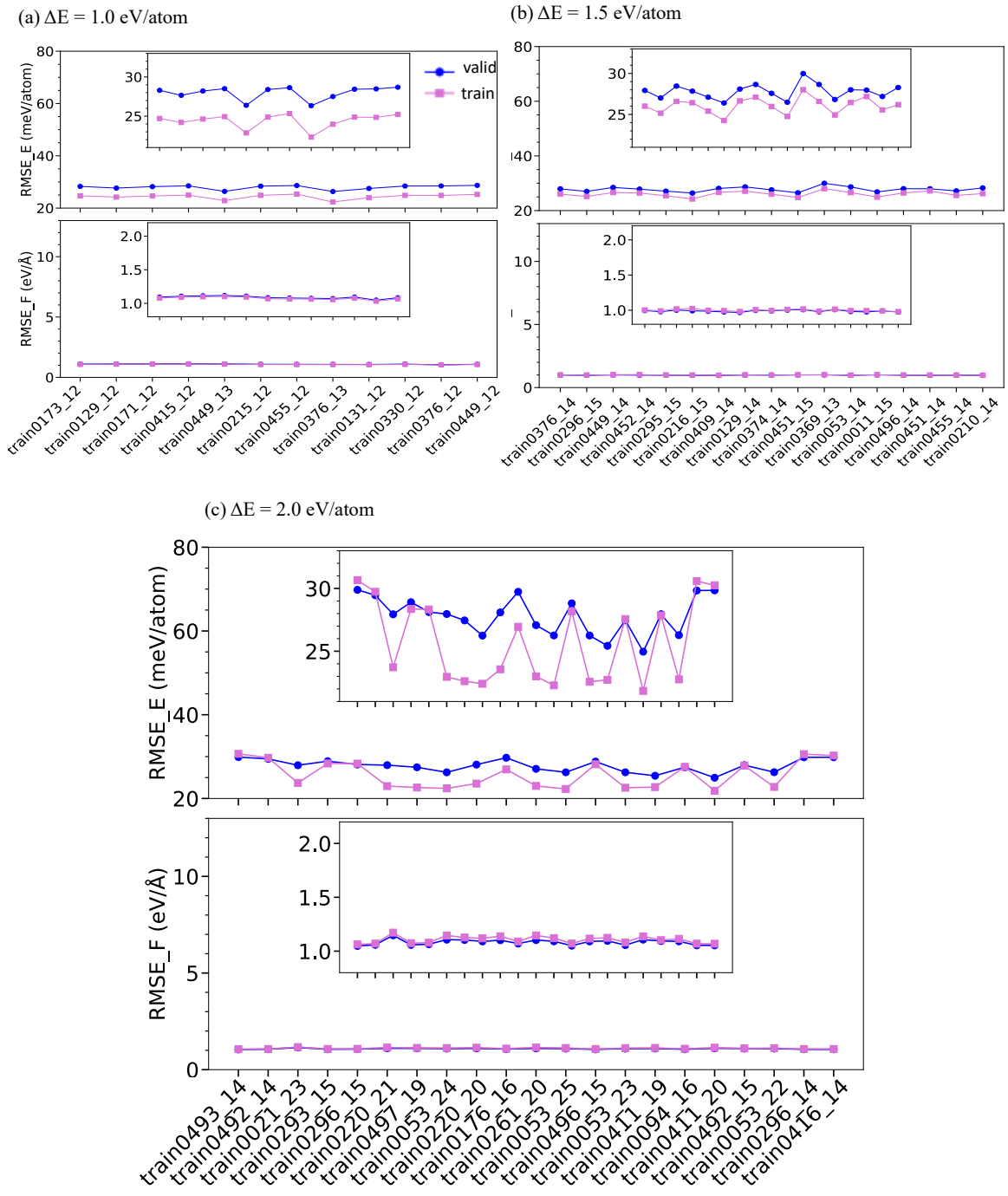

Figure S2: The RMSE of energy and forces of different trial potentials at different energy ranges (part I).

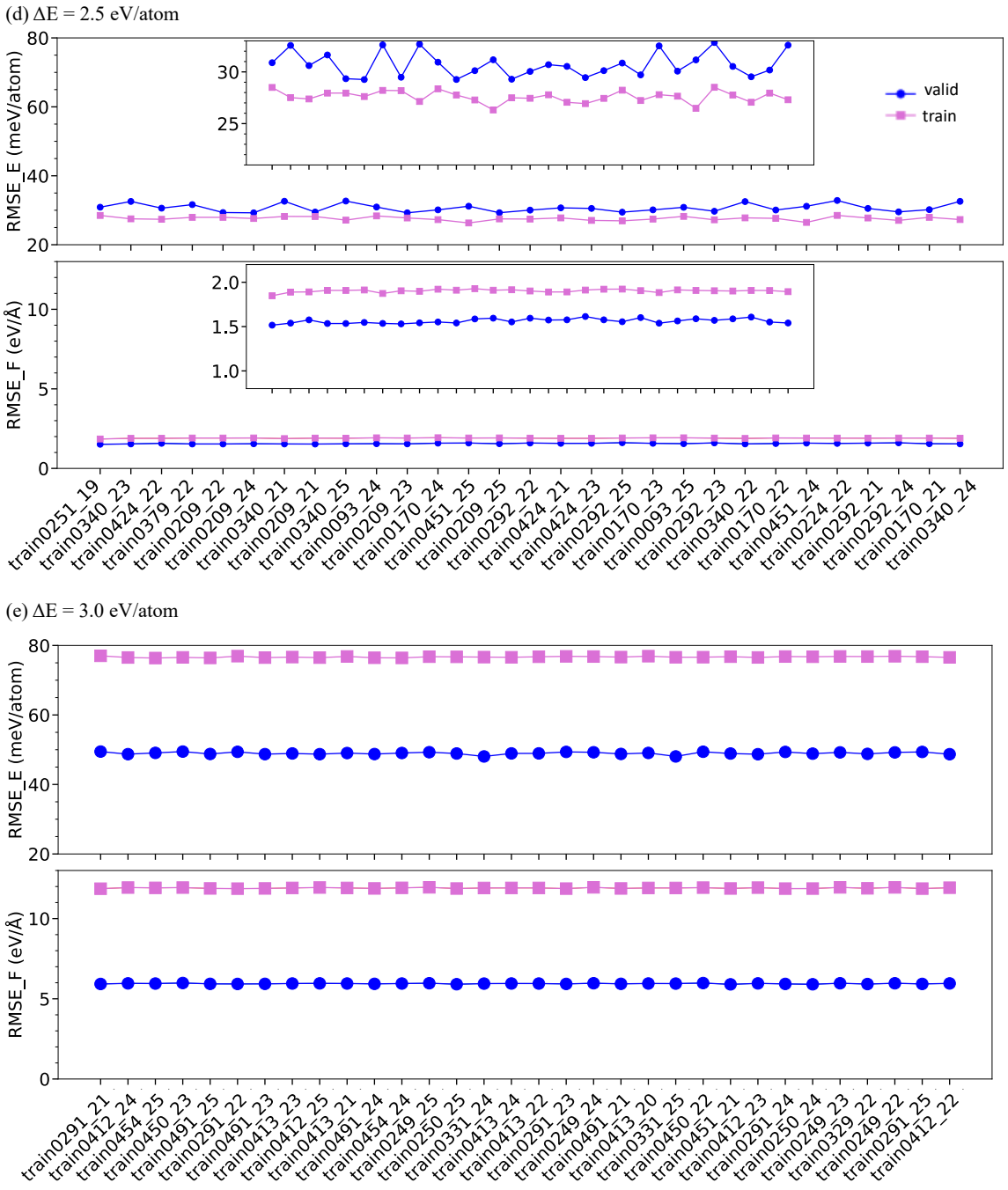

Figure S3: The RMSE of energy and forces of different trial potentials at different energy ranges (part2).

both the training and validation datasets, we achieved a reduction in large errors in the produced potentials. Figures S2-c and S4 illustrate the reduction in RMSEs before and after excluding these outliers from both training and validation.

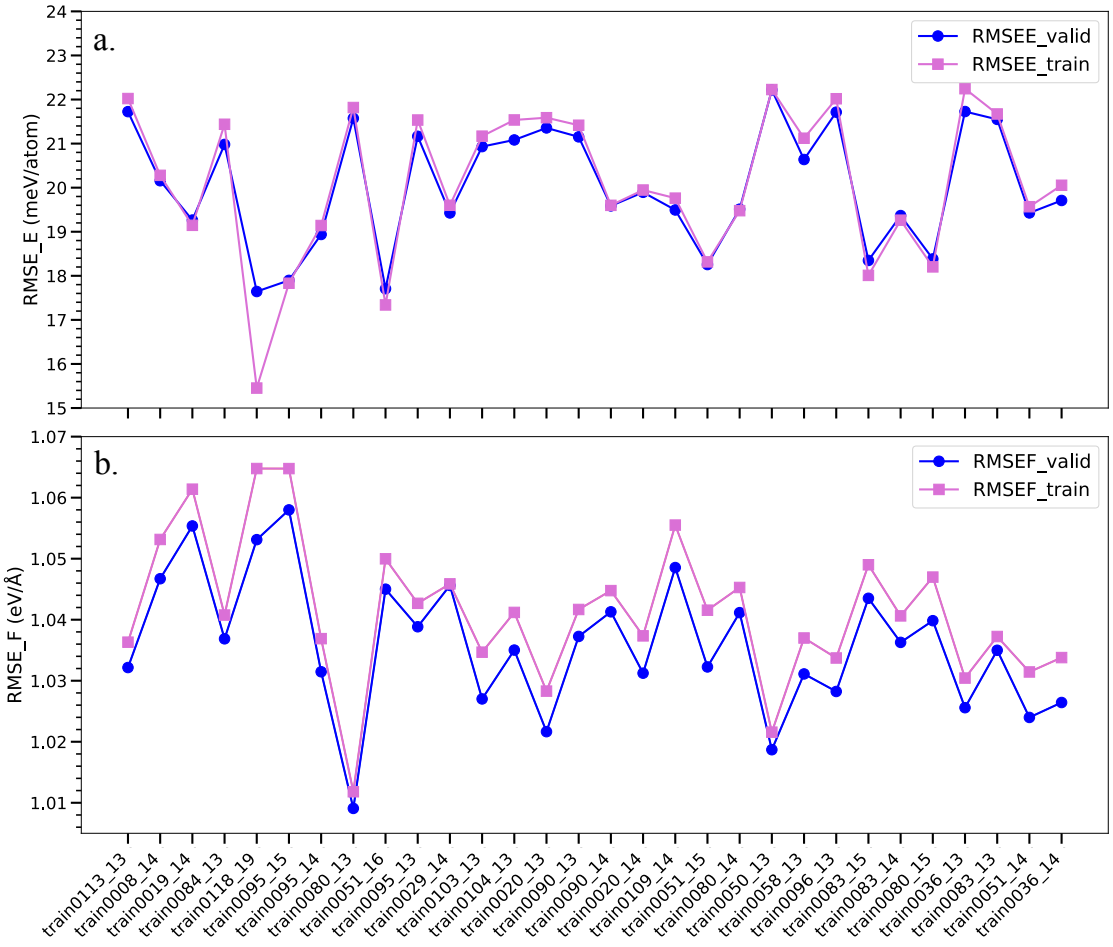

Figure S4: The RMSE of energy and forces of the selected potentials at energy range of 2.0 eV/atom.

Figure S5 illustrates the distribution of energy contributions across the dataset within the specified energy window of 2.0 eV/atom. This visual representation provides insights into how the energy values are spread across the range.

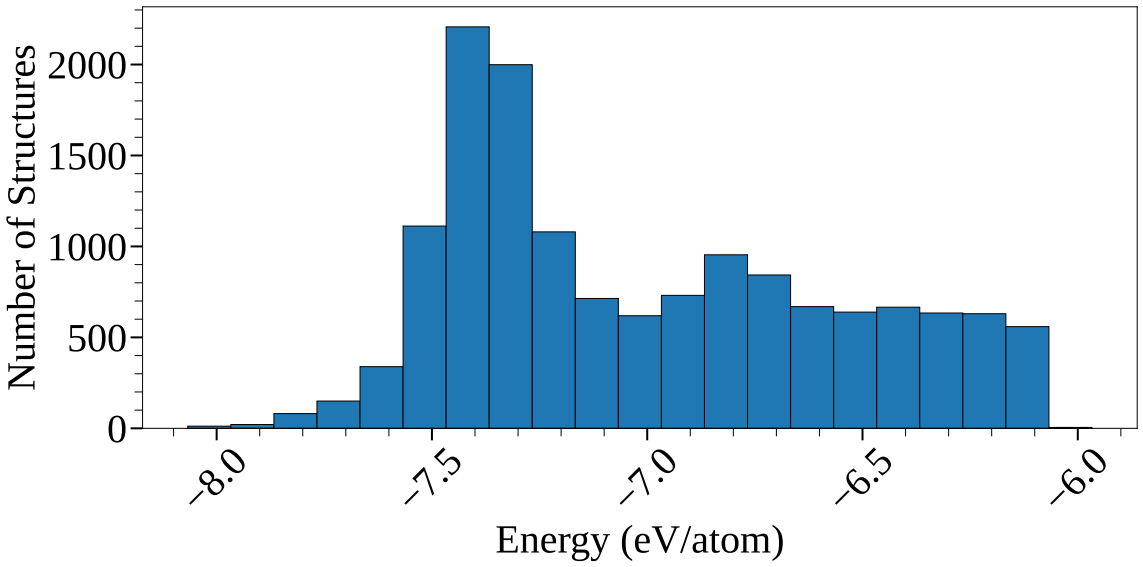

Figure S5: The energy contribution of dataset at energy windows of 2.0 eV/atom.

2 Impact of spin considerations on open-shell systems

As mentioned in the method section, the initial dataset was obtained by creating $1-24$ C vacancies in $C_{52}H_{18}$ zigzag flake. Therefore, all data initially had an even number of H atoms. To make the dataset more diverse, we added datapoints with an odd number of H atoms (i.e. H_x, x = 3, 5, 7, 9) and considered different numbers of C atoms for each odd number of H atoms during the data augmentation process. This resulted in having clusters with an odd number of H atoms, which are open-shell systems requiring spin-polarized calculations.

To assess whether the non-spin-polarized calculations can correctly compute the energies for these open-shell systems, we selected sample systems with 3, 5, 7 and 9 H atoms, along with varying odd and even numbers of C atoms and performed single-point calculations to compare the results of energies and forces. The results of spin- and non-spin-polarized DFT calculations with the PBE functional on these systems are summarized in Table S1. The results show that the total energies obtained from spin-polarized calculations are lower than those from non-spin-polarized calculations, with an MAE of total energy of 0.42 eV. However, it results in an MAE of 0.0188 eV/atom in E_f, which is smaller than the error in the ANN potential (0.02 eV/atom). Additionally, this test shows that the maximum RMSE of force components is 0.53 eV/Å, which is also a small value compared to the errors in our training and validation. Therefore, non-spin-polarized calculations of the open-shell systems do not introduce considerable error in the reference energies and forces of the dataset.

Table S1: ΔE_{tot} is the $E_{tot,spin}$ – $E_{tot,nonspin}$, MAE_F and mAE_F are the maximum and minimum absolute error in forces from spin and nonspin polarized single-point calculations, respectively, $RMSE_F$ is the root mean absolute error of forces, E_f is the formation energy, ΔE_f is the energy difference in formation energy obtained from spin and nonspin polarized calculations.

	ΔE_{tot} (eV)	MAE_F (eV/Å)	mAE_F (eV/Å)	$RMSE_F$ (eV/Å)	$E_{f,spin}$ (eV/atom)	$E_{f,nonspin}$ (eV/atom)	ΔE_f (eV/atom)
H07C18	-0.22551	0.371842	0.004620	0.140027	-11.72947	-11.72045	-0.00902
H07C19	-0.16484	0.731991	0.012878	0.249873	-11.96006	-11.95372	-0.00634
H07C42	-0.03888	1.250212	0.034761	0.503975	-13.80533	-13.80453	-0.00080
H07C43	-0.22167	1.533610	0.013094	0.398164	-14.27994	-14.27551	-0.00443
H03C09	-0.22580	0.327090	0.003396	0.172941	-12.46157	-12.44275	-0.01882
H03C18	-0.12670	0.560343	0.010566	0.186549	-14.18570	-14.17966	-0.00604
H03C27	-0.14312	0.303562	0.025608	0.137181	-14.00647	-14.00170	-0.00477
H03C42	-0.11255	0.275119	0.004444	0.086380	-15.42721	-15.42471	-0.00250
H05C12	-0.24764	0.417436	0.006959	0.188259	-11.72844	-11.71387	-0.01457
H05C29	-0.10720	0.145316	0.007962	0.073648	-14.21146	-14.20831	-0.00315
H05C41	-0.22839	1.191151	0.011326	0.437105	-14.77498	-14.77002	-0.00496
H09C22	-0.18302	1.375544	0.028908	0.496279	-12.07224	-12.06634	-0.00590
H09C23	-0.31419	0.429440	0.002054	0.141359	-12.41900	-12.40918	-0.00982
H09C38	-0.42436	0.822324	0.003721	0.232690	-13.61539	-13.60636	-0.00903
H09C46	-0.30921	1.583654	0.016955	0.396462	-14.12526	-14.11964	-0.00562

To investigate the smoothness of PES in the transition from closed- to open-shell systems, we conducted two tests. Firstly, we examined the C-C bond dissociation in C_2H_2. As shown in Figure S6-(a), using spin polarized wavefunctions in the calculations did not change the equilibrium C-C bond length and energy. The energy difference begins to diverge from the non-spin calculations at distances greater than 2 Å. At larger distances, ΔE_{tot} reaches up to 1.19 eV, with the spin-polarized energy being more favorable. However, there is no abrupt change or discontinuous in the transition from close to open shell systems. Secondly, we focused on the geometry of C_7H_5 and performed geometry optimization using PBE and PBE0, each with and without including spin-polarized calculations. Then, one H atom was removed from the optimized geometry and the remaining geometry was optimized like the previous step. This step was repeated until all the H atoms were removed. Figure S6-(b) shows the total energy versus the number of removed H atoms. As depicted in the figure, there is a smooth transition between closed- and open-shell systems. The absence of sharp changes or discontinuities indicates that the PBE can handle both states effectively. Additionally, the geometries obtained from PBE without spin polarization (blue cross) and those from PBE0 with spin polarization (green circle) shown in the figure are similar. Comparing the absolute values of energies reveals that both PBE and PBE0 exhibit a similar slope, with a noticeable shift in the absolute values, indicating that the energies obtained from PBE are reliable.

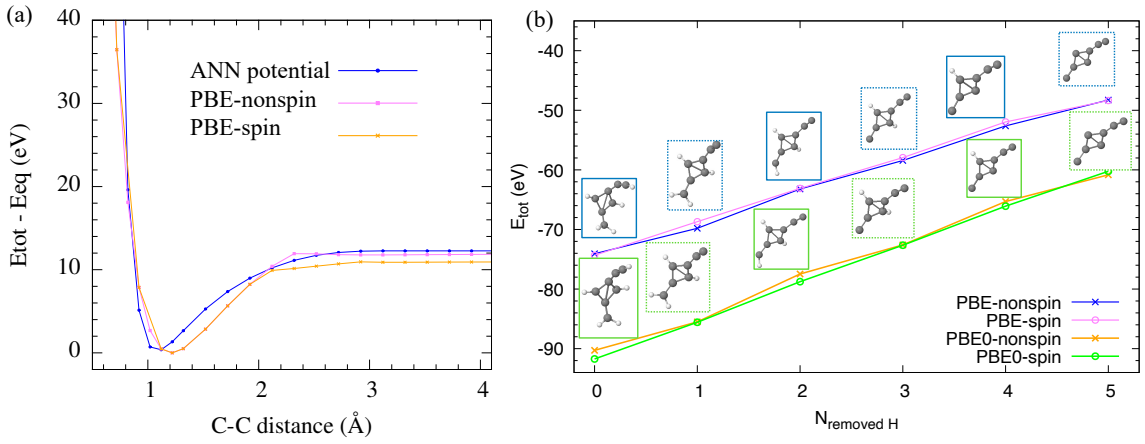

Figure S6: (a) Potential energy curves for C_2H_2 calculated by the ANN potential, PBE functional without and with including spin. (b) Total energies obtained by PBE and PBE0 functionals without and with including spin versus the number of removed H atoms from C_7H_5. The optimized geometries in the blue and green boxes show the results from PBE without spin and PBE0 with spin, respectively.

3 0D test cases

Table S2: Chemical name, number of H (n_H) and C (n_C) atoms in the geometry, total energies (E_{tot}), formation energies (E_f), and formation energy differences of DFT and the ANN potential ($E_{f,diff}$) of the examined 0D test cases. All the energies are in eV/atom.

ID	system	n_H	n_C	$E_{tot,DFT}$	$E_{tot,ANN}$	$E_{f,DFT}$	$E_{f,ANN}$	$E_{f,diff}$
01	methane	4	1	-4.80779	-4.83077	-0.280812	-0.325548	0.022971
02	ethane	6	2	-5.06129	-5.05908	-0.24895	-0.24675	-0.00221
03	cyclopropane	6	3	-5.38114	-5.44119	-0.09321	-0.15326	0.06005
04	propane	8	3	-5.18301	-5.17456	-0.24097	-0.23251	-0.00845
05	cyclobutane	8	4	-5.41396	-5.46376	-0.12603	-0.17583	0.04980
06	butane	10	4	-5.25342	-5.24421	-0.23726	-0.22805	-0.00921
07	isobutane	10	4	-5.25624	-5.23825	-0.24008	-0.22209	-0.01799
08	propellane	6	5	-5.79979	-5.66478	0.17992	0.31492	-0.13501
09	cyclopentane	10	5	-5.48841	-5.50076	-0.20048	-0.21283	0.01235
10	dimethylcyclopropane-c	10	5	-5.43828	-5.45257	-0.15035	-0.16464	0.01429
11	dimethylcyclopropane-t	10	5	-5.44229	-5.44797	-0.15436	-0.16004	0.00569
12	methylbutane	12	5	-5.29781	-5.28984	-0.23369	-0.22572	-0.00797
13	pentane	12	5	-5.29449	-5.28881	-0.23037	-0.22468	-0.00568
14	cyclohexane-boat	12	6	-5.48763	-5.49273	-0.19970	-0.20480	0.00510
15	cyclohexane-chair	12	6	-5.50365	-5.49537	-0.21572	-0.20744	-0.00828
16	hexane	14	6	-5.32871	-5.32045	-0.23101	-0.22275	-0.00826
17	norbornane	12	7	-5.66376	-5.67171	-0.17558	-0.18353	0.00795

ID	system	n_H	n_C	$E_{tot,DFT}$	$E_{tot,ANN}$	$E_{f,DFT}$	$E_{f,ANN}$	$E_{f,diff}$
18	cycloheptane	14	7	-5.49117	-5.47987	-0.20324	-0.19194	-0.01130
19	heptane	16	7	-5.35278	-5.34379	-0.23027	-0.22128	-0.00899
20	cyclooctane	16	8	-5.48280	-5.46976	-0.19487	-0.18183	-0.01304
21	isooctane	18	8	-5.36537	-5.36285	-0.22377	-0.22126	-0.00251
22	octane	18	8	-5.36901	-5.36175	-0.22742	-0.22016	-0.00726
23	spirononane	16	9	-5.62866	-5.64692	-0.18854	-0.20680	0.01826
24	cyclononane	18	9	-5.48065	-5.46853	-0.19272	-0.18060	-0.01213
25	nonane	20	9	-5.38289	-5.37600	-0.22616	-0.21927	-0.00689
26	adamantane	16	10	-5.77559	-5.77587	-0.19499	-0.19527	0.00028
27	decalin-cis	18	10	-5.62940	-5.61818	-0.20559	-0.19437	-0.01122
28	decalin-trans	18	10	-5.63438	-5.63268	-0.21057	-0.20886	-0.00171
29	cyclodecane	20	10	-5.48136	-5.47265	-0.19343	-0.18472	-0.00871
30	decane	22	10	-5.39417	-5.38757	-0.22514	-0.21854	-0.00660
31	cetane	34	16	-5.43340	-5.42785	-0.22157	-0.21602	-0.00555
32	dodecahedrane	20	20	-6.36100	-6.39322	-0.12189	-0.15411	0.03222
33	anthracene	10	14	-6.79043	-6.83212	-0.07573	-0.11742	0.04169
34	benzene	6	6	-6.33352	-6.38190	-0.09440	-0.14279	0.04839
35	benzopyrene	12	20	-7.03310	-7.07485	-0.08060	-0.12234	0.04175
36	buckminsterfullerene	0	60	-8.85075	-8.83164	0.24192	0.26103	-0.01911
37	butadiene-13	6	4	-5.68233	-5.69387	-0.01392	-0.02547	0.01154
38	butene-1	8	4	-5.42971	-5.42078	-0.14178	-0.13285	-0.00893
39	butene-2c	8	4	-5.43878	-5.46266	-0.15085	-0.17473	0.02388
40	butene-2t	8	4	-5.44292	-5.44838	-0.15499	-0.16046	0.00547
41	butyne-1	6	4	-5.63181	-5.59669	0.03660	0.07172	-0.03512
42	butyne-2	6	4	-5.66271	-5.62300	0.00570	0.04541	-0.03971
43	C_{70}	0	70	-8.88897	-8.88035	0.20370	0.21232	-0.00862
44	camphene	16	10	-5.72713	-5.74576	-0.14653	-0.16516	0.01862
45	caryophyllene	24	15	-5.70729	-5.73241	-0.12669	-0.15181	0.02512
46	cyclobutadiene	4	4	-5.82136	-6.09382	0.41775	0.14530	0.27245
47	cyclobutene	6	4	-5.65483	-5.72055	0.01357	-0.05214	0.06571
48	cyclohexene	10	6	-5.65311	-5.69548	-0.12738	-0.16975	0.04237
49	cyclooctatetraene	8	8	-6.14917	-6.29209	0.08994	-0.05297	0.14291
50	cyclooctene-cis	14	8	-5.60596	-5.61822	-0.14509	-0.15734	0.01225
51	cyclooctene-trans	14	8	-5.58386	-5.60775	-0.12299	-0.14687	0.02388
52	cyclopentene	8	5	-5.70147	-5.74831	-0.12087	-0.16771	0.04684
53	cyclopropene	4	3	-5.56636	-5.75567	0.26511	0.07580	0.18931
54	dibenzanthracene	14	22	-6.95236	-6.99128	-0.07912	-0.11804	0.03892
55	ethene	4	2	-5.32868	-5.30135	-0.04075	-0.01342	-0.02733
56	ethyne	2	2	-5.73412	-5.97970	0.50500	0.25942	0.24558
57	farnesene	24	15	-5.70273	-5.72768	-0.12213	-0.14708	0.02495

ID	system	n_H	n_C	$E_{tot,DFT}$	$E_{tot,ANN}$	$E_{f,DFT}$	$E_{f,ANN}$	$E_{f,diff}$
58	isoprene	8	5	-5.66527	-5.66707	-0.08467	-0.08647	0.00180
59	limonene	16	10	-5.72337	-5.74715	-0.14277	-0.16654	0.02378
60	lycopene	56	40	-5.87820	-5.90102	-0.11468	-0.13749	0.02281
61	muscalure	46	23	-5.49219	-5.48873	-0.20426	-0.20080	-0.00346
62	myrcene	16	10	-5.68687	-5.70678	-0.10627	-0.12618	0.01992
63	naphthalene	8	10	-6.64218	-6.68297	-0.08600	-0.12679	0.04079
64	phellandrene-a	16	10	-5.72407	-5.73884	-0.14347	-0.15824	0.01477
65	phellandrene-b	16	10	-5.72288	-5.76458	-0.14227	-0.18398	0.04171
66	phenanthrene	10	14	-6.79829	-6.83943	-0.08358	-0.12473	0.04114
67	pinene-a	16	10	-5.70997	-5.76748	-0.12937	-0.18688	0.05751
68	pinene-b	16	10	-5.70421	-5.74235	-0.12361	-0.16174	0.03814
69	propadiene-12	4	3	-5.70084	-5.67288	0.13062	0.15858	-0.02796
70	propene	6	3	-5.40625	-5.40136	-0.11832	-0.11343	-0.00489
71	pyrene	10	16	-6.98371	-7.02201	-0.08608	-0.12438	0.03830
72	selinene	24	15	-5.73372	-5.75062	-0.15312	-0.17002	0.01690
73	styrene	8	8	-6.30484	-6.33580	-0.06573	-0.09669	0.03096
74	thujene	16	10	-5.70572	-5.73221	-0.12511	-0.15160	0.02649
75	toluene	8	7	-6.17476	-6.21434	-0.12588	-0.16546	0.03958
76	xylene-m	10	8	-6.06909	-6.10267	-0.14704	-0.18061	0.03357
77	xylene-o	10	8	-6.06902	-6.10645	-0.14697	-0.18439	0.03742
78	xylene-p	10	8	-6.06856	-6.10224	-0.14651	-0.18019	0.03368
79	C_{11}-008	0	11	-7.80387	-7.88095	1.28880	1.21172	0.07708
80	C_{32}-255	0	32	-8.49852	-8.41582	0.59415	0.67685	-0.08270
81	C_{46}-377	0	46	-7.98483	-8.06876	1.10784	1.02391	0.08393
82	C_{48}-412	0	48	-8.47599	-8.45829	0.61668	0.63438	-0.01770
83	C_{60}-isomer01	0	60	-8.85062	-8.83164	0.24205	0.26103	-0.01897
84	C_{60}-isomer02	0	60	-8.82463	-8.81694	0.26804	0.27573	-0.00769
85	C_{60}-isomer03	0	60	-8.81194	-8.81028	0.28072	0.28239	-0.00166
86	C_{60}-isomer04	0	60	-8.81220	-8.80985	0.28047	0.28282	-0.00235
87	C_{60}-isomer05	0	60	-8.80369	-8.81048	0.28898	0.28219	0.00679

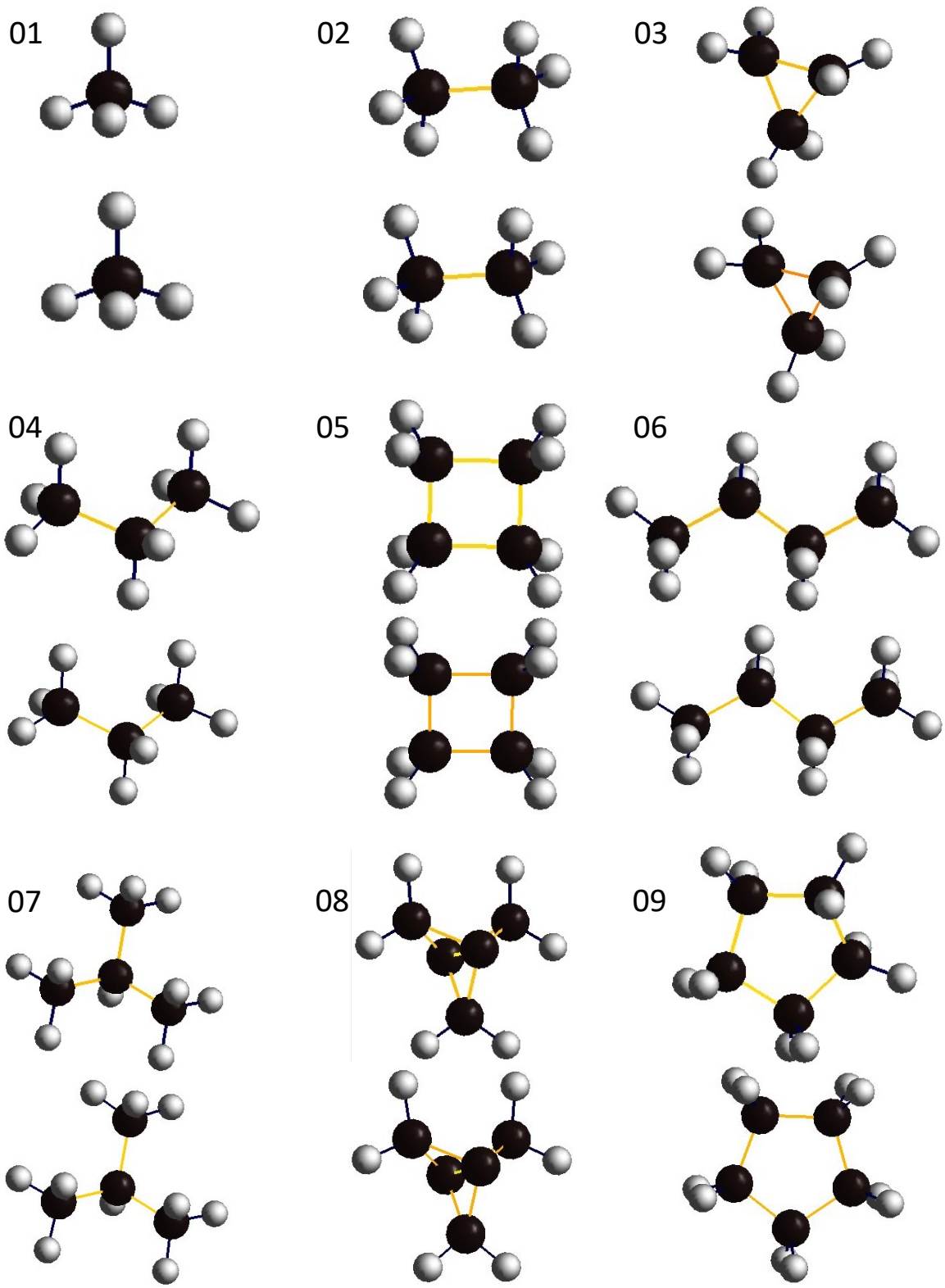

Figure S7: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure. (part 1)

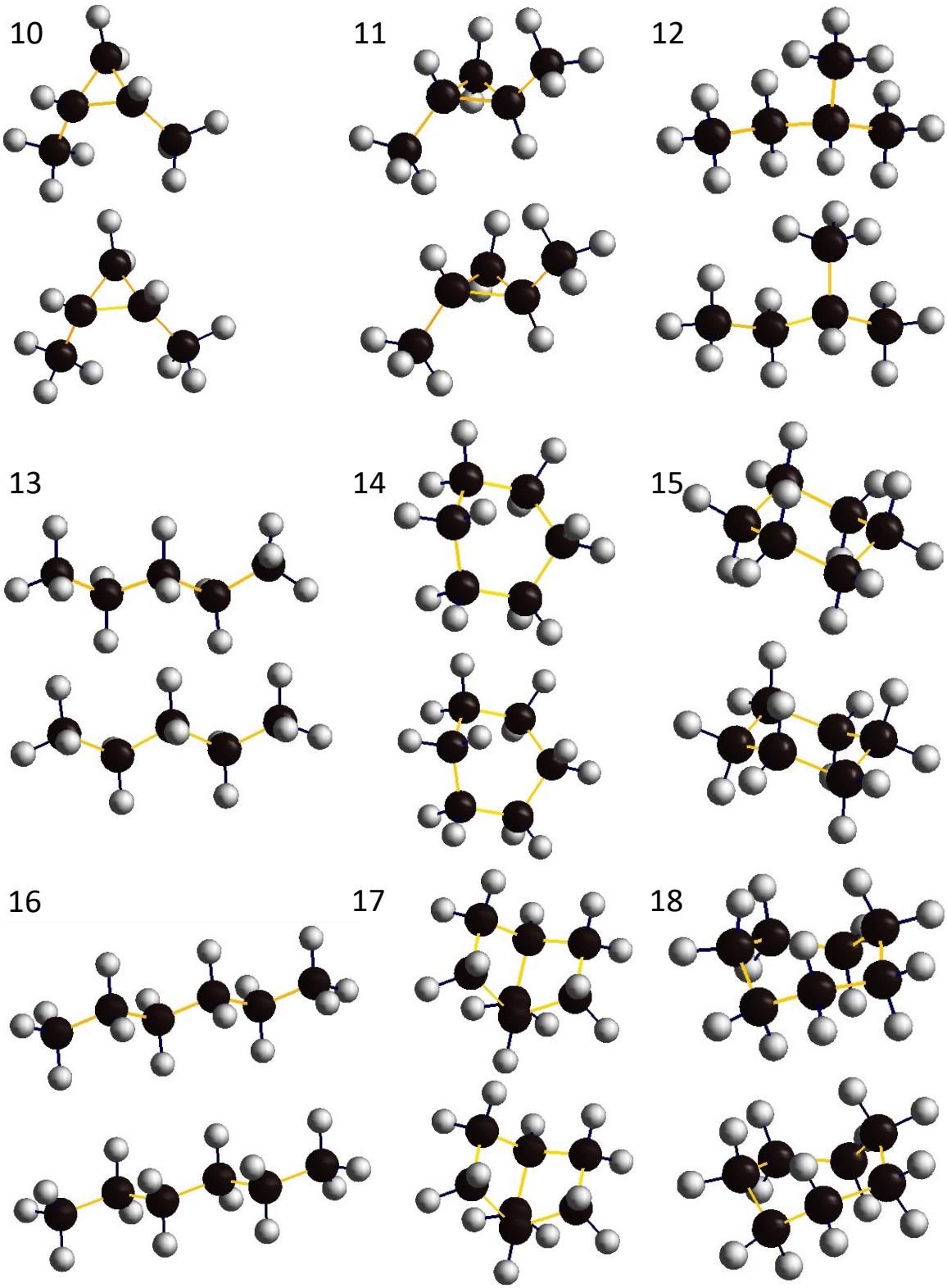

Figure S8: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 2).

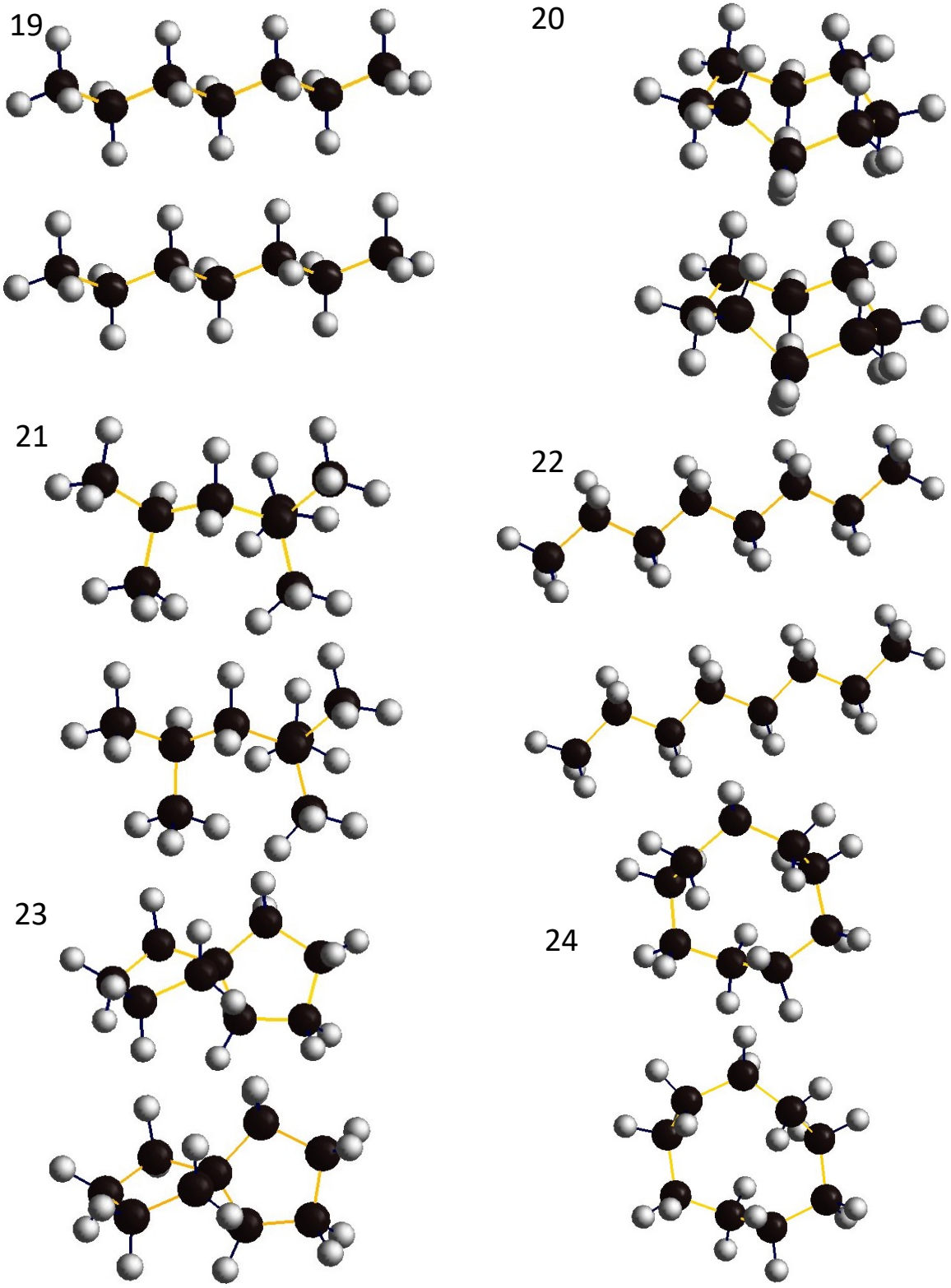

Figure S9: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 3).

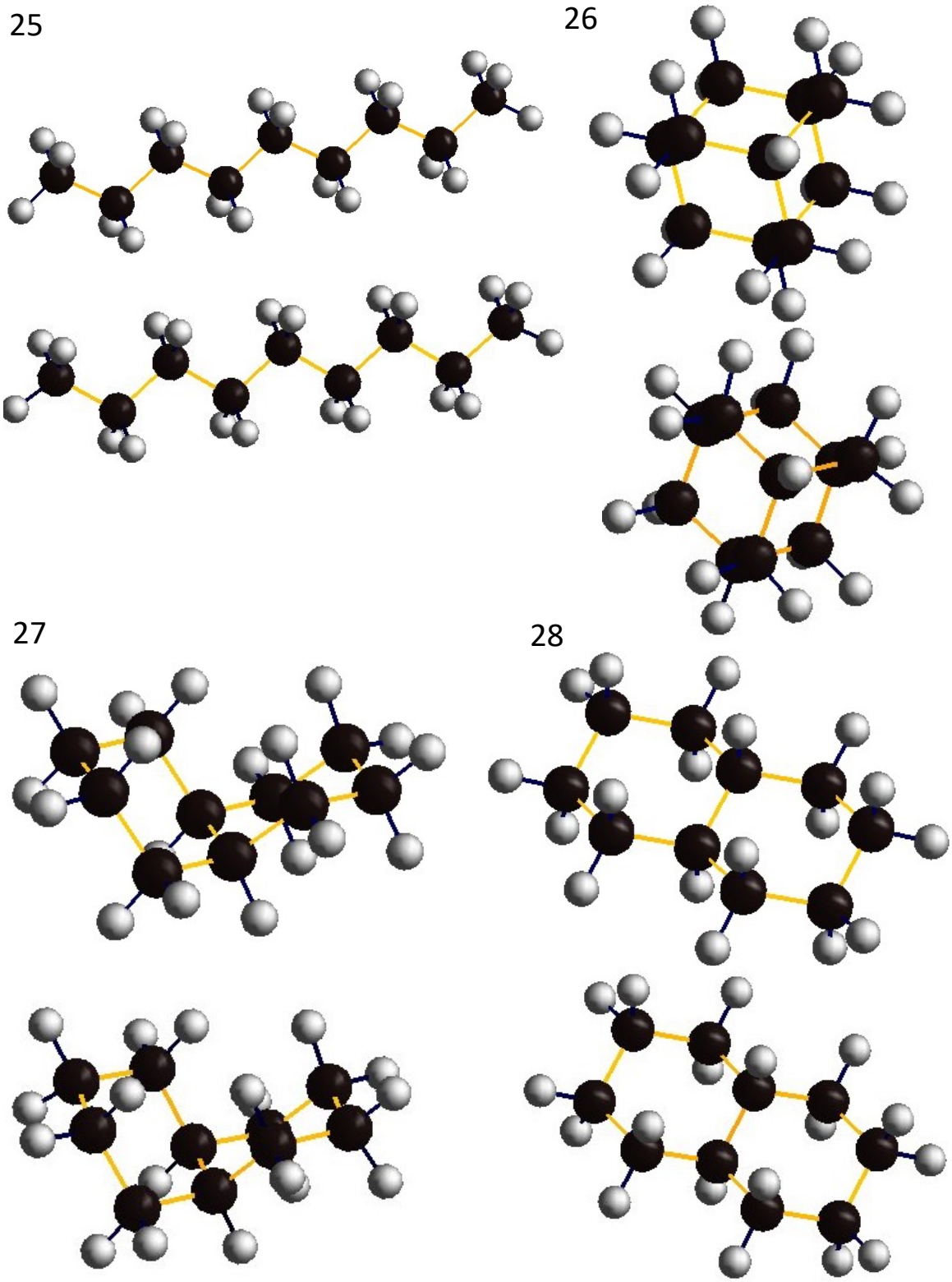

Figure S10: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 4).

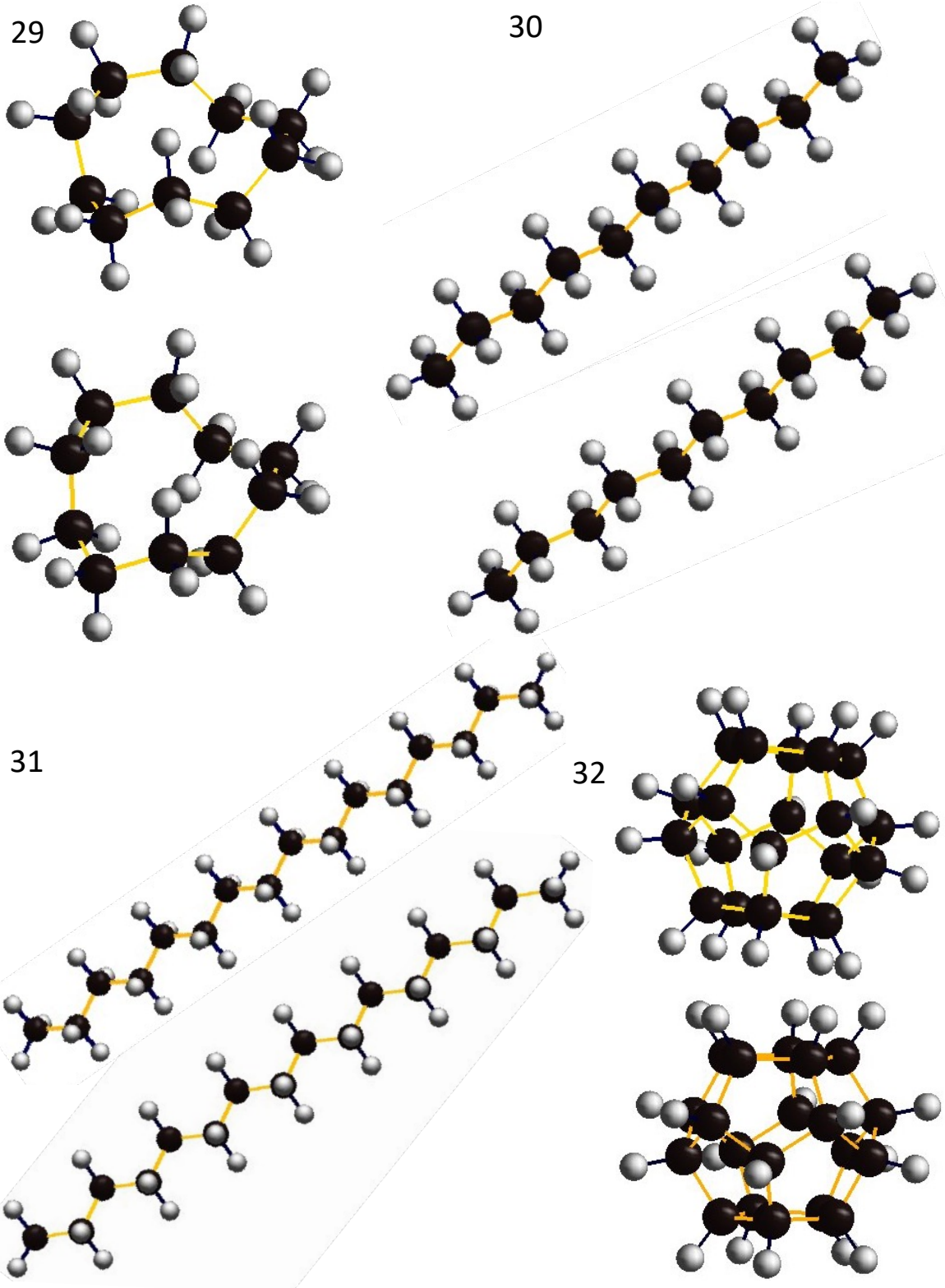

Figure S11: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 5).

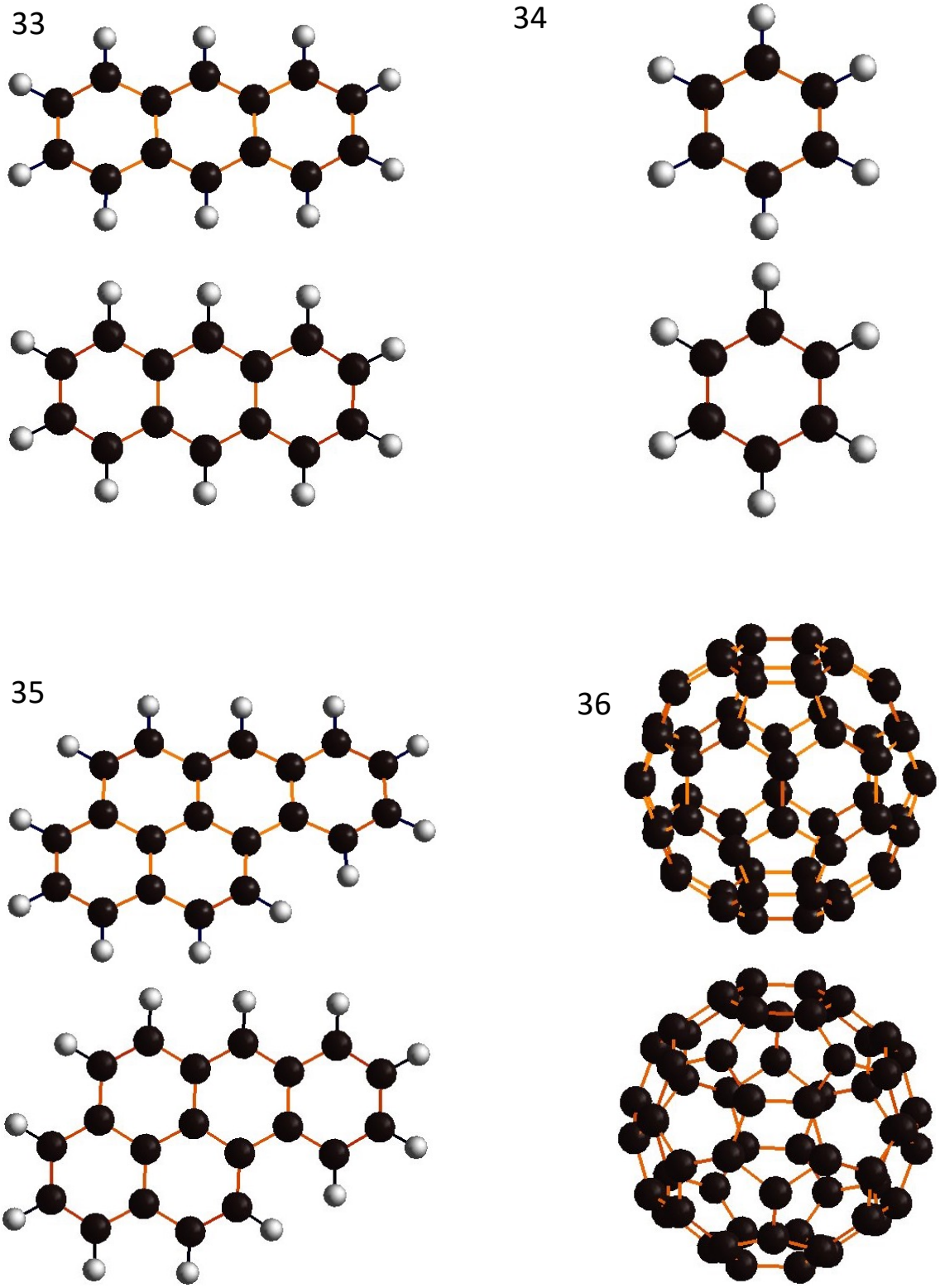

Figure S12: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 6).

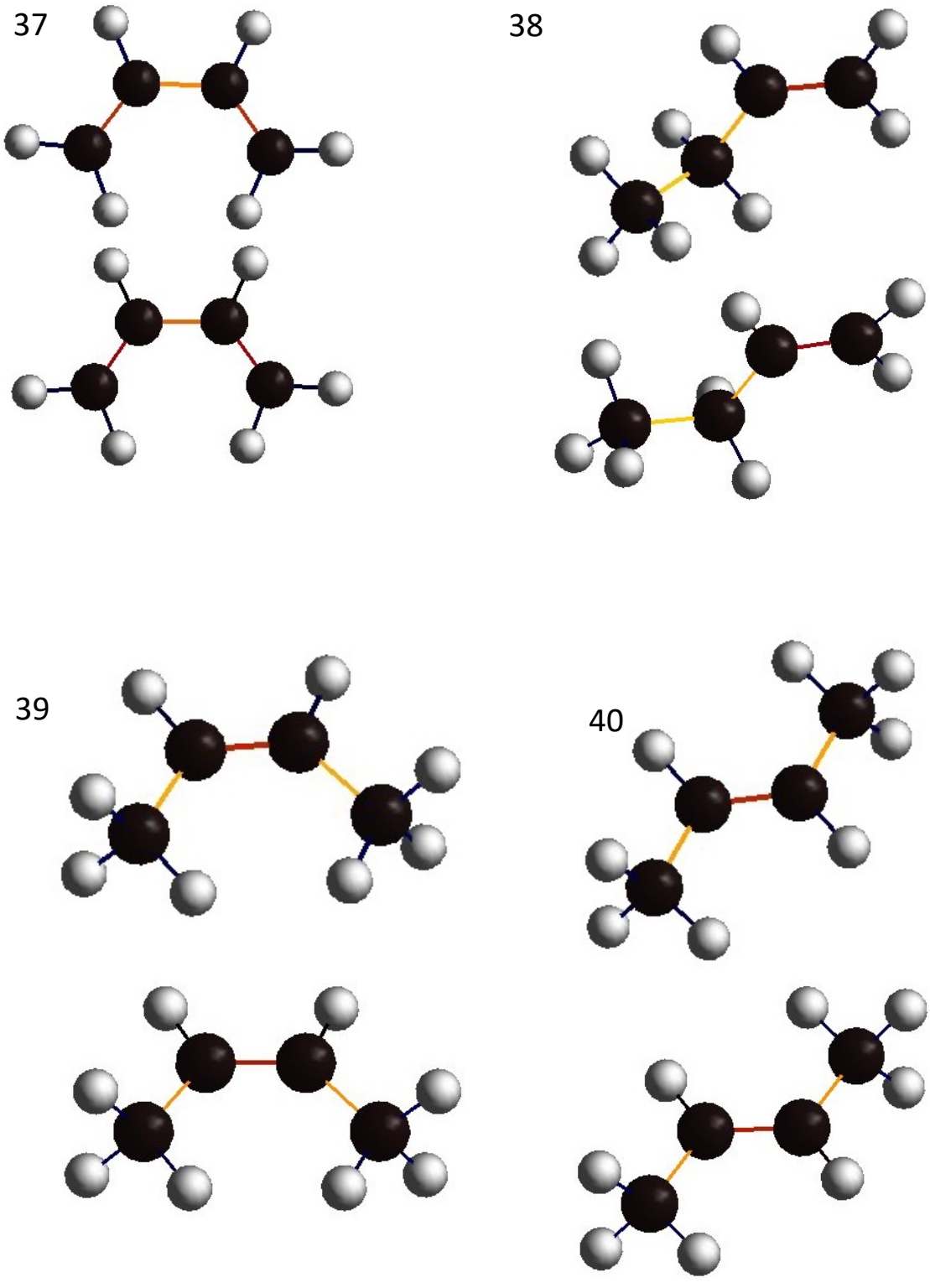

Figure S13: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 7).

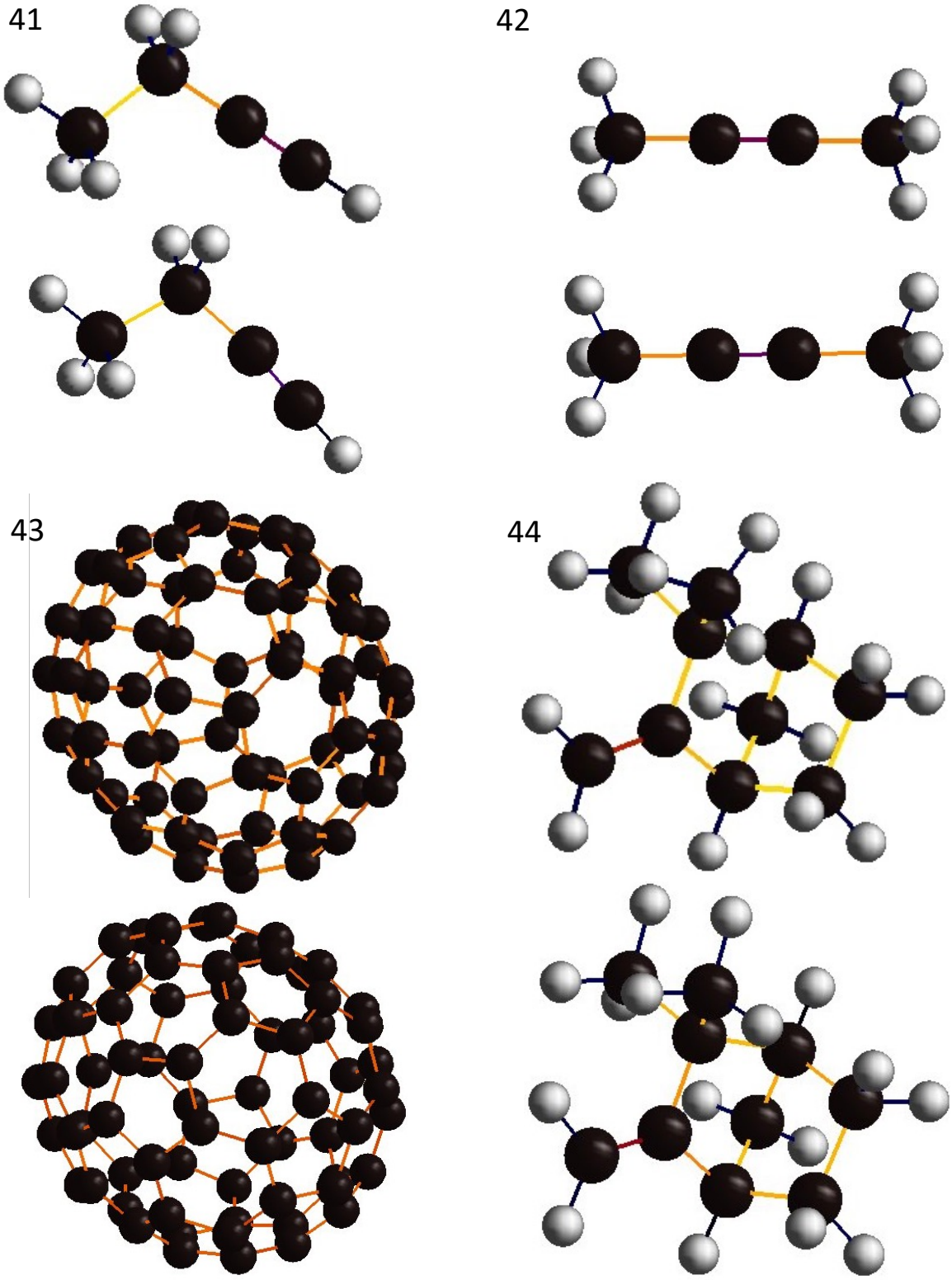

Figure S14: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 8).

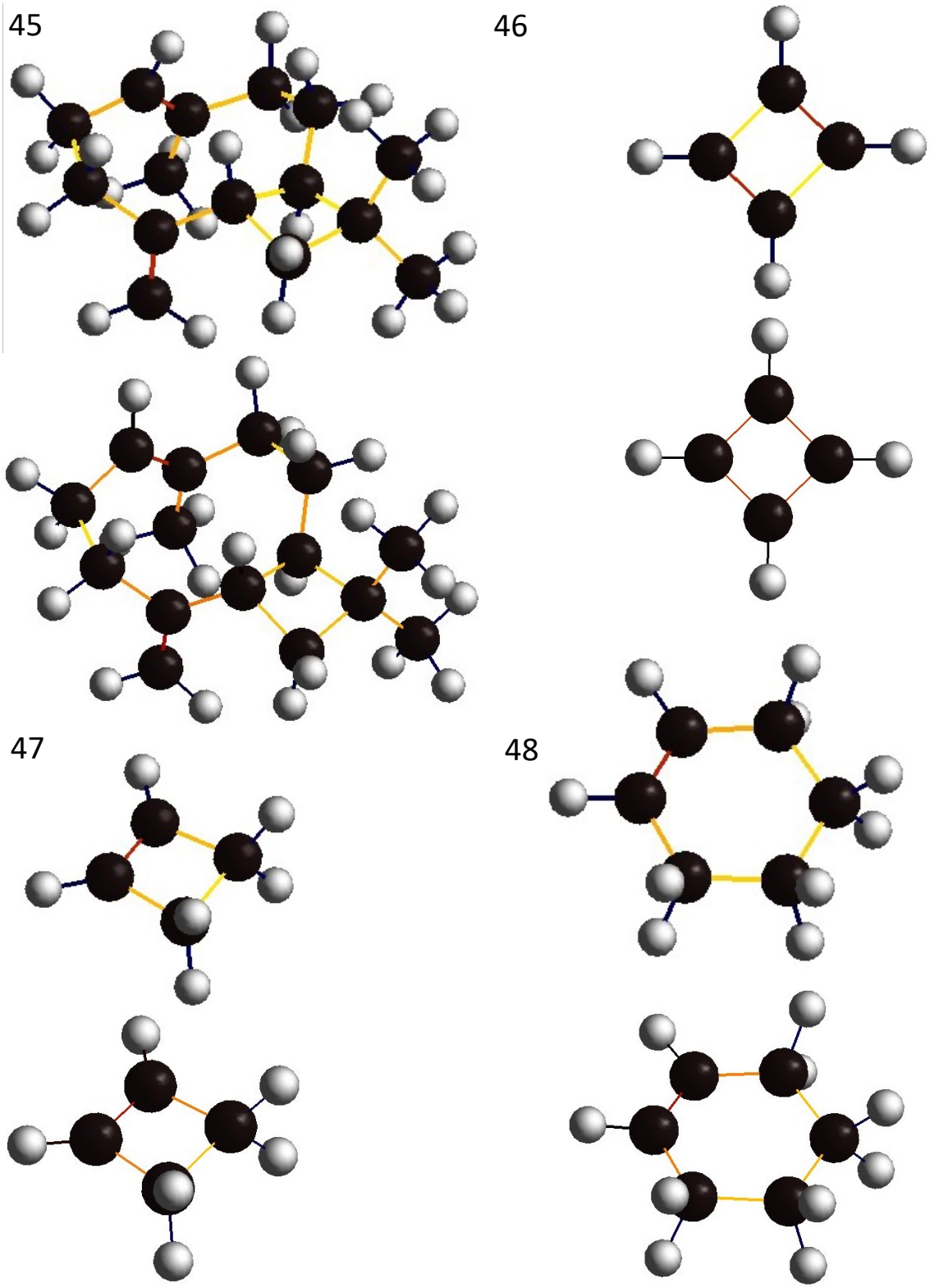

Figure S15: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 9).

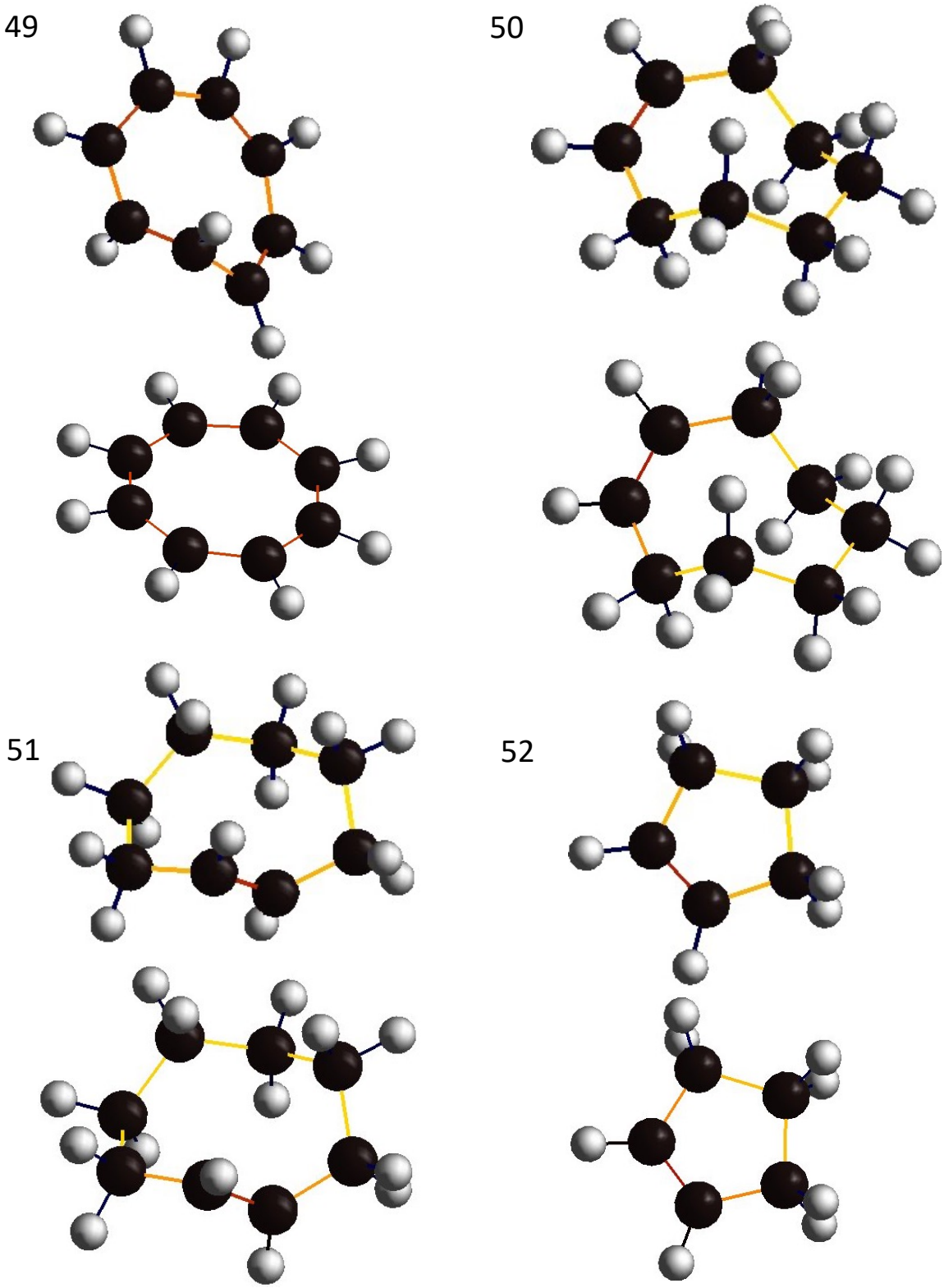

Figure S16: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 10).

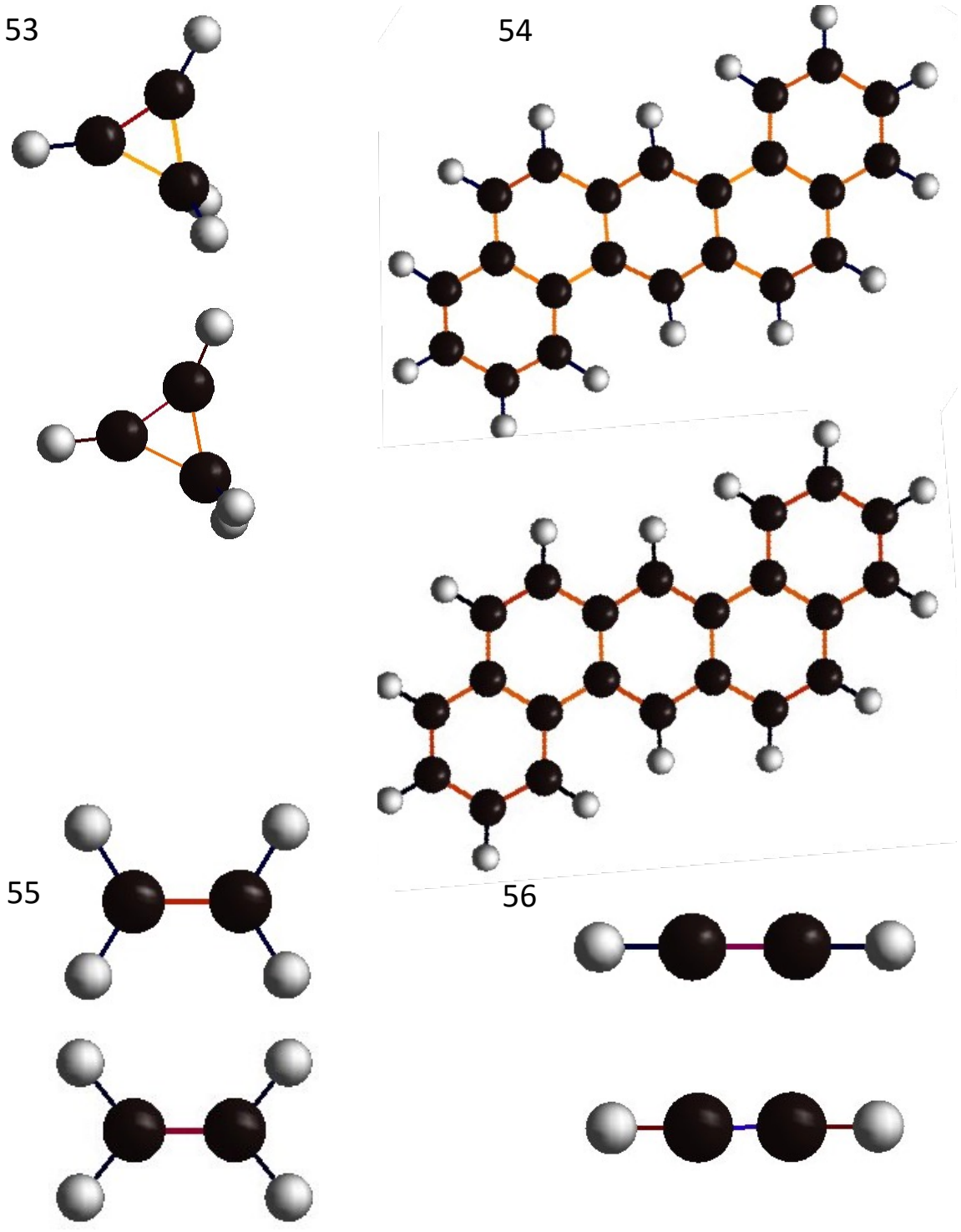

Figure S17: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 11).

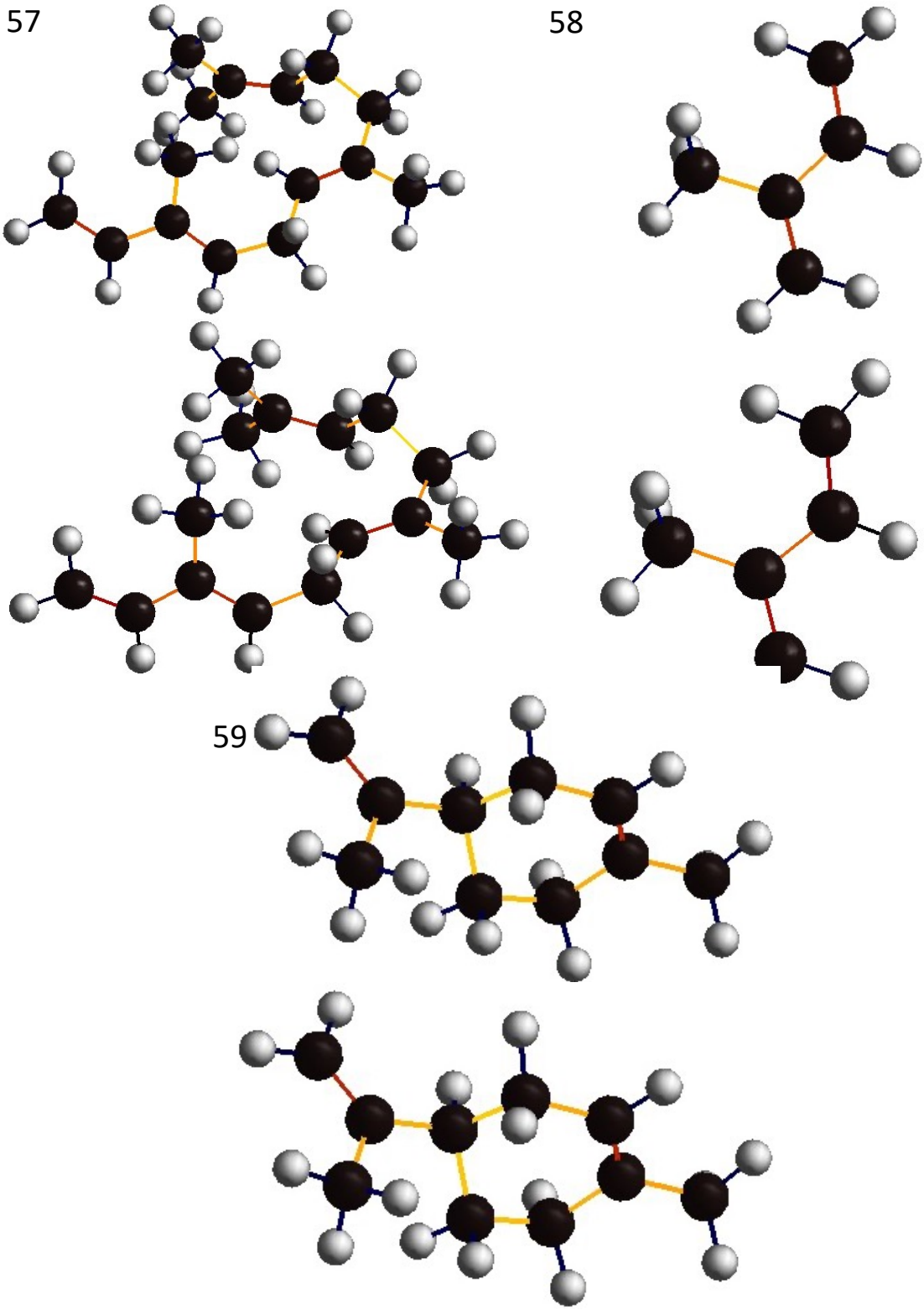

Figure S18: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 12).

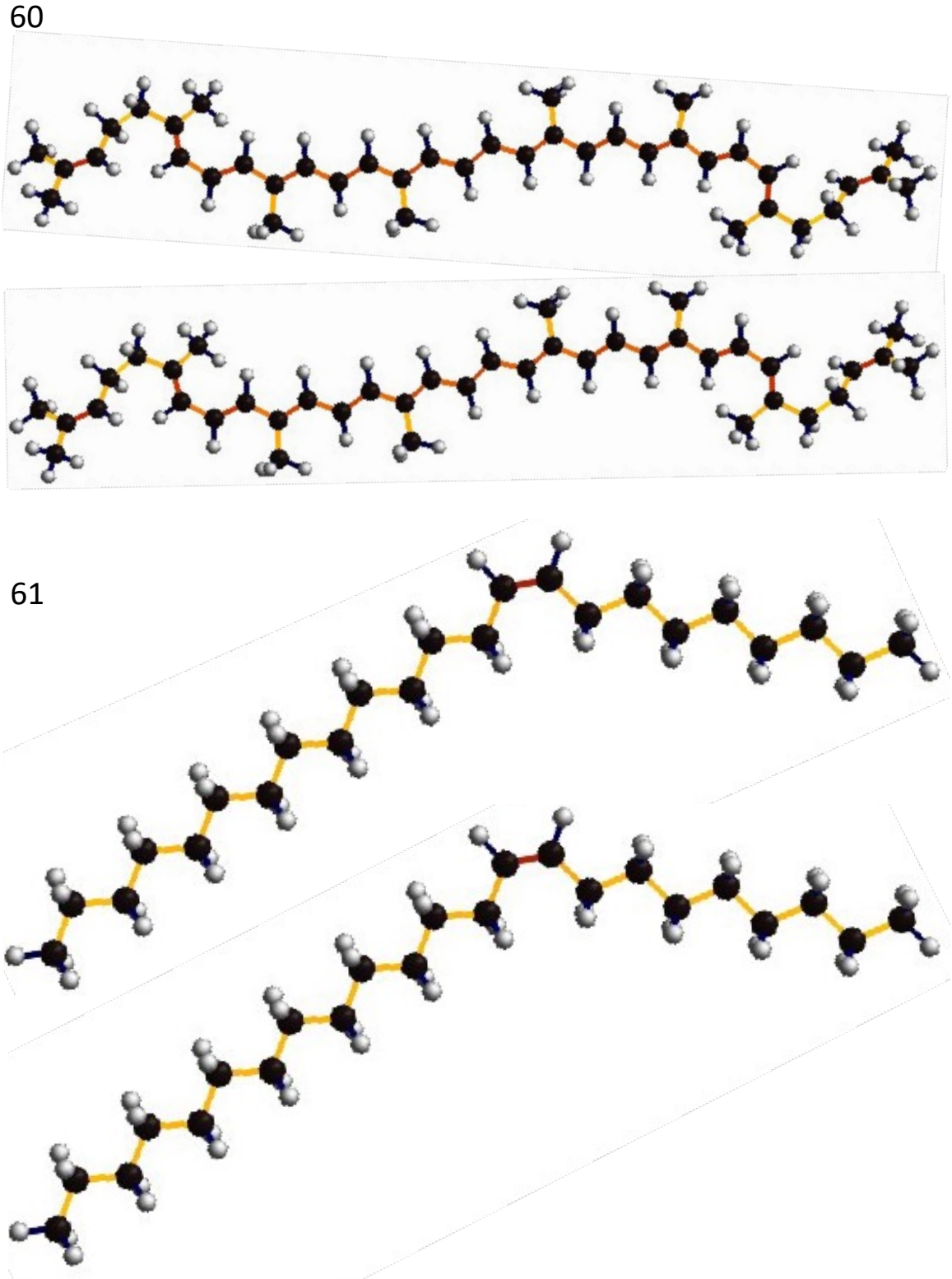

Figure S19: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 13).

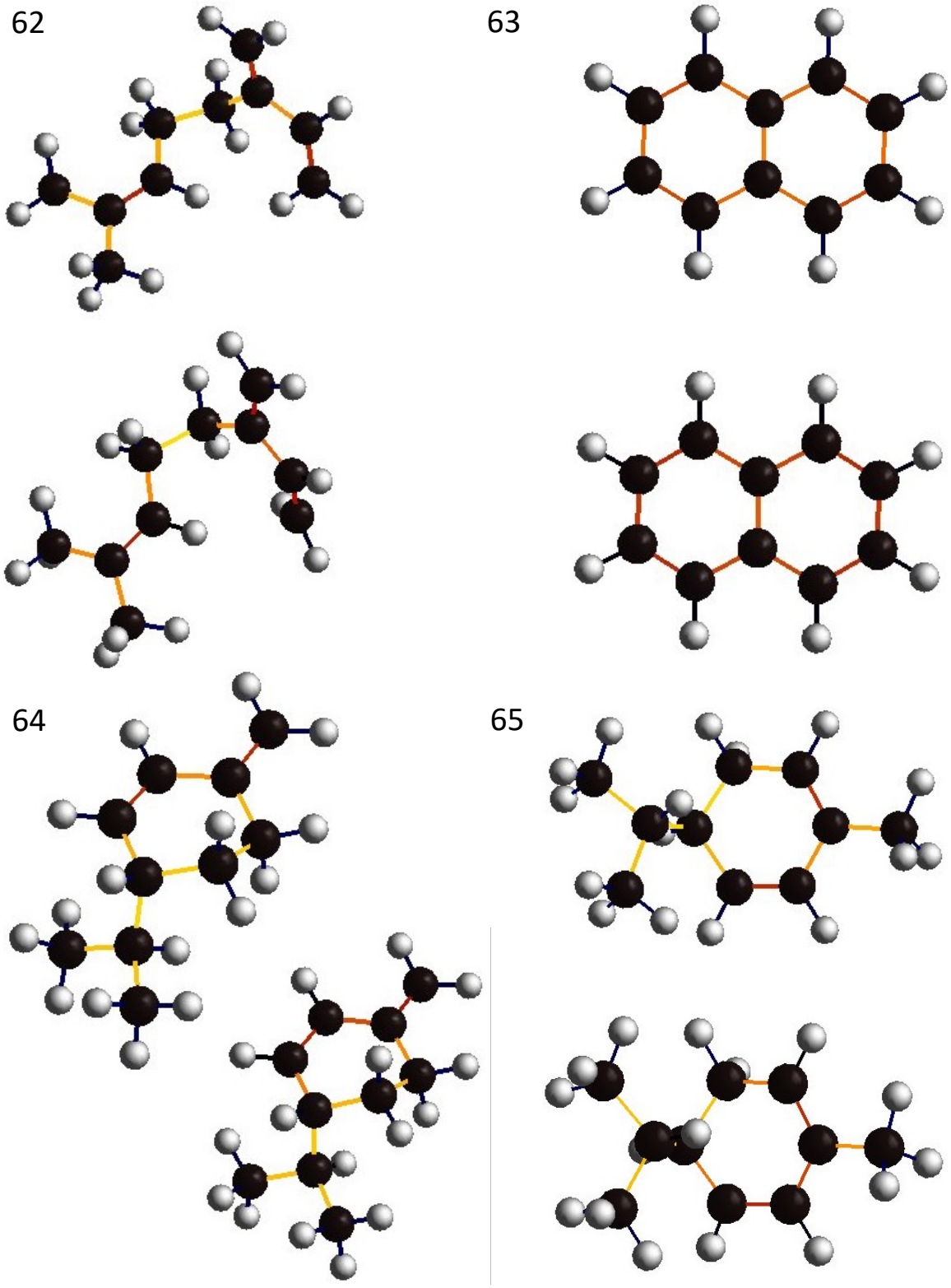

Figure S20: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 14).

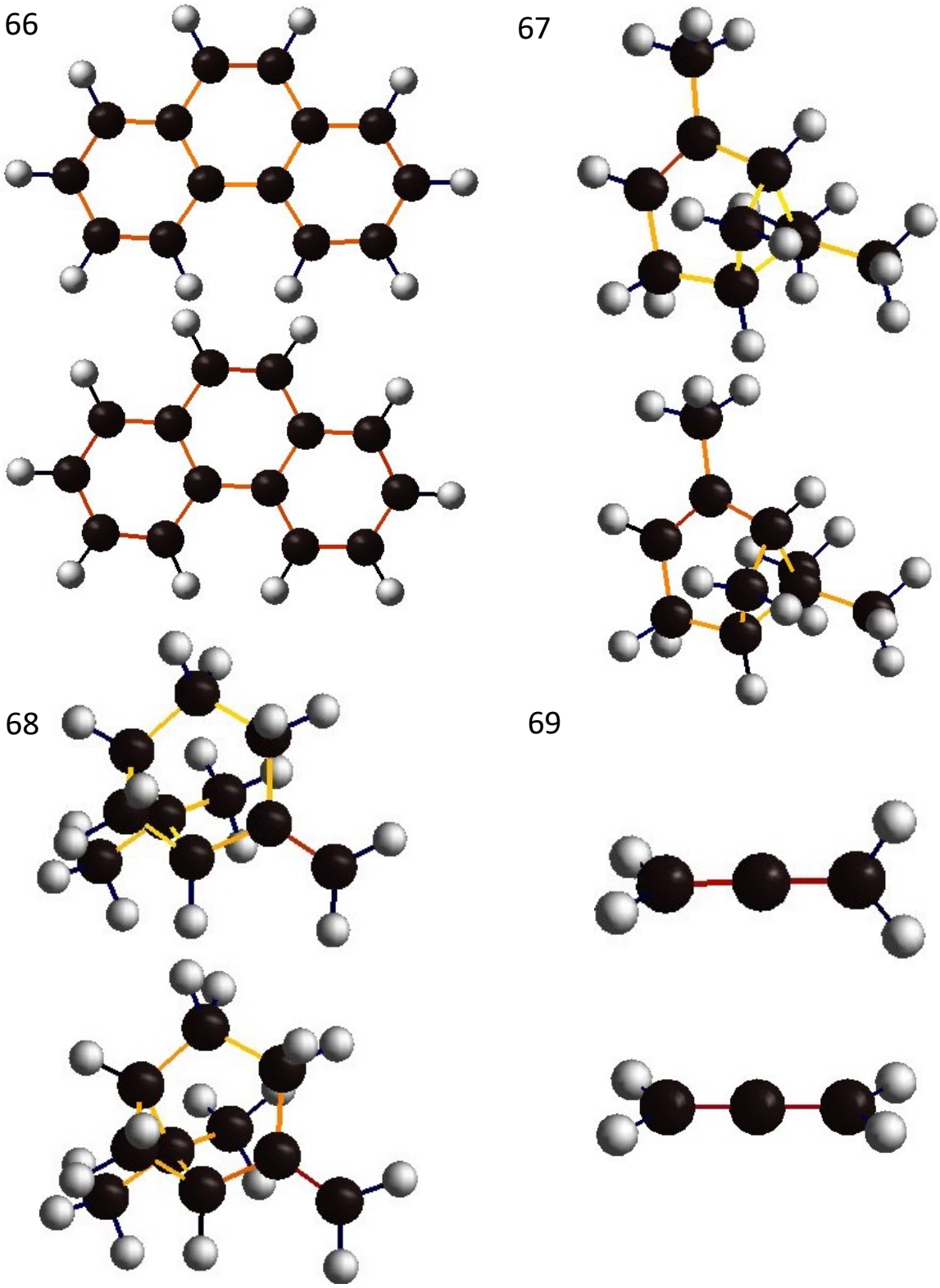

Figure S21: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 15).

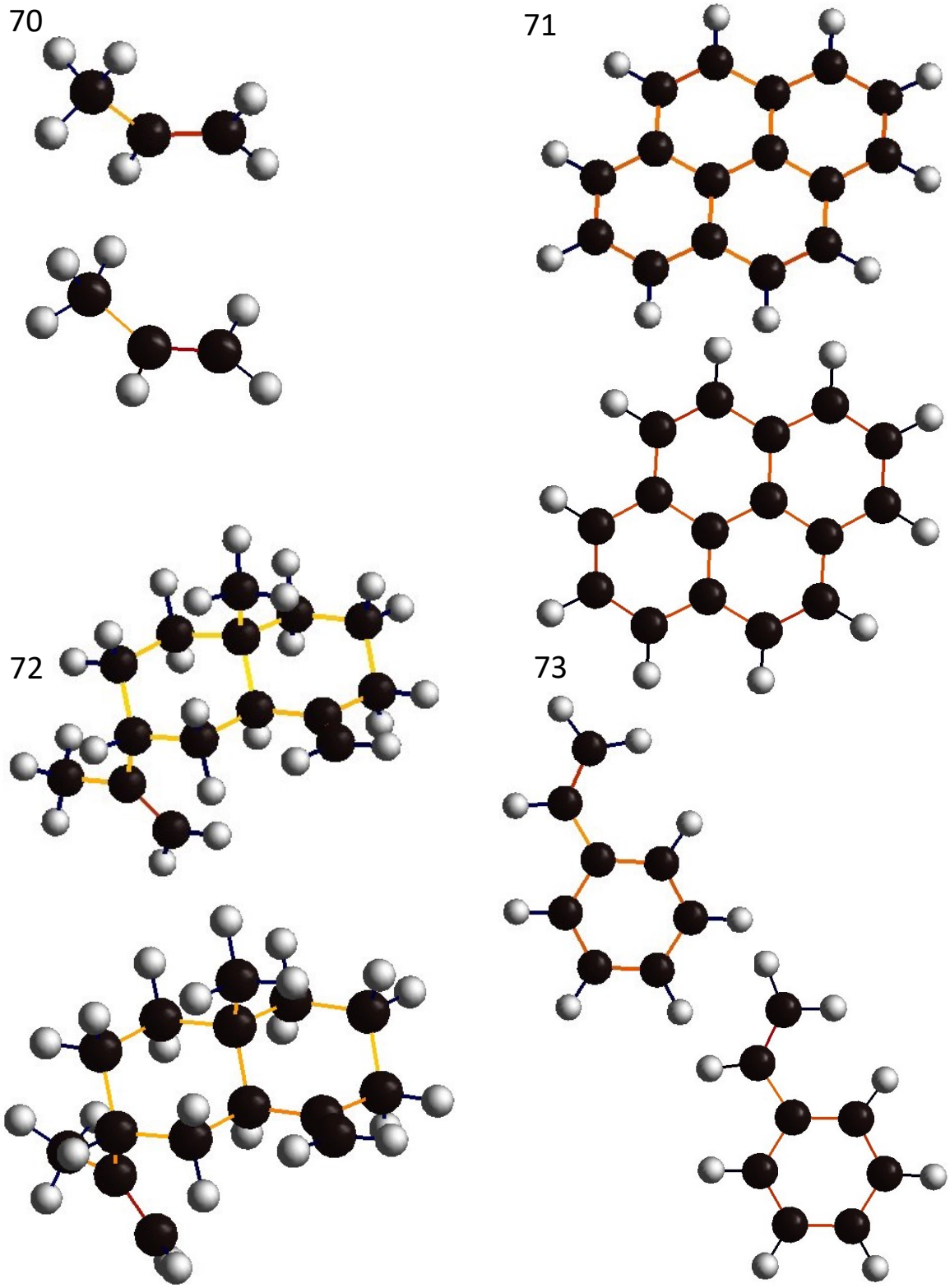

Figure S22: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 16).

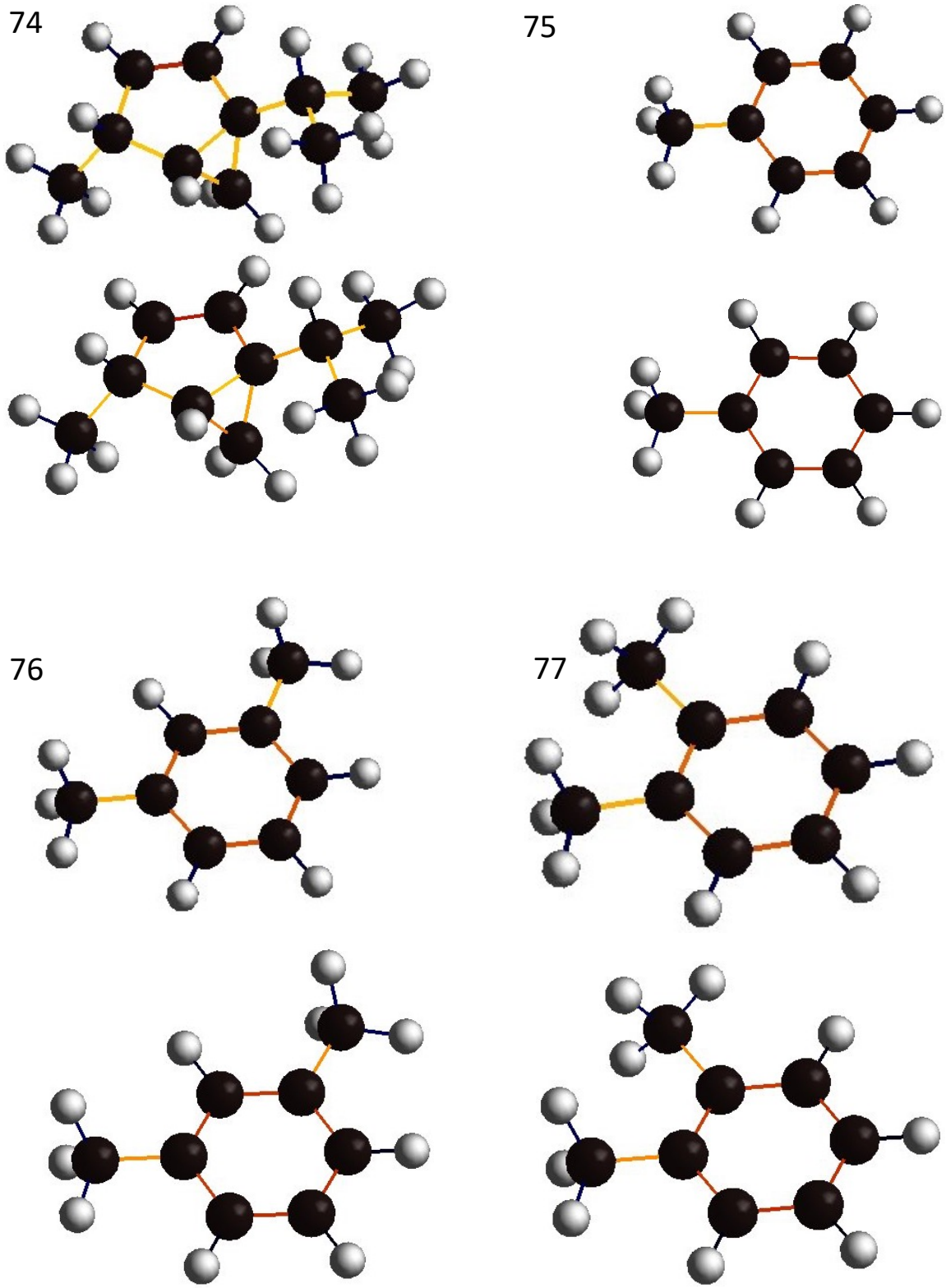

Figure S23: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 17).

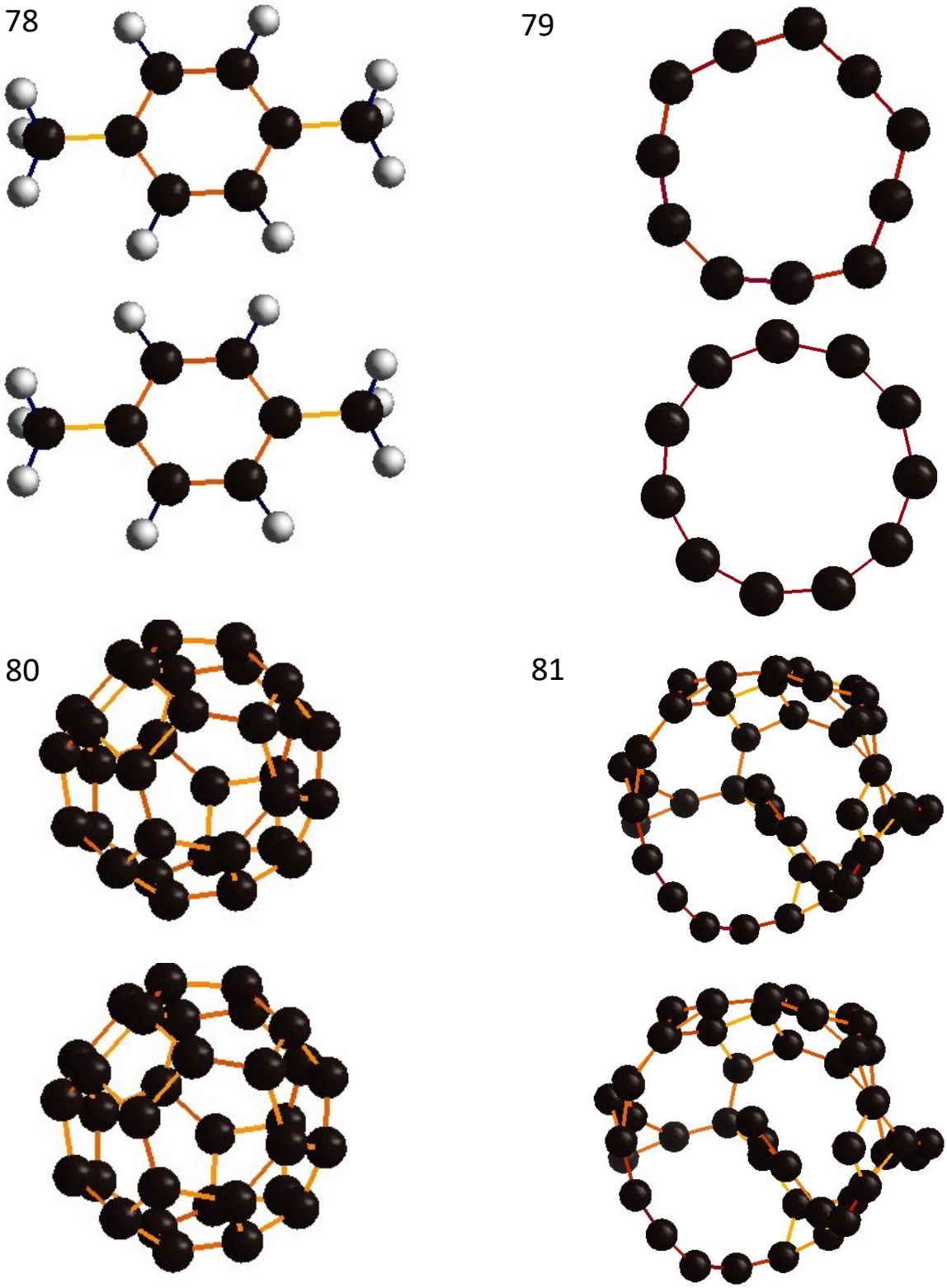

Figure S24: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 18).

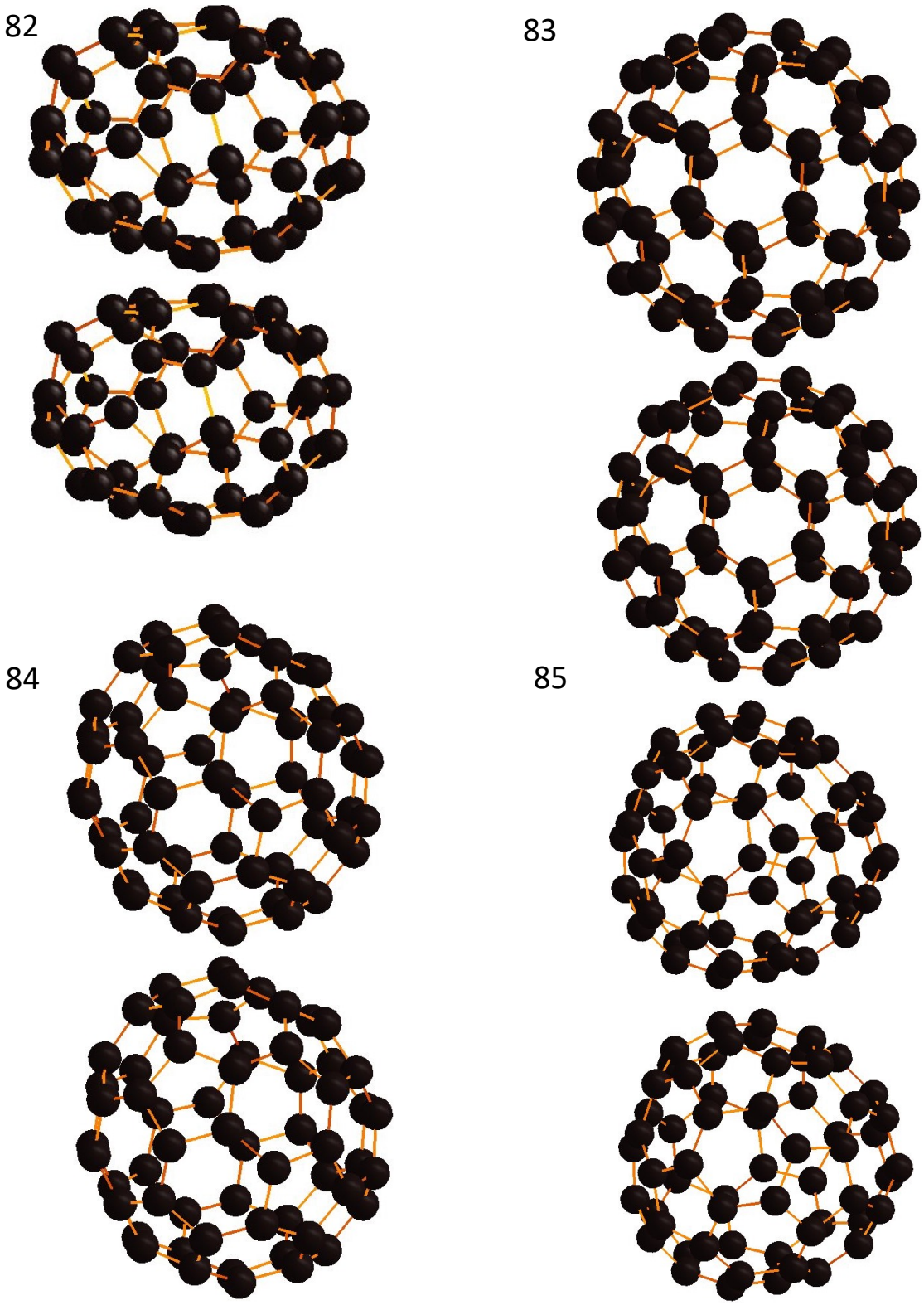

Figure S25: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 19).

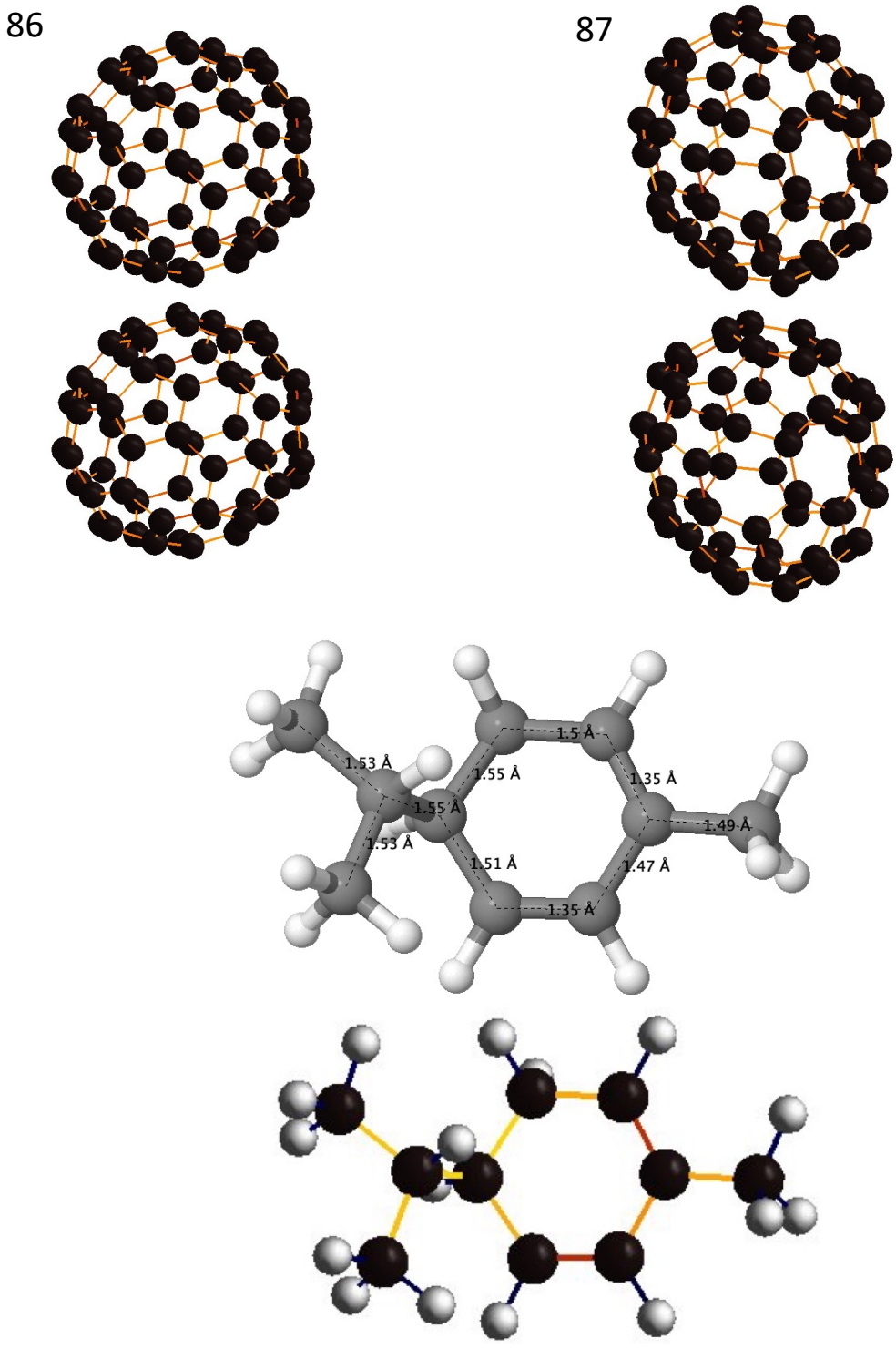

Figure S26: Optimized structures obtained from both DFT and the ANN potential. For each structural ID, the upper geometry is the DFT-optimized structure and the lower one is the ANN-optimized structure (part 20). The figure at the bottom is a sample showcasing bond lengths in one structure, offering insight into the color-coded representation denoting varying bond lengths.

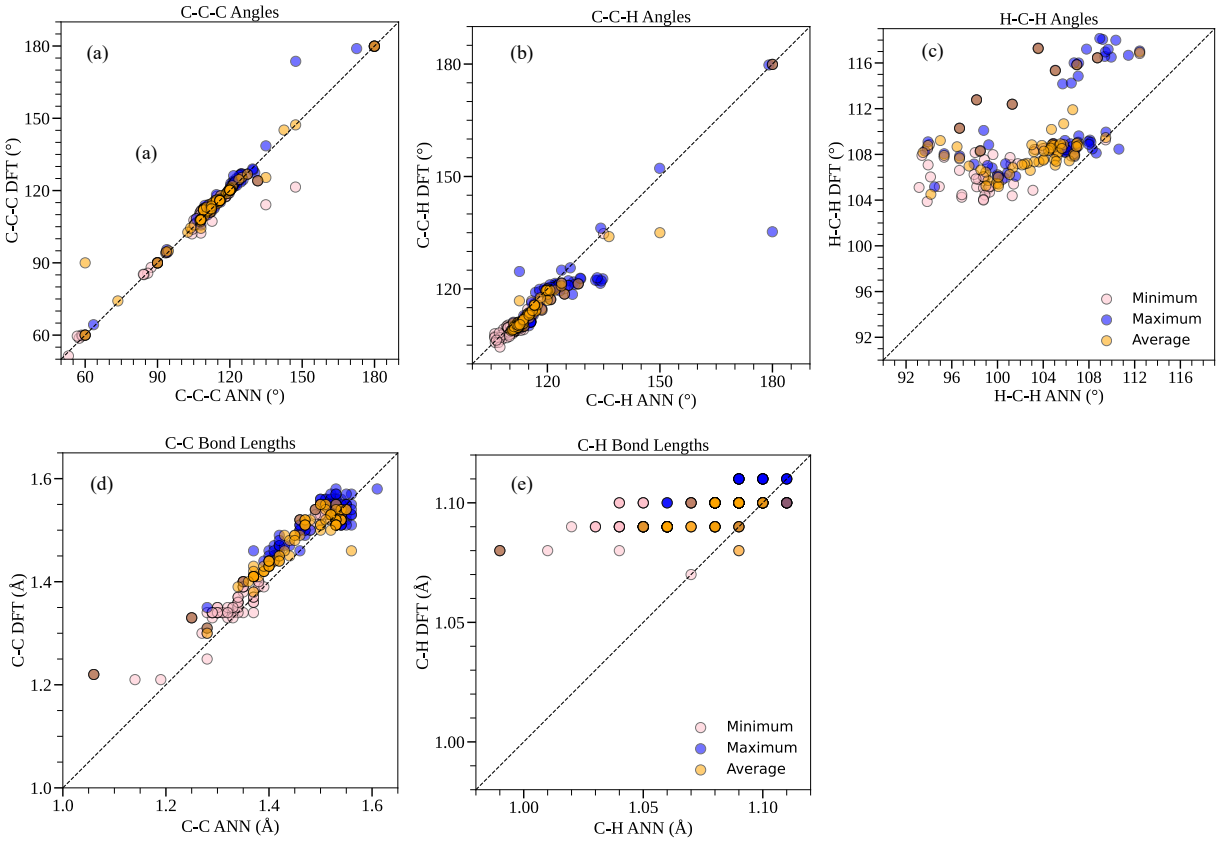

Figure S27: Bond length and bond angle analysis were conducted for the 0D test cases optimized by both DFT and the ANN potential. Panels (a)-(c) display the C-C-C, C-C-H, and H-C-H bond angles, while panels (d) and (e) represent the C-C and C-H bond lengths, respectively.

4 1D and 2D test cases

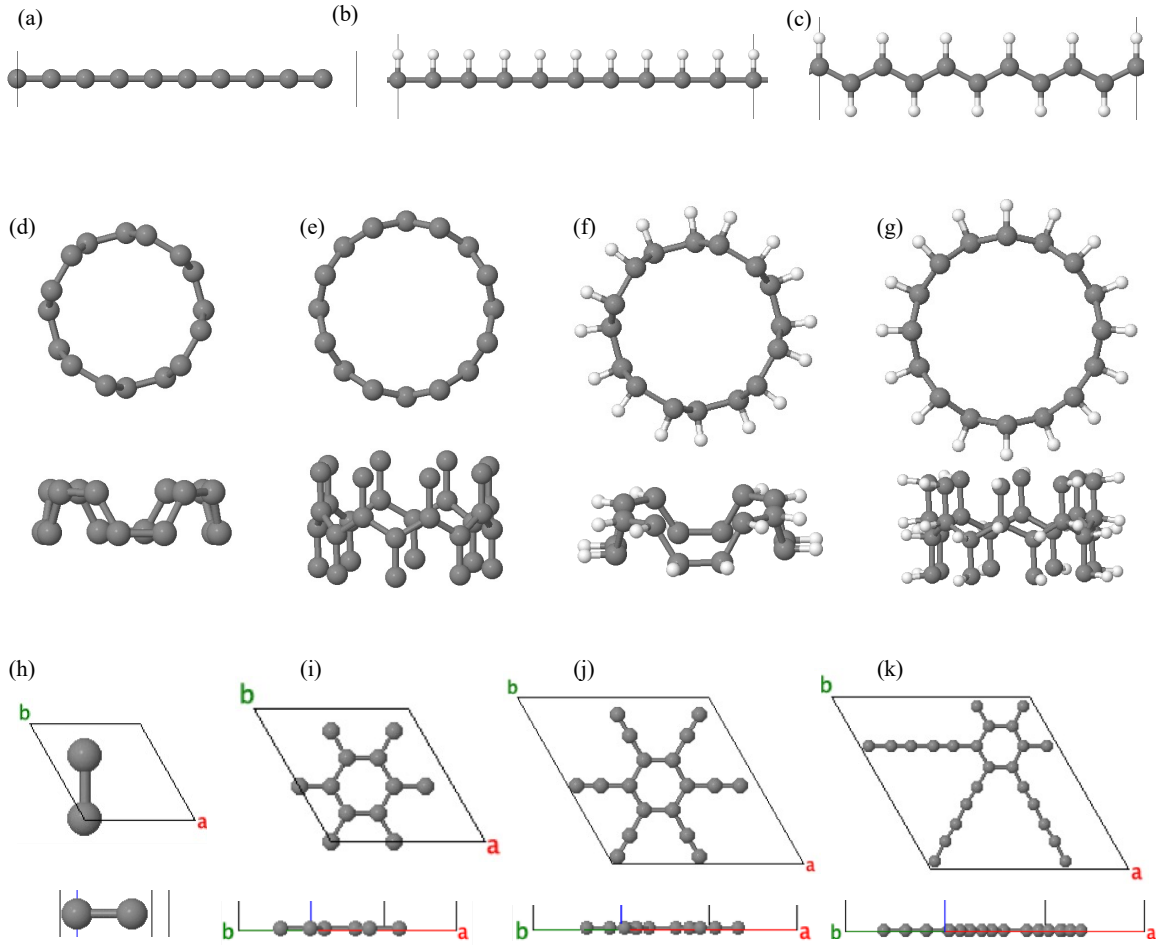

Figure S28: Top views and side views of the unitcells of the studied 1D and 2D systems: (a)-(c) 10-atom carbon chains,(d) SWCNT(4,4), (e) SWCNT (8,0), (f) H-saturated SWCNT(4,4), (g) H-saturated SWCNT (8,0), (h) graphene, (i) graphyne-1, (j) graphyne-2, and (k) graphyne-3.

5 3D test cases

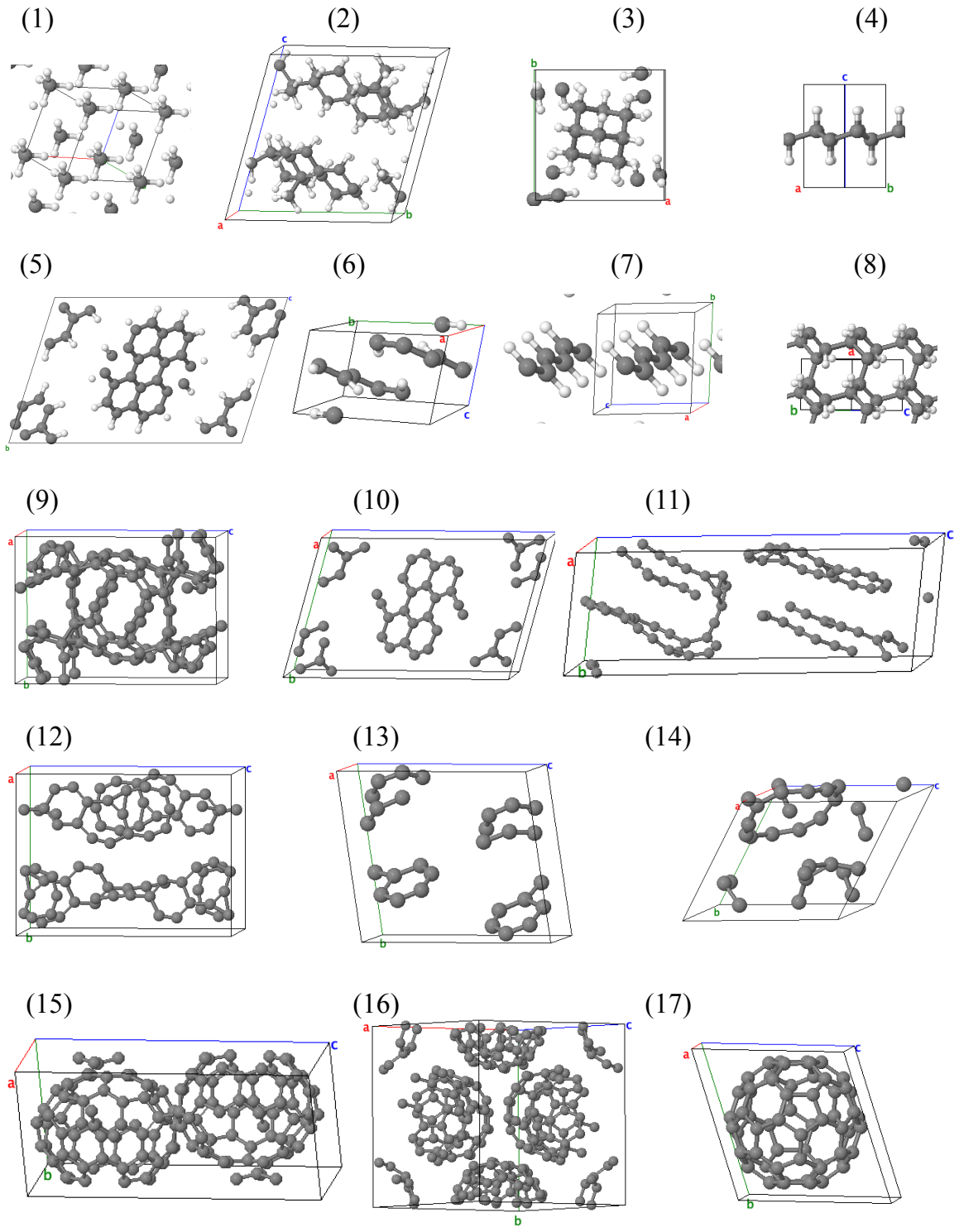

Figure S29: 3D C-H and C systems used as test case for geometry optimizations. The ID of each structure is the one reported in Table S2 and S3 (part 1).

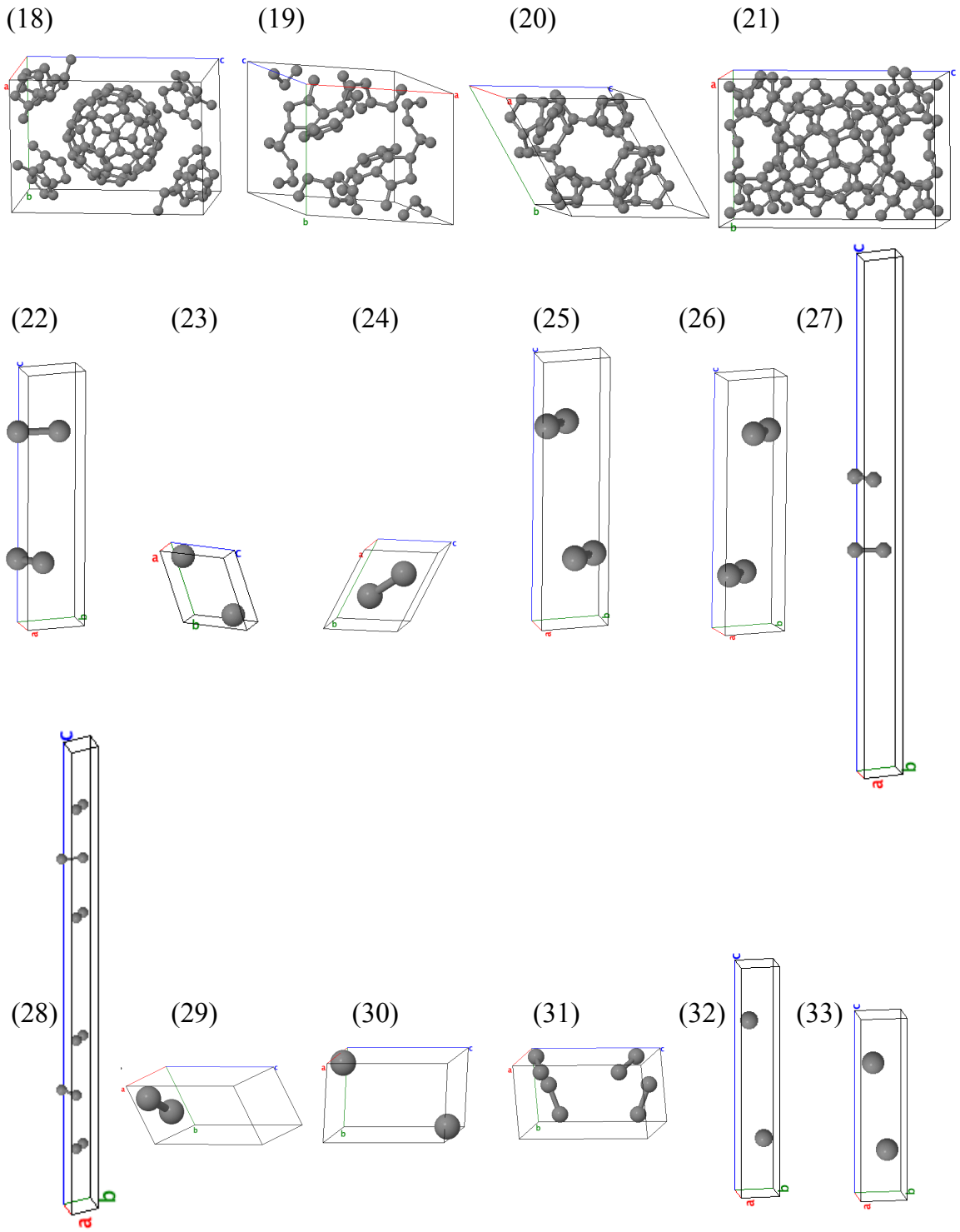

Figure S30: 3D C-H and C systems used as test case for geometry optimizations. The ID of each structure is the one reported in Table S2 and S3 (part 2).

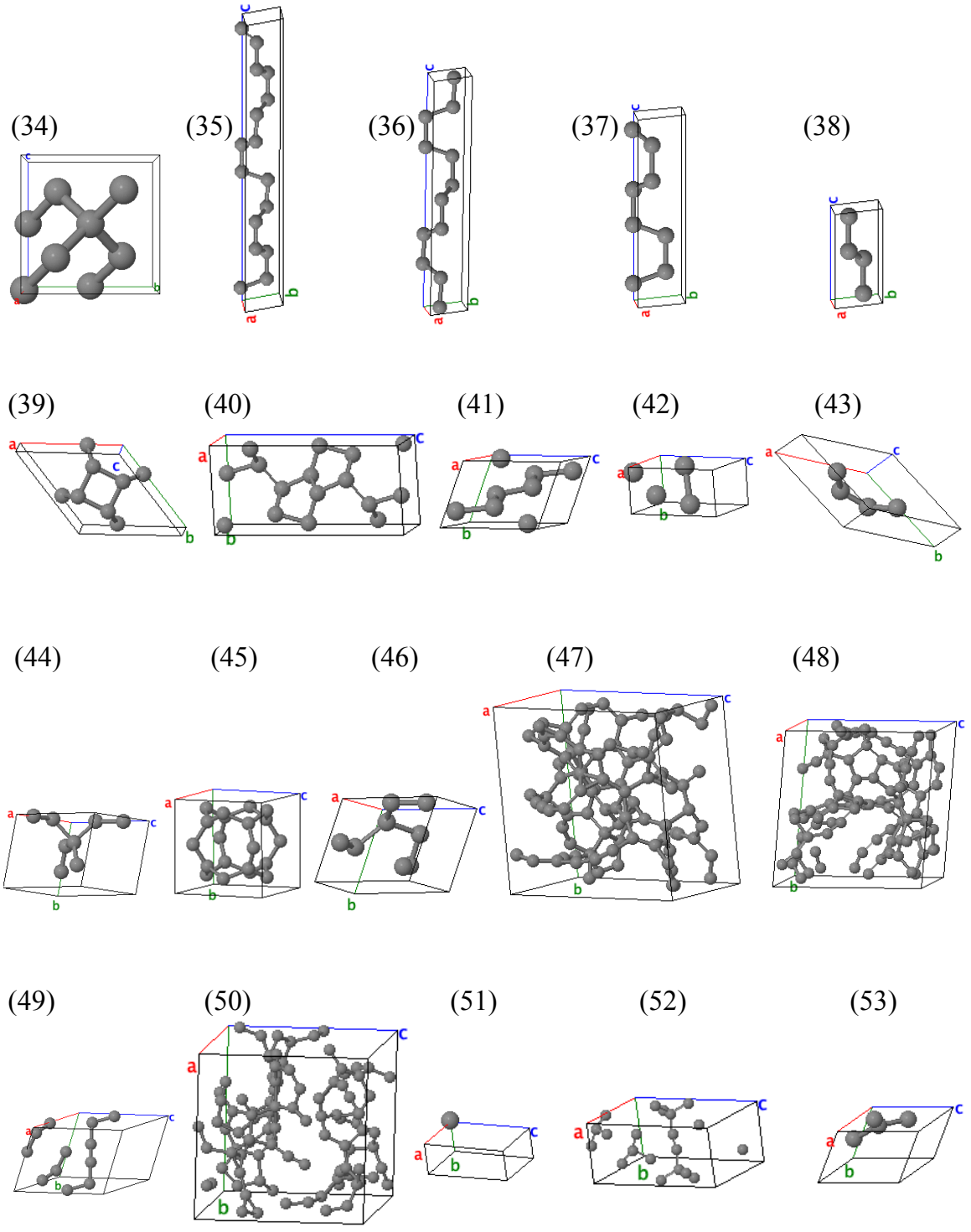

Figure S31: 3D C-H and C systems used as test case for geometry optimizations. The ID of each structure is the one reported in Table S2 and S3 (part 3).

Table S3: Chemical formula, space group (spg), total energies (E_{tot}), total energies differences of DFT and ANN (ΔE_{tot}), formation energies (E_f), and formation energy differences of DFT and ANN (E_f) of the examined 3D systems. All the energies are in eV/atom. The Materials Project'IDs of the structures are listed in the last column. The experimentally reported structures are marked by *.

ID	system	spg	$E_{tot,DFT}$	$E_{tot,ANN}$	ΔE_{tot}	$E_{f,DFT}$	$E_{f,ANN}$	ΔE_f	MP-ID
01	$H_{04}C_{01}$	$I\bar{4}3m$	-4.81413	-4.83068	0.01655	-0.28714	-0.30369	0.01655	mp-1021328
02	$H_{68}C_{38}$	$P12_11\bar{1}$	-5.63471	-5.64953	0.01482	-0.20320	-0.21802	0.01482	mp-30168*
03	$H_{32}C_{20}$	$P\bar{4}2_1c$	-5.77756	-5.77546	-0.00210	-0.19696	-0.19486	-0.00210	mp-1195106*
04	$H_{02}C_{02}$	$P\bar{3}m1$	-6.40774	-24.93188	18.5241	-0.16862	-18.6928	18.5242	mp-1096986
05	$H_{24}C_{48}$	$P12_1/c1$	-7.28579	-7.31600	0.03021	-0.09549	-0.12570	0.03021	mp-603334*
06	H_8C_{12}	$C12/c1$	-6.88042	-6.91393	0.03351	-0.07060	-0.10410	0.03350	mp-995217
07	$H_{02}C_{02}$	$Cmcm$	-6.30604	-6.24579	-0.06025	-0.06693	-0.00668	-0.06025	mp-995195
08	$H_{04}C_{04}$	$I2_1\bar{3}$	-6.15054	-6.17993	0.02939	0.08858	0.05919	0.02939	mp-1079612
	group I								
09	C_{120}	$Pnnm$	-8.53708	-8.59123	0.05415	0.55559	0.50144	0.05414	mp-1205283
10	C_{48}	$P2_1/c$	-8.19050	-8.13138	-0.05912	0.90217	0.96128	-0.05912	mp-1203645*
11	C_{80}	$P1$	-8.02929	-7.99525	-0.03404	1.06338	1.09742	-0.03404	mp-1197903
12	C_{80}	$P2_12_12_1$	-7.98126	-7.78202	-0.19924	1.11141	1.31065	-0.19924	mp-1182684
13	C_{28}	$P2_1$	-7.72848	-7.66183	-0.06665	1.36419	1.43084	-0.06665	mp-1194362*
14	C_{29}	$I\bar{4}3m$	-7.30820	-7.37988	0.07168	1.78447	1.71279	0.07168	mp-1192619
	group II								
15	C_{140}	$Cmcm$	-8.88456	-8.87055	-0.01401	0.20811	0.22212	-0.01401	mp-683919*
16	C_{240}	$Pa\bar{3}$	-8.85087	-8.83161	-0.01926	0.24180	0.26105	-0.01926	mp-1196583*
17	C_{60}	$Fm\bar{3}$	-8.84961	-8.83163	-0.01798	0.24306	0.26104	-0.01798	mp-667273*
18	C_{120}	$Pnnm$	-8.84576	-8.81973	-0.02603	0.24691	0.27293	-0.02603	mp-1147718
19	C_{60}	$Immm$	-8.83945	-8.80686	-0.03256	0.25322	0.28581	-0.03259	mp-630227*
20	C_{60}	$R\bar{3}m$	-8.82435	-8.77838	-0.04597	0.26831	0.31429	-0.04597	mp-680372*
21	C_{120}	$Pnnm$	-8.69964	-8.71730	0.01766	0.39303	0.37537	0.01766	mp-568028*
	group III								
22	C_4	$P6_3/mmc$	-9.22515	-9.28328	0.05813	-0.13249	-0.19061	0.05813	mp-48*
23	C_2	$R\bar{3}m$	-9.22459	-9.28328	0.05869	-0.13192	-0.19061	0.05870	mp-169*
24	C_2	$Fmmm$	-9.22424	-9.28328	0.05904	-0.13157	-0.19061	0.05904	mp-937760
25	C_4	$Cmme$	-9.22406	-9.28328	0.05922	-0.13139	-0.19061	0.05923	mp-568286
26	C_4	$Cmme$	-9.22396	-9.28328	0.05932	-0.13129	-0.19061	0.05933	mp-568363
27	C_4	$P\bar{3}m1$	-9.22303	-9.28328	0.06025	-0.13036	-0.19061	0.06025	mp-990424
28	C_{12}	$P6_3/mmc$	-9.22292	-9.28328	0.06036	-0.13025	-0.19061	0.06036	mp-606949
29	C_2	$P6/mmm$	-9.22073	-9.28328	0.06255	-0.12806	-0.19061	0.06256	mp-568806
30	C_2	$C2/m$	-9.21831	-9.28328	0.06497	-0.12564	-0.19061	0.06497	mp-632329
31	C_8	$Cmmm$	-8.28535	-8.57616	0.29081	0.80732	0.51651	0.29082	mp-579909
32	C_2	$Cmcm$	-6.69705	-7.20123	0.50418	2.39562	1.89144	0.50418	mp-1097832

ID	system	spg	$E_{tot,DFT}$	$E_{tot,ANN}$	ΔE_{tot}	$E_{f,DFT}$	$E_{f,ANN}$	ΔE_f	MP-ID
33	C_2	$Cmcm$	-6.68170	-7.20123	0.51953	2.41097	1.89144	0.51953	mp-1056957*
	group IV								
34	C_8	$Fd\bar{3}m$	-9.09267	-8.90824	-0.18443	0.00000	0.18442	-0.18442	mp-66*
35	C_{16}	$P6_3/mmc$	-9.08823	-8.91262	-0.17561	0.00444	0.18005	-0.17561	mp-616440
36	C_{12}	$P6_3/mmc$	-9.08682	-8.91441	-0.17241	0.00585	0.17826	-0.17241	mp-611448
37	C_8	$P6_3/mmc$	-9.08352	-8.91890	-0.16462	0.00915	0.17376	-0.16461	mp-611426
38	C_4	$P6_3/mmc$	-9.06806	-8.93745	-0.13061	0.02461	0.15522	-0.13061	mp-47*
39	C_8	$Cmmm$	-8.96336	-8.83272	-0.13064	0.12930	0.25994	-0.13064	mp-1078845
40	C_{16}	$Pnma$	-8.93867	-8.86150	-0.07717	0.15400	0.23117	-0.07718	mp-1190171
41	C_8	$C2/m$	-8.93054	-8.85154	-0.07900	0.16213	0.24113	-0.07900	mp-1080826
42	C_4	$I4/mmm$	-8.89623	-8.74316	-0.15307	0.19644	0.34950	-0.15306	mp-1008395
43	C_4	$Cmmm$	-8.79138	-8.63055	-0.16083	0.30129	0.46212	-0.16083	mp-1008374
44	C_8	$Im\bar{3}m$	-8.46581	-8.79792	0.33211	0.62686	0.29475	0.33212	mp-570002*
45	C_{20}	$Pm\bar{3}m$	-8.46507	-8.55512	0.09005	0.62760	0.53755	0.09005	mp-1188817
46	C_8	$Ia\bar{3}$	-8.39587	-8.39042	-0.00545	0.69680	0.70225	-0.00544	mp-24
47	C_{100}	$P1$	-8.29905	-8.43063	0.13158	0.79362	0.66203	0.13158	mp-1244964
48	C_{100}	$P1$	-8.23585	-8.37504	0.13919	0.85682	0.71762	0.13920	mp-1244913
49	C_{12}	$I4/mmm$	-8.23249	-8.00509	-0.22740	0.86018	1.08758	-0.22740	mp-1095633
50	C_{100}	$P1$	-8.21852	-8.34462	0.12610	0.87415	0.74805	0.12610	mp-1245190
51	C_1	$P2/m$	-8.21158	-8.38223	0.17065	0.88109	0.71044	0.17065	mp-1182029
52	C_{20}	$I4/mmm$	-8.06062	-7.79945	-0.26117	1.03205	1.29322	-0.26117	mp-1205417
53	C_4	$I4_132$	-7.91997	-8.68112	0.76115	1.17270	0.41155	0.76115	mp-1018088

Table S4: Chemical formula, space group (spg), lattice constants (a, b, and c in Å), volumes (in Å^3/atom), volume differences (Δ V in Å^3/atom) obtained from DFT and ANN.

ID	system	spg	a_{DFT}	a_{ANN}	b_{DFT}	b_{ANN}	c_{DFT}	c_{ANN}	V_{DFT}	V_{ANN}	Δ V
01	$H_{04}C_{01}$	$I\bar{4}3m$	4.31	4.29	4.31	4.30	4.31	4.30	12.303	12.219	0.084
02	$H_{68}C_{38}$	$P12_11\bar{1}$	7.74	7.79	11.04	10.67	11.38	11.84	8.862	8.994	-0.133
03	$H_{32}C_{20}$	$P\bar{4}2_1c$	6.87	7.30	6.87	7.30	9.34	9.76	8.477	9.992	-1.515
04	$H_{02}C_{02}$	$P\bar{3}m1$	2.54	1.78	2.54	1.78	5.49	5.36	7.670	1.684	5.986
05	$H_{24}C_{48}$	$P12_1/c1$	5.10	5.00	10.35	10.48	16.53	15.99	11.289	10.990	0.299
06	H_8C_{12}	$C12/c1$	7.41	4.22	7.41	7.76	4.15	7.76	10.489	11.121	-0.632
07	$H_{02}C_{02}$	$Cmcm$	4.12	4.21	4.12	4.21	2.47	2.41	10.357	10.608	-0.251
08	$H_{04}C_{04}$	$I2_1\bar{3}$	3.74	3.61	3.74	3.61	3.74	3.61	5.027	4.544	0.483
	group I										
09	C_{120}	$Pnnm$	9.12	8.86	9.61	9.47	12.98	13.32	9.666	9.073	0.593
10	C_{48}	$P2_1/c$	4.98	4.63	10.03	10.03	15.65	15.70	15.705	14.568	1.137
11	C_{80}	$P1$	7.27	7.37	7.92	9.52	22.53	23.55	16.041	19.165	-3.123
12	C_{80}	$P2_12_12_1$	8.19	9.35	12.14	11.33	14.31	14.52	16.540	18.949	-2.409
13	C_{28}	$P2_1$	6.09	6.19	8.65	8.54	9.42	9.46	17.611	17.583	0.028
14	C_{29}	$I\bar{4}3m$	7.97	7.73	7.97	7.74	7.74	7.76	13.457	12.319	1.139
	group II										
15	C_{140}	$Cmcm$	10.52	9.87	10.52	9.87	17.62	18.00	12.627	10.700	1.927
16	C_{240}	$Pa\bar{3}$	14.88	14.02	14.88	14.02	14.02	14.30	13.718	11.479	2.240
17	C_{60}	$Fm\bar{3}$	10.46	9.91	10.46	9.91	9.915	9.92	13.500	11.504	1.995
18	C_{120}	$Pnnm$	9.11	8.97	10.74	10.00	14.41	14.65	12.697	10.772	1.924
19	C_{60}	$Immm$	10.41	8.92	10.41	8.97	9.81	9.97	11.307	10.013	1.294
20	C_{60}	$R\bar{3}m$	10.20	9.02	10.20	9.02	9.68	9.80	10.660	9.594	1.066
21	C_{120}	$Pnnm$	8.31	8.19	9.60	9.34	13.56	14.08	9.365	8.644	0.722
	group III										
22	C_4	$P6_3/mmc$	2.47	2.39	2.47	2.39	8.69	8.69	11.451	10.731	0.720
23	C_2	$R\bar{3}m$	4.27	2.39	4.27	2.39	4.25	4.27	10.607	9.935	0.672
24	C_2	$Fmmm$	4.53	4.16	4.18	4.16	2.39	2.47	10.524	9.860	0.665
25	C_4	$Cmme$	2.47	2.39	2.47	2.39	8.03	8.03	10.590	9.919	0.671
26	C_4	$Cmme$	2.47	2.39	2.47	2.39	8.15	8.15	10.752	10.070	0.682
27	C_4	$P\bar{3}m1$	2.47	2.39	2.47	2.39	25.83	25.83	34.267	31.913	2.354
28	C_{12}	$P6_3/mmc$	2.47	2.39	2.47	2.39	31.98	31.98	14.080	13.172	0.908
29	C_2	$P6/mmm$	2.47	2.39	2.47	2.39	3.83	3.83	10.238	9.464	0.773
30	C_2	$C2/m$	2.47	2.39	2.47	2.39	3.74	3.74	9.843	9.215	0.628
31	C_8	$Cmmm$	3.68	3.57	3.68	3.57	6.30	6.27	10.572	9.853	0.719
32	C_2	$Cmcm$	1.62	1.55	1.62	1.55	9.88	9.88	12.900	11.925	0.975
33	C_2	$Cmcm$	1.62	1.55	1.62	1.55	6.40	6.40	8.362	7.723	0.639
	group IV										

ID	system	spg	a_{DFT}	a_{ANN}	b_{DFT}	b_{ANN}	c_{DFT}	c_{ANN}	V_{DFT}	V_{ANN}	Δ V
34	C_8	$Fd\bar{3}m$	3.57	3.51	3.57	3.51	3.51	3.56	5.705	5.401	0.304
35	C_{16}	$P6_3/mmc$	2.52	2.47	2.52	2.47	16.35	16.56	5.706	5.416	0.290
36	C_{12}	$P6_3/mmc$	2.52	2.47	2.52	2.47	12.30	12.43	5.707	5.420	0.287
37	C_8	$P6_3/mmc$	2.52	2.47	2.52	2.47	8.24	8.31	5.711	5.428	0.283
38	C_4	$P6_3/mmc$	2.51	2.46	2.51	2.46	4.15	4.18	5.718	5.442	0.276
39	C_8	$Cmmm$	4.87	4.19	4.87	4.76	2.50	2.51	5.866	5.586	0.280
40	C_{16}	$Pnma$	2.53	2.41	4.15	4.18	9.95	9.07	5.961	5.645	0.316
41	C_8	$C2/m$	4.77	2.41	4.77	4.17	4.68	4.15	5.975	5.655	0.320
42	C_4	$I4/mmm$	3.34	2.52	3.34	3.23	3.28	3.34	6.024	5.770	0.254
43	C_4	$Cmmm$	4.14	4.12	4.14	4.12	2.43	2.51	6.633	6.427	0.206
44	C_8	$Im\bar{3}m$	4.22	4.06	4.22	4.06	4.06	4.22	7.250	6.431	0.818
45	C_{20}	$Pm\bar{3}m$	5.22	5.17	5.22	5.17	5.17	5.22	7.096	6.916	0.180
46	C_8	$Ia\bar{3}$	3.88	3.80	3.88	3.80	3.80	3.88	5.606	5.269	0.338
47	C_{100}	$P1$	10.18	9.99	11.00	10.66	9.79	10.12	11.268	10.387	0.881
48	C_{100}	$P1$	10.65	10.31	10.73	10.50	10.27	10.43	11.843	11.070	0.773
49	C_{12}	$I4/mmm$	5.77	5.67	5.77	5.67	5.67	5.80	12.289	11.674	0.616
50	C_{100}	$P1$	10.19	9.77	10.42	10.10	10.32	19.54	11.166	10.174	0.992
51	C_1	$P2/m$	3.34	3.34	1.29	1.24	3.43	3.43	14.622	14.067	0.555
52	C_{20}	$I4/mmm$	8.96	6.69	8.96	8.79	8.80	8.96	23.364	22.131	1.233
53	C_4	$I4_132$	3.57	3.53	3.57	3.53	3.53	3.57	8.783	8.467	0.317

Table S5: Mean percentage error (MPE) (in Å) of lattice constants a, b, and c in the 3D systems. The values for C-H* are after excluding the layered system.

	C-H	C-H*	C-I	C-II	C-III	C-IV
MPE_a	2.751	-1.130	-0.720	6.606	7.131	7.359
MPE_b	2.820	-1.052	-1.325	7.462	6.490	2.503
MPE_c	0.163	-0.152	-1.597	-1.773	-0.256	-4.907

6 Structural and mechanical properties of the novel carbon polymorph

The structure of the discovered carbon polymorph is shown in Figure S32. Details of the structural parameters and atomic positions are summarized in Table S6.

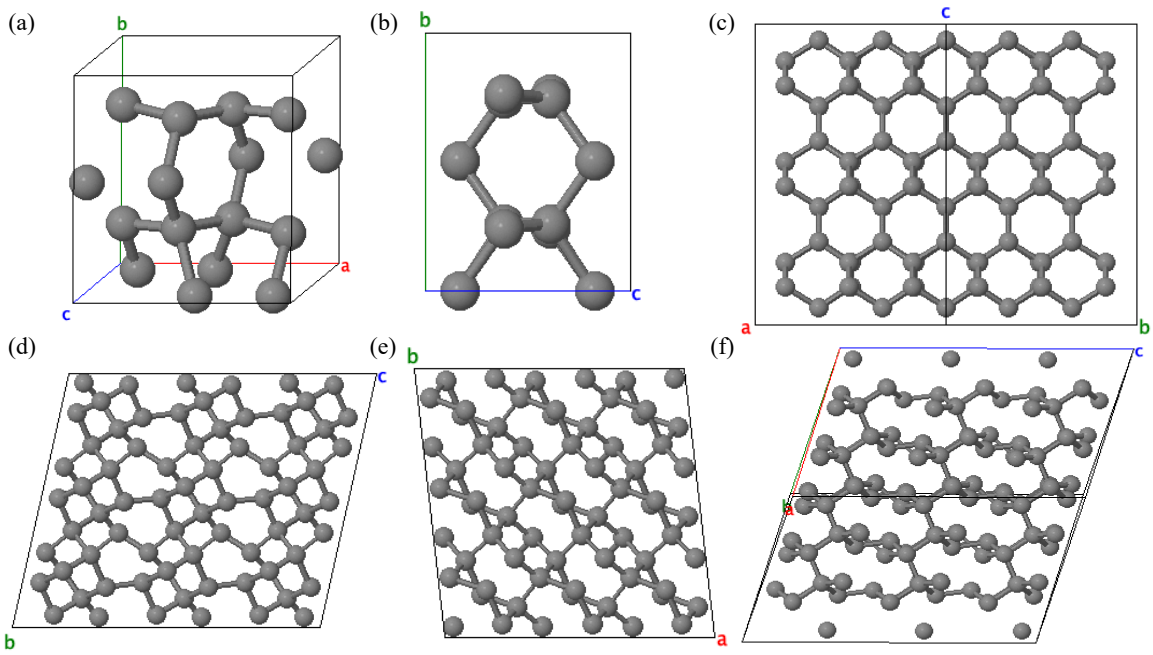

Figure S32: Different views of the unit cell (panels (a) and (b)) and $3 \times 3 \times 3$ supercell (panels (c)-(f)) of the new carbon polymorph.

Table S6: Lattice vectors a, b, and c along with the reduced atomic positions of the $C2/m$ carbon polymorph.

	x	y	z
a (Å)	2.36357	-2.52698	0.00000
b (Å)	2.36357	2.52698	0.00000
c (Å)	-1.21528	0.00000	3.99645
C1	0.52701	0.52701	0.82729
C2	0.11464	0.11464	0.16998
C3	0.33627	0.85558	0.38327
C4	0.14442	0.66373	0.61673
C5	0.85558	0.33627	0.38327
C6	0.66373	0.14442	0.61673
C7	0.88536	0.88536	0.83002
C8	0.47299	0.47299	0.17271

The 13 independent elastic constants are listed in Table S7. As shown in the table, the C_{11} and C_{22} are smaller, but comparable with C_{11} in diamond; however, C_{33} is close to that one for diamond.

Therefore, the new structure displays strong resistance to compression deformation along the x-, y-, and z-axis. The bulk modulus (B), Young's modulus (E) and shear modulus (G) are also provided in the table. Based on the values, B/G ratio for $C2/m$ is 0.89 comparable to diamond which is 0.84, smaller than 1.75, indicating the brittleness of $C2/m$ phase. Based on harmonic approximation,

Table S7: Calculated elastic constants C_{ij}, bulk modulus B, Young's modulus E, and shear modulus G for the $C2/m$ carbon polymorph along with other experimental values for diamond. All values are in GPa.

	$C2/m$	Diamond	
		Present	Experiments
C_{11}	1034.3	1050.0	1074.8 [1], 1079 [2]
C_{12}	27.6	127.4	125.6 [1], 124 [2]
C_{13}	86.0		
C_{15}	-3.0		
C_{22}	1023.3		
C_{23}	131.8		
C_{25}	-17.3		
C_{33}	1074.9		
C_{35}	-3.5		
C_{44}	505.1	559.3	720.6 [1], 578 [2]
C_{46}	-15.7		
C_{55}	440.8		
C_{66}	372.2		
B	401.8	434.9	442 [1, 2]
E	988.1	1112.1	1048.5 [1]
G	453.2	517.8	609.5 [1], 535 [2]

its molar heat capacity (C_v) at room temperature (300 K) is determined to be 6.94 J/mol/K. This value is comparable to the heat capacities of diamond (6.40 J/mol/K) and mp-47 (6.43 J/mol/K), as shown in Figure S33. As the last considered properties, we calculated the electronic structure of this polymorph. It was found that the structure is insulator with an indirect band gap of 5.06 eV. The calculated electronic band structure is shown in Figure S32.

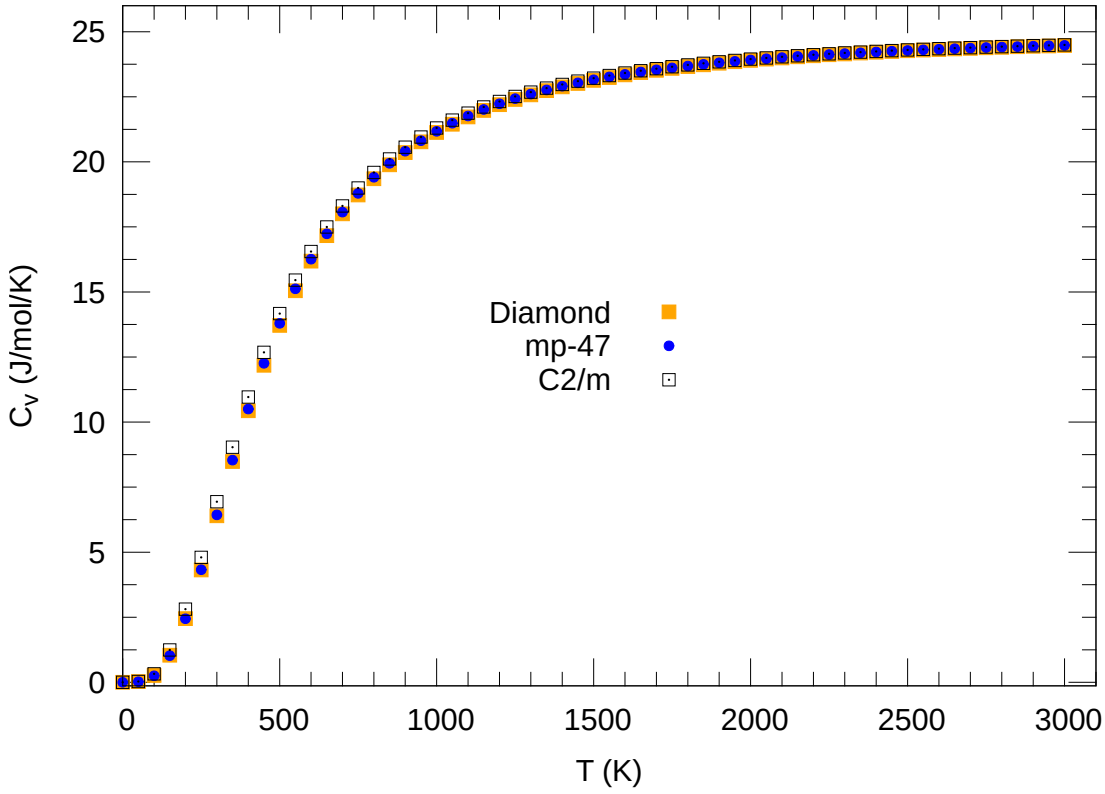

Figure S33: The molar heat capacity of the new carbon polymorph along with diamond and two other reported structures.

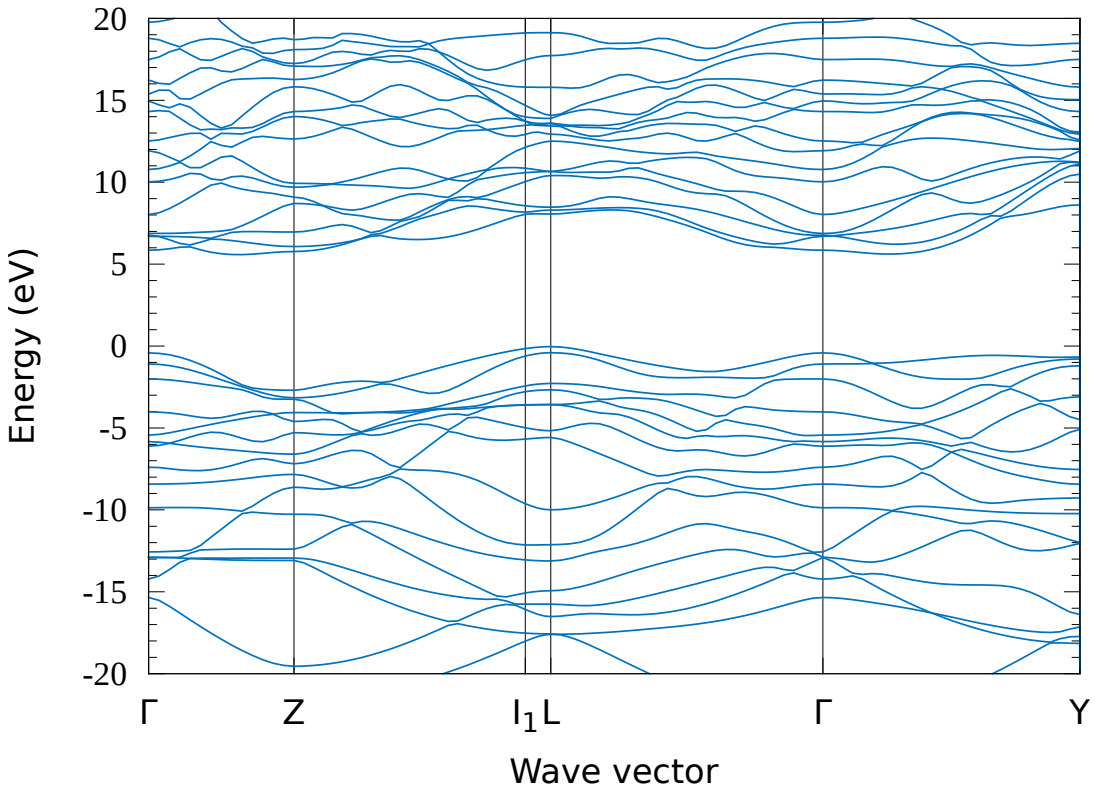

Figure S34: The electronic band structure of the new carbon polymorph with symmetry $C2/m$. The Fermi energy is set to zero.

References

1. Güler E and Gueler M. Elastic and mechanical properties of cubic diamond under pressure. Chinese Journal of physics 2015;53:195–205.

2. McSkimin H and Andreatch Jr P. Elastic moduli of diamond as a function of pressure and temperature. Journal of Applied Physics 1972;43:2944–8.