Electronic Supplementary Material (ESI) for Chemical Science.
This journal is © The Royal Society of Chemistry 2021

Supplementary Information for: Troubleshooting Unstable Molecules in Chemical Space

Salini Senthil, Sabyasachi Chakraborty, and Raghunathan Ramakrishnan*

Tata Institute of Fundamental Research, Centre for Interdisciplinary Sciences, Hyderabad 500107, India

E-mail: ramakrishnan@tifrh.res.in

Table of Contents

Technical Details:

Geometry optimizations of the 3k ambiguous molecules were done through the `ConnGO` workflow, consisting of 4 tiers. With the best control parameters, all tier-1 calculations involve generation of Cartesian coordinates in the SDF format from initial SMI using MMFF94[1] forcefield with steepest descent (SD) minimizer with a tight threshold of 10^{-8} kcal/mol for energy convergence using OpenBabel 2.3.2.[2] Further, the tier-2 geometry optimizations involve HF calculations with the STO-3G basis set. Tier-3 calculations were done with the 3-21G basis set and the B3LYP functional;[3] tier-4 involves geometry optimization at the B3LYP/6-31G(2*df*,*p*) level. 229 molecules that fail `ConnGO` were resubmitted to the workflow with improved tier-4 settings: ωB97-XD[4]/6-31G(2*df*,*p*), ωB97-XD/def2-TZVPP and CCSD[5]/6-31G(2*df*,*p*).

PM6,[6] HF, DFT, MP2[7] and CCSD calculations were performed using Gaussian16[8] and PM7[9] calculations were performed with MOPAC.[10] All Gaussian16 calculations used the following settings: SCF iterations with `maxcycles=100` and `tight` geometry optimization with `calcall`, along with a `Ultrafine` integration grid. To allow connectivities in the input file, `geom=connectivity` keyword was used. Final geometries at each tier was checked for being a local minimum on the PES through harmonic frequency analysis. An exception was made in the case of calculations performed with CCSD/6-31G(2*df*,*p*) tier-4 owing to prohibitive computational costs; here, for the 43 molecules passed vibrational analysis was done at the ωB97-XD/def2-TZVPP level. Out of 43, one molecule (QM9 index, 020482) showed an imaginary wavenumber of small magnitude ($|\overline{\nu}| \approx 20$ cm^{-1}), which was still considered a *pass* in `ConnGO`. 3054 uncharacterized molecules can be classified into 1262 zwitterions and 1792 non-zwitterions, the latter set was used for benchmarking. GAFF[11] and UFF[12] results are based on the implementations in OpenBabel.

Figures S4 and S5 present pass/fail statistics for various workflows. Successful optimization at a tier is quantified using the maximum absolute deviation (MaxAD) of bond lengths corresponding to covalent bonding connectivities and their mean percentage absolute deviation (MPAD). Molecules showing MaxAD < 0.2 Å and MPAD $< 5\%$ are deemed *pass*. For the initial tiers (HF, PM7, PM6) MaxAD criteria is waived to allow for unusual structures with long bond lengths (> 1.7 Å). Flow-1 (Figure S4) depicts a typical geometry optimization workflow, where force field geometries are used as an initial guess for DFT-level relaxation. In the best performing workflow (Flow-3a), 98% of 1792 non-zwitterions pass at tier-2 incurring a computational overhead of only 3% compared to Flow-1. The remaining 2% of the molecules require curation at a tier-3 level costing 20% in addition. On an average, the ConnGO workflow (Flow-3a) ensures maximum number of successful geometry relaxations preserving the intended covalent bonding connectivities with an overhead of only 3%.

LASSO regression

Least Absolute Shrinkage and Selection Operator (LASSO) method is a linear regression which uses L_1-regularization. For a common linear regression model the cost function is as follows:

$$\text{cost} = \langle \mathbf{y} - \mathbf{y}^{\text{est.}} | \mathbf{y} - \mathbf{y}^{\text{est.}} \rangle = \langle \mathbf{y} - \mathbf{A}\mathbf{c} | \mathbf{y} - \mathbf{A}\mathbf{c} \rangle$$

where $\mathbf{y}$ is a vector of target values, $\mathbf{y}^{\text{est.}}$ is the estimated target, $\mathbf{c}$ is a vector of regression coefficients while $\mathbf{A}$ is the covariance matrix. Linear regression results in a dense model with all elements of the coefficient vector taking non-zero values.

To avoid over-fitting ridge and LASSO regressions were developed. Here an additional penalty/regularization term is added to the cost. In the ridge regression, the penalty term is an L_2 norm of $\mathbf{c}$, while in LASSO it is the L_1 norm

$$\text{cost}^{\text{LASSO}} = \langle \mathbf{y} - \mathbf{y}^{\text{est.}} | \mathbf{y} - \mathbf{y}^{\text{est.}} \rangle = \langle \mathbf{y} - \mathbf{A}\mathbf{c} | \mathbf{y} - \mathbf{A}\mathbf{c} \rangle + \lambda ||\mathbf{c}||_1,$$

where λ is the regularization strength. With increase in the magnitude of λ, the number of non-zero elements of $\mathbf{c}$ decreases making the LASSO model sparser. Hence, LASSO is a preferred method for feature selection. In this study, we use LASSO regression as implemented in the Scikit-learn python package[13] to identify key features or chemical signatures necessary for a successful classification between stable and uncharacterized subsets of the 133885 QM9 molecules.

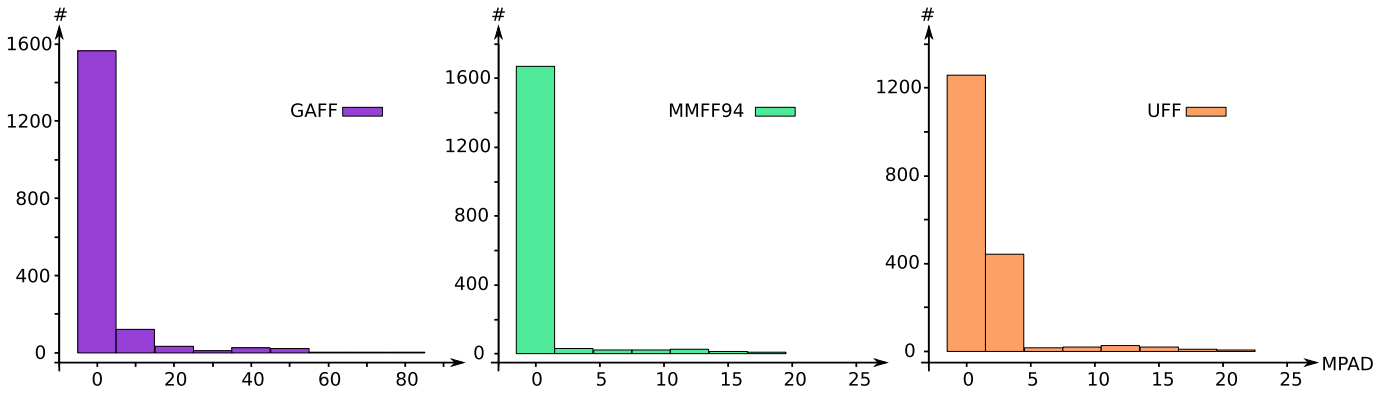

Figure S1: Performance of the force fields—GAFF, MMFF94 and UFF—with OpenBabel default settings. Mean Percentage Absolute Deviation (MPAD) of force field bond lengths from PM7 reference values are reported.

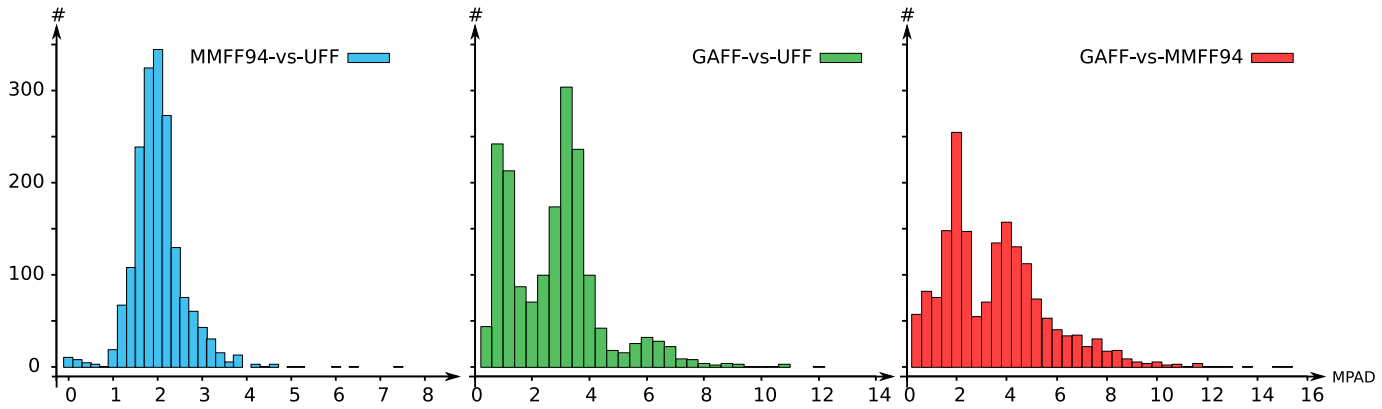

Figure S2: Pairwise Mean Percentage Absolute Deviation (MPAD) of bond distances between force fields. Results are shown for 1,792 subset of 3,054 unstable molecules.

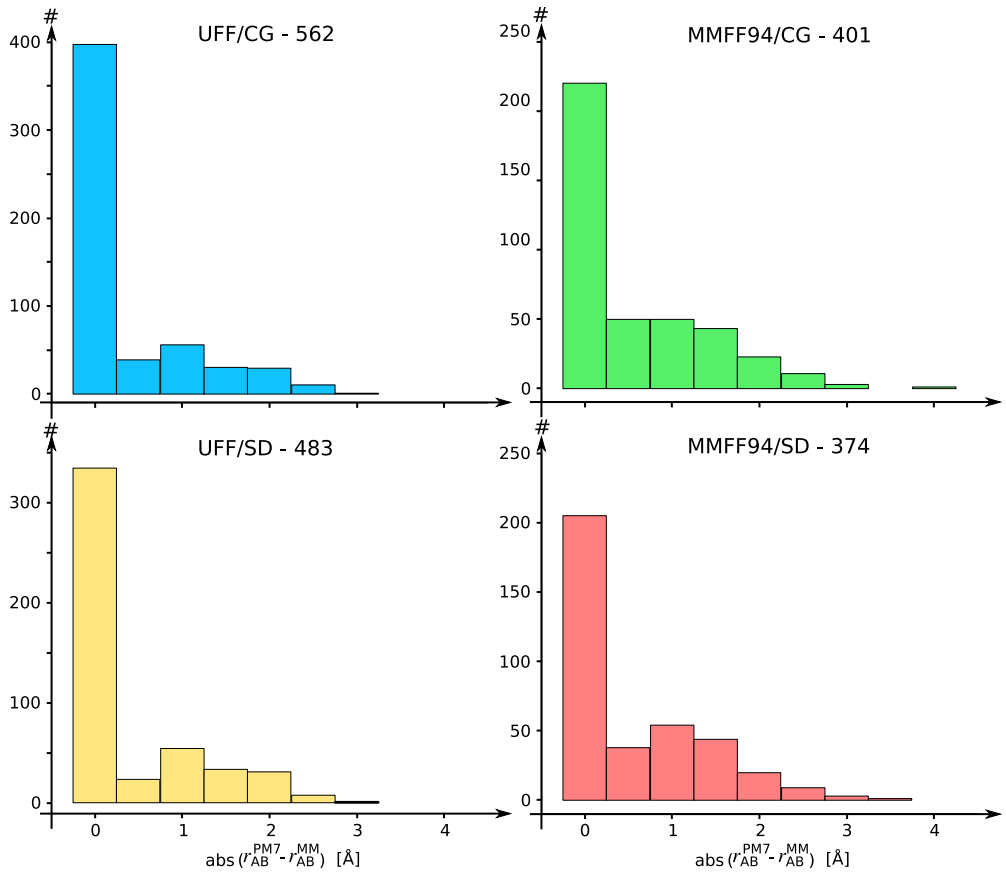

Figure S3: Joint performance of the minimizers: conjugate gradient (CG) and steepest descent (SD); and force fields: UFF and MMFF94 for tier-1 calculations. Absolute deviation of force field bond lengths from PM7 reference values are collected. Number of bond length values showing more than 10% deviation is reported at the top of each panel.

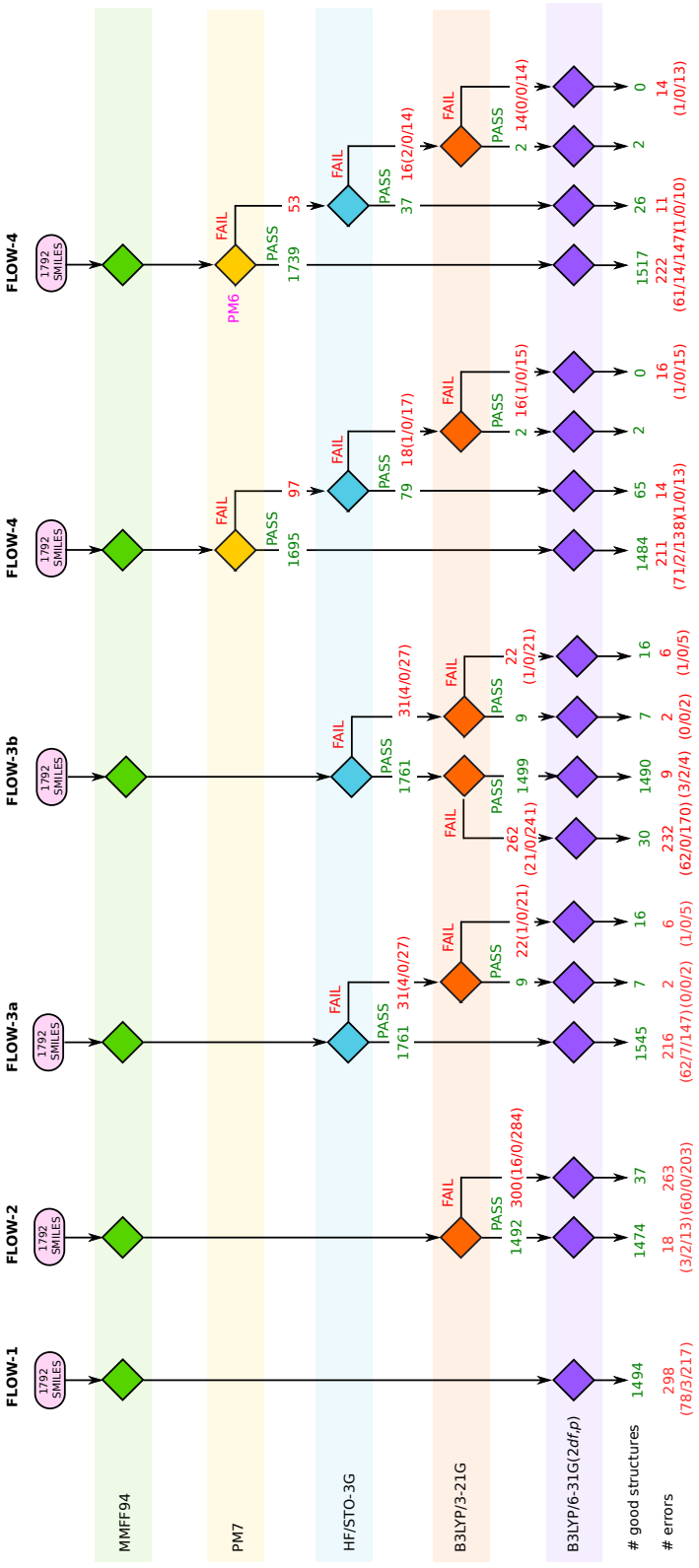

Figure S4: Performance of combinations of Tiers for 1,792 non-zwitterions. `ConnGO` pass/fail statistics are reported: Number of passes is denoted in green and number of fails are in red. For failed cases a breakdown of the number is also given in parenthesis, (Geometry error terminations/Presence of imaginary frequencies/Failing `ConnGO` criteria). PM6[6] values are based on an implementation in Gaussian-16.[8]

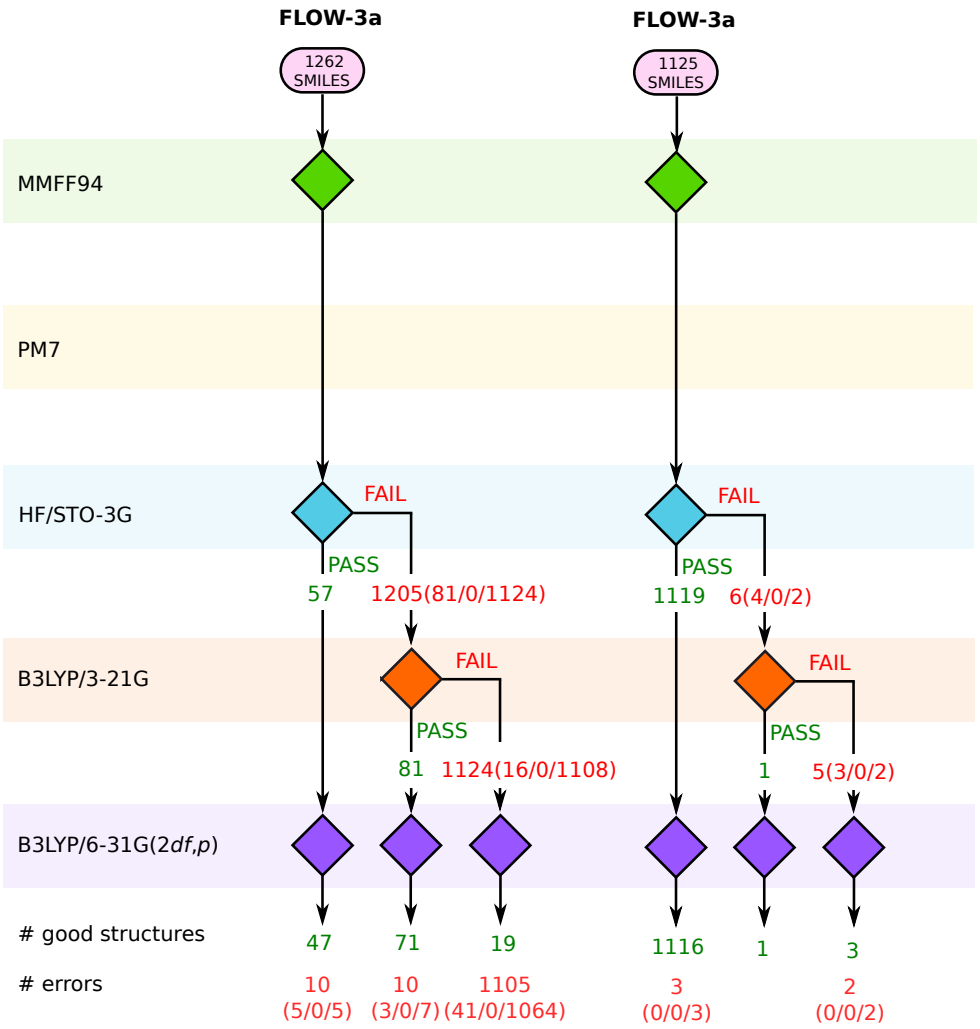

Figure S5: `ConnGO` pass/fail statistics for the 1,262 zwitterions: Left flow corresponds to results with original QM9 zwitterionic SMILES and the right corresponds to statistics based on the modified SMILES of those failing the left flow. Number of passes is denoted in green and number of fails are in red. For failed cases a breakdown of the number is also given in parenthesis, (Geometry error terminations/Presence of imaginary frequencies/Failing `ConnGO` criteria).

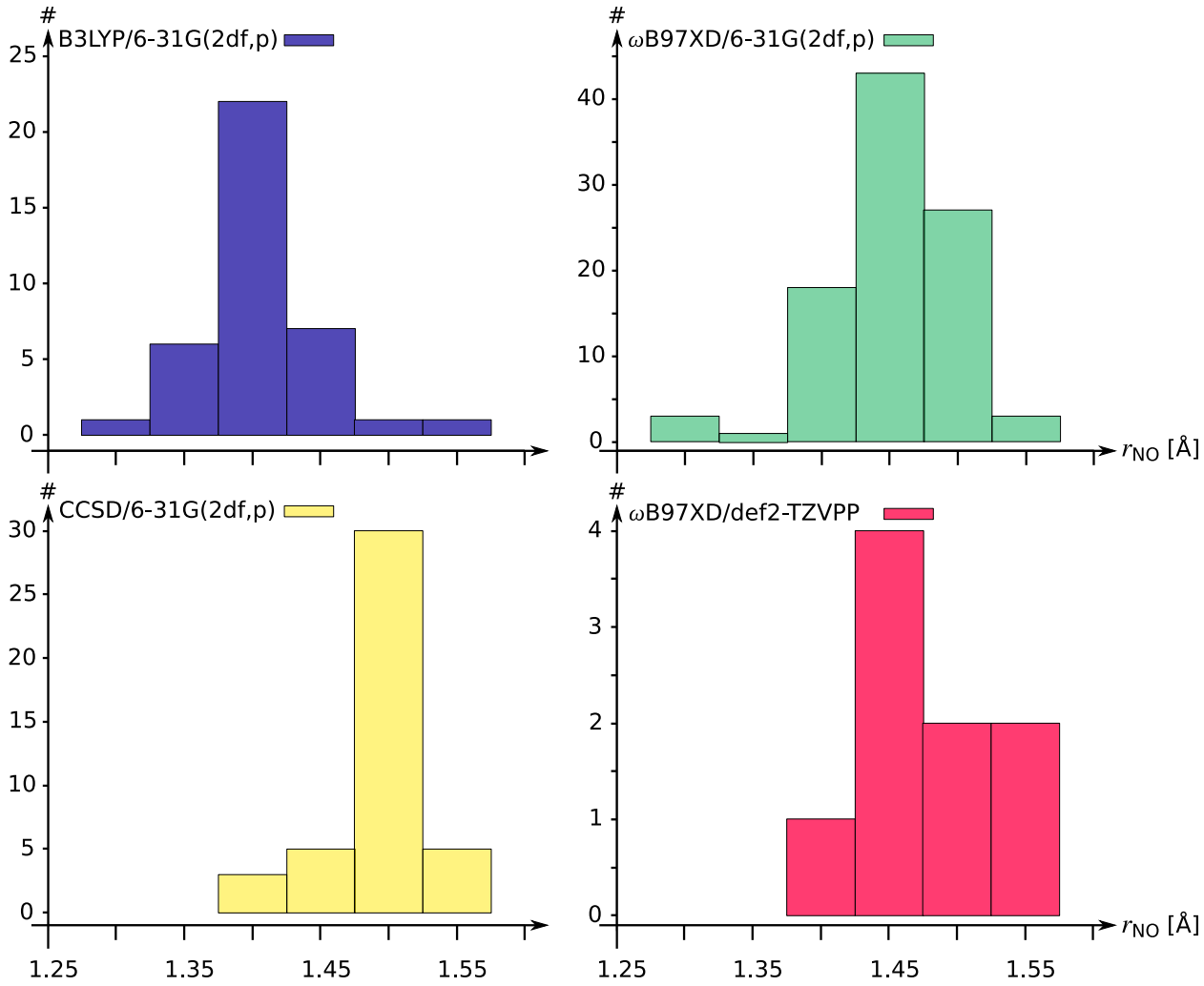

Figure S6: NO bond distances in 185 NNO containing molecules from the 3,054 set that passed through `ConnGO` collected for various tier-4 choices. 38 systems passed at B3LYP/6-31G(2*df*,*p*), 95 systems passed at ωB97XD/6-31G(2*df*,*p*), 43 systems passed at CCSD/6-31G(2*df*,*p*), an 9 passed at ωB97XD/def2-TZVPP level.

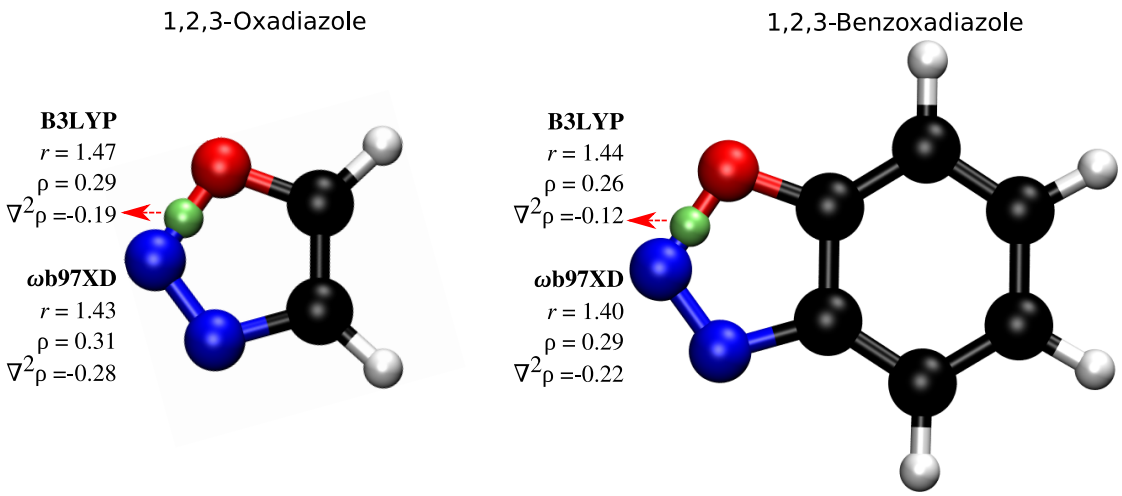

Figure S7: B3LYP and ωB97XD values of NO bond lengths and topological features of the NO bond in 1,2,3-oxadiazole and 1,2,3-benzoxadiazole: Bond lengths (r in Å), density (ρ) and laplacian of density ($\nabla^2\rho$). Green sphere indicates a bond critical point. All calculations were performed using the def2-TZVPP basis set.

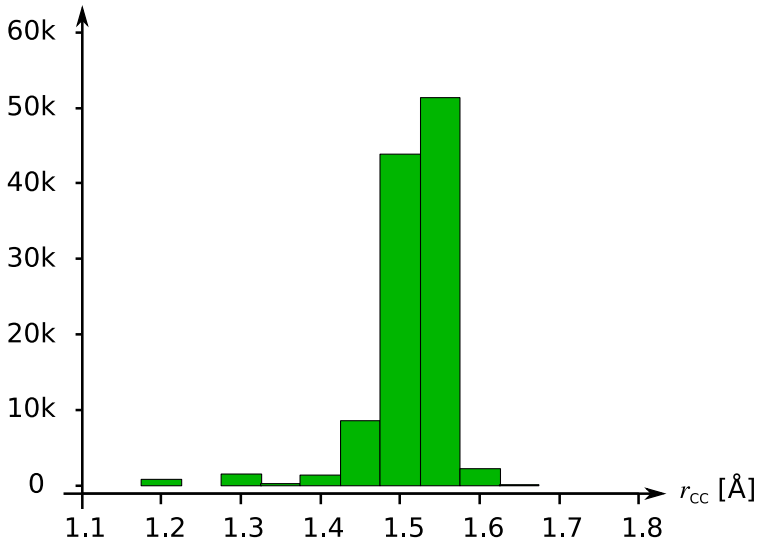

Figure S8: All CC bond distances (r_{CC}) for molecules in 130,831 set converged at B3LYP/6-31G(2*df*,*p*).

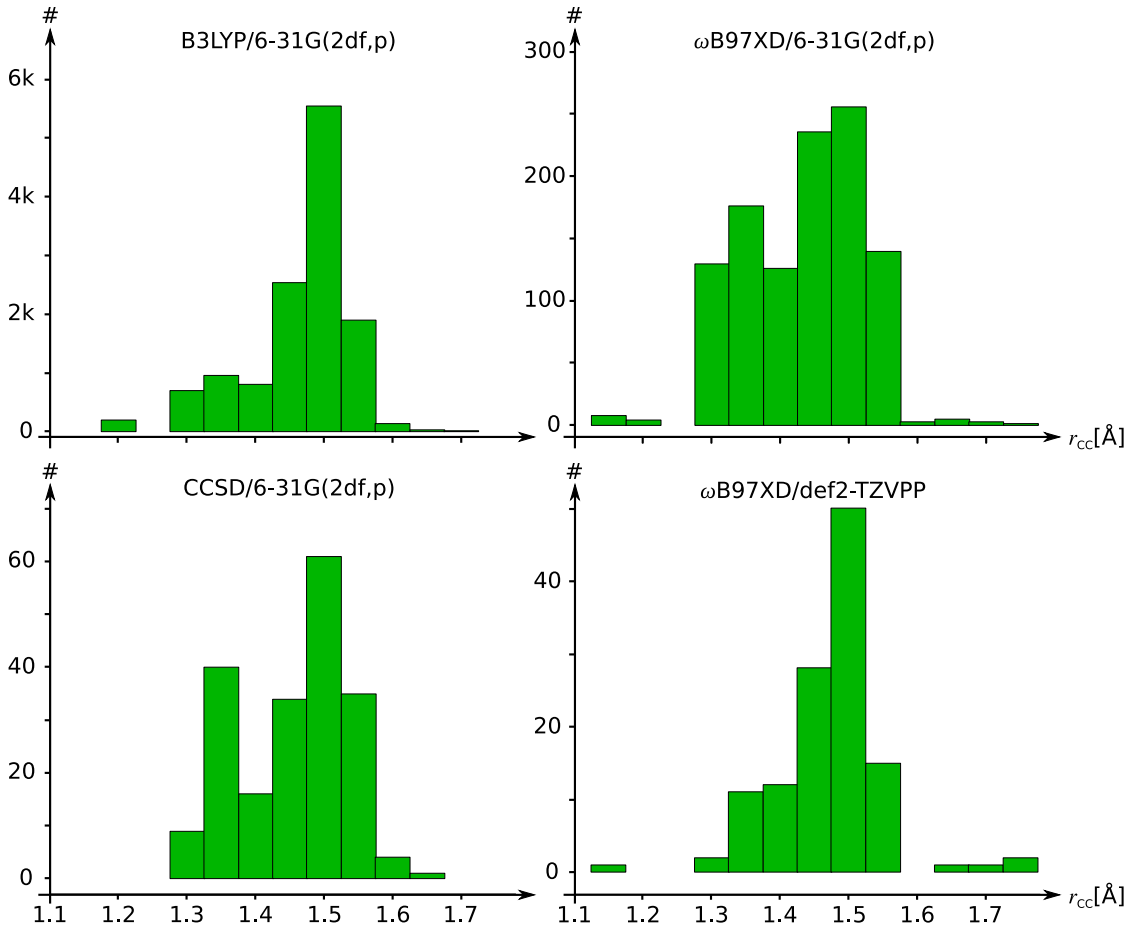

Figure S9: CC bond distances (r_{CC} in Å) for 2,988 molecules from the 3,054 set converged with `ConnGO` using various tier-4 levels.

Table S1: All possible 2-atom fragments in 134k QM9 molecules and their distribution in 1792 non-zwitterionic molecules in the 3k uncharacterized subset of QM9. Distribution in 1792 molecules of similar stoichiometries in the stable 131k subset of QM9 is given in parenthesis. Triple bonds are denoted by #.

fragments	3k (131k)
H-C	1660(1668)
H-N	1283(1534)
H-O	653(317)
C-C	1471(1508)
C-N	1507(1603)
C-O	757(1189)
C-F	234(235)
C=C	993(978)
C=N	1533(1550)
C=O	605(434)
C#C	133(95)
C#N	208(112)
N-N	159(323)
N-O	551(353)
N=N	188(304)
N=O	1(3)

Table S2: All possible 3-atom fragments in 134k QM9 molecules and their distribution in 1792 non-zwitterionic molecules in the 3k uncharacterized subset of QM9. Distribution in 1792 molecules of similar stoichiometries in the stable 131k subset of QM9 is given in parenthesis. Triple bonds are denoted by #.

fragments	3k (131k)	fragments	3k (131k)	fragments	3k (131k)
H-C-H	1081(1172)	C-C-N	1013(1120)	C=N-N	79(241)
H-C-C	1270(1264)	C-C-O	362(657)	C=N-O	512(136)
H-C-N	1001(934)	C-C-F	117(130)	N-C-N	601(404)
H-C-O	412(604)	C-C=C	694(823)	N-C-O	399(532)
H-C=C	707(742)	C-C=N	1048(971)	N-C-F	115(75)
H-C=N	562(444)	C-C=O	327(304)	N-C=N	922(1232)
H-C=O	92(168)	C-C#C	133(95)	N-C=O	457(238)
H-C#C	111(80)	C-C#N	208(112)	N-N-N	79(14)
H-C#N	0(0)	C-N-C	932(929)	N-N=N	128(87)
H-N-H	647(528)	C-N-N	133(314)	N-O-N	31(14)
H-N-C	1140(1047)	C-N=C	897(579)	N=C-O	416(750)
H-N-N	68(201)	C-N=N	139(271)	N=C-F	103(91)
H-N=C	448(1383)	C-N=O	1(3)	N=N-O	65(229)
H-N=N	0(0)	C-O-C	498(773)	O-C-O	70(224)
H-O-H	0(0)	C-O-N	55 (339)	O-C-F	24(48)
H-O-C	251(317)	C=C-N	839(665)	O-C=O	130(54)
H-O-N	453(0)	C=C-O	397(637)	O=N=O	1(3)
C-C-C	793(715)	C=C-F	141(152)	F-C-F	1(2)

Table S3: All possible 4-atom fragments in 134k QM9 molecules and their distribution in 1792 non-zwitterionic molecules in the 3k uncharacterized subset of QM9. Distribution in 1792 molecules of similar stoichiometries in the stable 131k subset of QM9 is given in parenthesis. Triple bonds are denoted by #.

fragments	3k (131k)	fragments	3k (131k)	fragments	3k (131k)	fragments	3k (131k)
H-C-C-H	498(552)	H-O-C=C	107(116)	C-O-C=O	106(51)	N-C=N-O	27(80)
H-C-C-C	760(649)	H-O-C=N	83(84)	C-O-N=C	124(300)	N-N-C-O	18(64)
H-C-C-N	659(644)	H-O-N=C	453(0)	C-O-N=N	41(218)	N-N-C-F	9(6)
H-C-C-O	218(439)	C-C-C-C	455(320)	C=C-C-N	271(415)	N-N-C=N	42(259)
H-C-C-F	44(41)	C-C-C-N	457(482)	C=C-C-O	134(186)	N-N-C=O	26(31)
H-C-C=C	398(501)	C-C-C-O	167(274)	C=C-C-F	41(30)	N-N-N=N	73(8)
H-C-C=N	794(506)	C-C-C-F	18(21)	C=C-C=C	186(128)	N-N=C-O	13(56)
H-C-C=O	170(160)	C-C-C=C	298(293)	C=C-C=N	354(491)	N-N=C-F	7(14)
H-C-C#C	71(41)	C-C-C=N	474(321)	C=C-C=O	122(63)	N-N=N-N	41(18)
H-C-C#N	115(35)	C-C-C=O	122(138)	C=C-C#C	15(32)	N-N=N-O	7(14)
H-C-N-H	553(342)	C-C-C#C	117(47)	C=C-C#N	24(23)	N-O-C-O	3(89)
H-C-N-C	709(723)	C-C-C#N	135(58)	C=C-N-N	28(48)	N-O-C-F	1(12)
H-C-N-N	74(153)	C-C-N-C	575(586)	C=C-N=C	526(168)	N-O-C=N	23(156)
H-C-N=C	356(228)	C-C-N-N	51(132)	C=C-N=N	25(147)	N-O-C=O	21(4)
H-C-N=N	15(27)	C-C-N=C	287(219)	C=C-N=O	0(2)	N-O-N=N	24(11)
H-C-O-H	58(90)	C-C-N=N	44(150)	C=C-O-N	13(179)	N=C-C-O	146(174)
H-C-O-C	349(495)	C-C-N=O	0(0)	C=N-C-N	365(222)	N=C-C-F	61(99)
H-C-O-N	9(37)	C-C-O-C	251(474)	C=N-C-O	158(63)	N=C-C=N	187(186)
H-C=C-H	310(219)	C-C-O-N	18(68)	C=N-C-F	59(33)	N=C-C=O	118(155)
H-C=C-C	465(580)	C-C=C-C	231(314)	C=N-C=N	318(345)	N=C-C#N	32(28)
H-C=C-N	519(323)	C-C=C-N	520(442)	C=N-C=O	128(27)	N=C-N=N	100(76)
H-C=C-O	174(308)	C-C=C-O	209(430)	C=N-N-N	53(23)	N=C-N=O	1(1)
H-C=C-F	60(109)	C-C=C-F	93(102)	C=N-N=C	85(31)	N=N-C-O	17(16)
H-C=N-H	64(35)	C-C=N-C	405(310)	C=N-N=N	49(8)	N=N-C-F	11(5)
H-C=N-C	481(272)	C-C=N-N	40(181)	C=N-O-N	30(37)	N=N-C=O	16(51)
H-C=N-N	27(119)	C-C=N-O	497(83)	C#C-C-N	28(40)	N=N-N=N	28(2)
H-C=N-O	13(31)	C-C#C-C	32(16)	C#C-C-O	13(27)	N#C-C-O	33(34)
H-C#C-H	0(0)	C-N-C-N	450(280)	C#C-C-F	0(0)	N#C-C-F	0(0)
H-C#C-C	111(80)	C-N-C-O	176(408)	C#C-C=N	50(16)	N#C-C=O	7(3)
H-N-C-C	682(627)	C-N-C-F	41(32)	C#C-C=O	5(11)	N#C-C#N	0(0)
H-N-C-N	521(272)	C-N-C=C	515(411)	C#C-C#C	6(1)	O-C-C-O	33(100)
H-N-C-O	236(309)	C-N-C=N	564(756)	C#C-C#N	3(1)	O-C-C-F	11(9)
H-N-C-F	43(35)	C-N-C=O	295(168)	N-C-C-N	264(253)	O-C-C=O	23(60)
H-N-C=C	627(413)	C-N-N=C	179(380)	N-C-C-O	98(267)	O-C-N=O	1(2)
H-N-C=N	660(840)	C-N-N=N	114(87)	N-C-C-F	42(61)	O-C=C-O	51(105)
H-N-C=O	323(74)	C-N=C-N	657(340)	N-C-C=N	431(280)	O-C=C-F	24(30)
H-N-N-C	0(0)	C-N=C-O	298(271)	N-C-C=O	128(185)	O-C=N-O	8(48)
H-N-N=C	40(156)	C-N=C-F	96(78)	N-C-C#N	89(54)	O-N=C-F	1(0)
H-N-N=N	44(48)	C-N=N-C	48(11)	N-C-N-N	36(101)	O-N=N-O	1(1)
H-N=C-C	261(696)	C-N=N-N	80(56)	N-C-N=N	60(82)	O=C-C-F	17(6)
H-N=C-N	350(1110)	C-N=N-O	57(214)	N-C-N=O	0(1)	O=C-C=O	7(18)
H-N=C-O	147(518)	C-O-C-N	258(355)	N-C-O-N	27(133)	F-C-C-F	3(3)
H-N=N-C	0(0)	C-O-C-O	78(228)	N-C=C-N	448(169)	F-C=C-F	10(10)
H-O-C-C	99(187)	C-O-C-F	23(36)	N-C=C-O	263(389)		
H-O-C-N	133(79)	C-O-C=C	307(451)	N-C=C-F	76(51)		
H-O-C-O	24(58)	C-O-C=N	328(590)	N-C=N-N	57(72)		

Table S4: Long CC bonds (in Å) for molecules highlighted in Fig. 6 and Table. 1 computed at B2PLYP,[14] PBE0,[15] M06-2X[16] and MP2 methods with def2-TZVPP basis set. Missing values indicate dissociating molecules. QM9 indices and SMILES are also given.

Index	SMI	B2PLYP	PBE0	M06-2X	MP2
037488	C1C2C3C1C1=CCN3C21	1.735	1.692	1.701	1.718
037497	C1C2C3N1C1CC3=CC21	1.747	1.701	1.717	1.708
037750	C1C2C3C=C4C=CC13C24			1.751	
037993	C1C2N=C3C=C1C1C2N31	1.686	1.659		1.736
038078	C1C2C3OC4C=CC2C134	1.731	1.682	1.683	1.699
039170	C1C2C3OC4COC2C134		1.725	1.722	1.759
066495	CC12C3C4OC(=NC14)N23		1.699	1.701	
066503	CC12C3C4N=C(OC14)N23		1.698	1.7	1.73
133863	C1C2CC3CC4CC1C234	1.806	1.734	1.738	
133816	N1C2C3OC4C3C1=NC24	1.725	1.691	1.695	1.712

List S1: List of QM9 indices and SMILES of the 66 unstable molecules

003899 O=C1N=CON=N1
020692 CC1=NC(=O)N=NO1
021567 NC1=CC(=O)N=NO1
021611 NC1=NC(=N)N=NO1
021620 NC1=NC(=O)N=NO1
021748 O1C=NC2=C1ON=N2
021764 O1N=NC2=C1ON=N2
021859 OC1=CC(=O)N=NO1
021870 OC1=NC(=N)N=NO1
021875 OC1=NC(=O)N=NO1
023793 FC1=CC(=N)N=NO1
023799 FC1=CC(=O)N=NO1
023822 FC1=NC(=N)N=NO1
023828 FC1=NC(=O)N=NO1
037079 C1C2=CC3C=C1C1C2N31
037080 C1C2=CC3C=C1C1C3C21
037082 N1C2=CC3N=C1C1C3C21
038196 C1C2CC11C=CC3C1N23
125055 CC1=C(N)ON=NC1=O
125471 CC1=NC2=C(ON=N2)O1
125793 CC1=NOC2=C1N=NO2
126360 CCC1=NC(=O)N=NO1
127356 CN1C=NC2=C1ON=N2
127778 CN1N=NC2=C1ON=N2
127894 CN=C1N=NOC(N)=N1
127895 CN=C1N=NOC(O)=N1
128195 CNC1=CC(=O)N=NO1
128252 CNC1=NC(=N)N=NO1
128260 CNC1=NC(=O)N=NO1
128558 COC1=CC(=N)N=NO1
128568 COC1=CC(=O)N=NO1
128619 COC1=NC(=N)N=NO1
128627 COC1=NC(=O)N=NO1
129106 N=C1CC2=C(N1)ON=N2

129137 N=C1N=CON=NC1=O
129146 N=C1N=NOC(=N1)C#N
129654 NC1=C(N)C(=O)N=NO1
130009 NC1=NC2=C(ON=N2)O1
130424 NCC1=NC(=O)N=NO1
130603 O=C1C=CON=NC1=O
130630 O=C1CC2=C(N1)ON=N2
130662 O=C1N=CON=NC1=O
130664 O=C1N=NOC(=N1)C#C
130665 O=C1N=NOC(=N1)C#N
130667 O=C1N=NOC=C1C#N
130879 O=CC1=CON=NC1=O
130883 O=CC1=NC(=O)N=NO1
131555 OC1COC2=C1N=NO2
131696 OCC1=NC(=N)N=NO1
131705 OCC1=NC(=O)N=NO1
133168 CC1=C(F)ON=NC1=O
133263 CN=C1C=C(F)ON=N1
133274 CN=C1N=NOC(F)=N1
133311 NC1=C(F)C(=O)N=NO1
133403 OC1=C(F)C(=O)N=NO1
133404 FC1=C(F)C(=O)N=NO1
133828 CC12C3C1C1=CC2CC31
133829 CC12C3C1C1=CC2OC31
133840 CC12C3C1C1=CC3CC21
133843 CC12C3N1C1CC2C3=C1
133844 OC12C3C1C1=CC3CC21
133846 CC12C3C1C1CC3C2=C1
133847 CC12C3C1C1OC3C2=C1
133849 CC12C3C4CC(C=C14)N23
133851 OC12C3C1C1CC3C2=C1
133852 OC12C3C1C1OC3C2=C1

List S2: Coordinates of Top-10 molecules with long CC bonds illustrated in Fig. 6.

```
18
dsgdb9_037488_b3lyp_2dfp
C       -0.09085457       1.543870320       -0.05421690
C       1.428627360       1.450595230       0.108221900
C       1.397701730       0.032166670       0.721561070
C       -0.06191308       0.016641720       -0.23499368
C       0.783257450       -0.32446067       -1.40780213
C       1.216724210       -1.58942338       -1.58409744
C       2.580019480       -1.64800841       -0.88739028
N       2.590129210       -0.37116734       -0.08865773
C       1.998872680       0.596117940       -1.05823714
H       -0.57749124       1.843867260       0.879104340
H       -0.49194178       2.137217540       -0.88140593
H       2.056444250       2.248240480       0.500614750
H       1.323802270       -0.24980421       1.771295320
H       -0.89188143       -0.63015586       0.042330410
H       0.548436530       -2.44534885       -1.56132781
H       2.632625100       -2.54313208       -0.25789713
H       3.481969720       -1.63009416       -1.51161295
H       2.626659620       0.982675260       -1.86246856
18
dsgdb9_037497_b3lyp_2dfp
C       -0.04464006       1.478751550       -0.06822999
C       1.483508890       1.400965620       0.093611320
C       1.240442500       0.050937530       0.833179340
N       -0.04867996       -0.01582377       -0.04561453
C       0.646560530       -0.43224188       -1.25047917
C       1.317937040       -1.79038962       -0.73587274
C       2.209627490       -0.99019095       0.255451440
C       2.942459810       -0.31862114       -0.71324301
C       1.987047790       0.708548650       -1.17952557
H       -0.56017502       1.867147670       0.814195420
H       -0.45451479       1.935391330       -0.97555313
H       2.088365880       2.219103690       0.486610010
```

```
H       1.006830140       0.048373570       1.896008280
H       0.003081470       -0.42217420       -2.13575630
H       0.554039680       -2.43819863       -0.29585573
H       1.868007130       -2.33222987       -1.50684817
H       3.583664990       -0.85287949       -1.41193926
H       2.197708740       1.316136790       -2.05951262
17
dsgdb9_037750_wb97xd_2dfp
C       -1.82078600       -0.86943000       -0.16506800
C       -1.50423400       0.567164000       0.393514000
C       -0.57371300       0.898634000       -0.77275800
C       0.720867000       1.508906000       -0.35647300
C       1.075283000       0.678650000       0.668419000
C       1.926559000       -0.45820600       0.177108000
C       1.071919000       -1.32384100       -0.43961900
C       -0.29233400       -0.74917600       -0.29245900
C       -0.22941200       -0.00507600       1.042803000
H       -2.40672700       -0.91630400       -1.08753500
H       -2.18437900       -1.59954600       0.564145000
H       -2.21973300       1.259709000       0.840707000
H       -0.97914400       1.109964000       -1.76246600
H       1.427707000       1.901433000       -1.07974800
H       3.005818000       -0.47204900       0.099616000
H       1.380222000       -2.15719600       -1.05680900
H       -0.26865300       -0.61177000       1.949282000
15
dsgdb9_037993_b3lyp_2dfp
C       -1.09831400       0.620634000       1.082334000
C       -0.19237900       1.287335000       -0.20728500
N       -0.58531100       0.568381000       -1.39673300
C       -0.02487100       -0.58668500       -1.05788100
C       -0.96120800       -1.31969900       -0.25426900
C       -0.52735200       -0.74945800       0.963417000
C       0.985024000       -0.55720100       0.901843000
C       1.215384000       0.743873000       0.184054000
N       1.316872000       -0.54022900       -0.54427100
```

```
H       -2.16807100       0.732378000       0.899445000
H       -0.78424800       1.165453000       1.979154000
H       -0.33659800       2.364390000       -0.27187900
H       -1.99475400       -1.42655800       -0.55403500
H       1.681750000       -1.02409700       1.587749000
H       2.103288000       1.358580000       0.273320000
17
dsgdb9_038078_b3lyp_2dfp
C       1.474340000       -0.83788000       0.934414000
C       1.552649000       -0.56769300       -0.61435000
C       1.041169000       1.040645000       -0.21456500
O       -0.26747700       1.699432000       -0.51054700
C       -0.99996100       0.829735000       0.596901000
C       -1.95795700       -0.31462900       0.241308000
C       -1.30134000       -1.34714600       -0.34034900
C       0.028932000       -0.73439100       -0.73931600
C       0.308318000       0.098360000       0.614022000
H       2.262099000       -0.42286300       1.568292000
H       1.251129000       -1.87512300       1.193254000
H       2.339904000       -0.92076000       -1.28796800
H       1.785542000       1.770222000       0.106351000
H       -1.34764800       1.544279000       1.343856000
H       -3.01361200       -0.29765100       0.487389000
H       -1.74148200       -2.29587100       -0.62660600
H       -0.27302400       -0.09968600       -1.56858600
17
dsgdb9_039170_wb97xd_def2tzvpp
C       1.846348000       0.166897000       -0.68821000
C       1.544051000       -0.48930900       0.670208000
C       0.352705000       0.462865000       0.864306000
O       -0.23129400       1.771669000       0.464647000
C       -0.86177500       0.959835000       -0.59554200
C       -1.83631400       -0.19490800       -0.12619400
O       -1.09437800       -1.40249900       0.255227000
C       0.247210000       -1.30369900       -0.17801600
C       0.329989000       0.079067000       -0.66121200
```

```
H       2.342028000       -0.43774900       -1.44753300
H       2.273894000       1.164338000       -0.62196100
H       2.265395000       -0.87225600       1.394869000
H       -0.35160800       0.064099000       1.580007000
H       -1.27350900       1.590648000       -1.38358200
H       -2.47863200       -0.45804300       -0.96884400
H       -2.45343200       0.097342000       0.721829000
H       0.547955000       -2.18622600       -0.74581600
15
dsgdb9_066495_b3lyp_2dfp
C       -0.02178584       1.536179490       -0.04032011
C       0.034124270       0.046500140       0.005615030
C       0.306010160       -1.00365210       -1.02682766
C       0.775875700       -2.06824149       -0.03114051
O       -0.36243550       -2.94228614       0.151032050
C       -0.98954529       -1.78863783       0.637094420
N       -0.85082035       -1.39357895       1.856272070
C       0.419550880       -0.89209955       1.199637000
N       -1.06755091       -0.81704136       -0.48746404
H       0.978726090       1.966431950       0.074387680
H       -0.44002951       1.878486820       -0.99067802
H       -0.65297992       1.914346820       0.769634630
H       0.479164520       -0.94306481       -2.09403110
H       1.708742210       -2.62101235       -0.06883603
H       1.179045270       -0.70602120       1.953512240

15
dsgdb9_066503_b3lyp_2dfp
C       0.006224030       1.549972660       -0.02468285
C       0.054356350       0.056799240       0.012216530
C       0.302978770       -0.98624023       -1.03274116
C       0.692737830       -2.09680364       -0.01182568
N       -0.48230282       -3.03123117       0.170605690
C       -0.99314901       -1.89050067       0.482733240
O       -0.80961687       -1.31556538       1.743493900
C       0.449371070       -0.85971067       1.185313350
```

```
N       -1.06499592       -0.77596148       -0.50370698
H       1.006450820       1.975025620       0.109032020
H       -0.39196244       1.898530860       -0.98096205
H       -0.63747007       1.932791480       0.773739930
H       0.510295440       -0.91217678       -2.09273108
H       1.623671220       -2.65456387       0.018262640
H       1.172626380       -0.63268887       1.962602960
15
dsgdb9_133816_b3lyp_2dfp
N       1.422553000       1.151831000       0.138189000
C       0.247317000       0.904135000       -0.77421800
C       -0.95361400       0.984033000       0.176181000
O       -1.93192500       0.052030000       -0.32954200
C       -1.00107100       -0.98026200       0.069914000
C       -0.37141800       -0.07634400       1.162351000
C       1.099261000       -0.16297900       0.639746000
N       1.481395000       -1.22660200       0.003687000
C       0.251027000       -0.83285900       -0.82875300
H       2.279938000       1.043636000       -0.40627600
H       0.280249000       1.463653000       -1.70532600
H       -1.34684500       1.956606000       0.461310000
H       -1.43962500       -1.96378000       0.214627000
H       -0.61451200       -0.11566200       2.217348000
H       0.339550000       -1.29163300       -1.80979900
21
dsgdb9_133863_b3lyp_2dfp
C       -1.25582376       1.489228780       -0.50362797
C       -1.32147518       0.012540650       -0.00507766
C       -1.28015789       -1.47310540       -0.47897955
C       -0.01617924       -1.74284580       0.385217040
C       1.264230300       -1.49343024       -0.46068534
C       1.322736290       -0.00876695       0.014526500
C       1.288556450       1.468913640       -0.48420228
C       0.012049270       1.752618590       0.356797190
C       -0.00420331       0.007276040       0.66266113
H       -2.13546851       2.081460510       -0.23981309
```

H	-1.06199846	1.624361810	-1.57094562
H	-2.05268427	0.025314930	0.812384920
H	-1.08916459	-1.62902682	-1.54396446
H	-2.16917134	-2.04649933	-0.20513520
H	-0.02790630	-2.57986159	1.083485390
H	1.086092600	-1.64604593	-1.52837980
H	2.139829970	-2.08115987	-0.17431012
H	2.041861330	-0.00806347	0.842737150
H	1.113250610	1.607346310	-1.55428963
H	2.173516770	2.046826690	-0.20675304
H	0.013619440	2.600959490	1.041360520

References

(1) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490–519.

(2) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **2011**, *3*, 33.

(3) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785.

(4) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(5) Scuseria, G. E.; Janssen, C. L.; Schaefer Iii, H. F. An efficient reformulation of the closed-shell coupled cluster single and double excitation (CCSD) equations. *J. Chem. Phys.* **1988**, *89*, 7382–7387.

(6) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Modeling* **2007**, *13*, 1173.

(7) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153*, 503–506.

(8) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. et al. Gaussian16 Revision C.01. 2016; `https://gaussian.com`, Gaussian Inc. Wallingford CT 2016.

(9) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Modeling* **2013**, *19*, 1.

(10) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **1990**, *4*, 1–103.

(11) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comp. Chem.* **2004**, *25*, 1157–1174.

(12) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

(13) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.

(14) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(15) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(16) Zhao, Y.; Truhlar, D. G. 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. Chim. Acta* **2008**, *120*, 215–241.