Supporting Information:

Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation

WooSeok Jeong[a,1], Samuel J. Stoneburner[a,b,1], Daniel King[a], Ruye Li[a,c], Andrew Walker[d], Roland Lindh[e], Laura Gagliardi[a]*

[a]Department of Chemistry, Nanoporous Materials Genome Center, Minnesota Supercomputing Institute, and Chemical Theory Center, University of Minnesota, 207 Pleasant Street Southeast, Minneapolis, Minnesota 55455, United States

[b]Current address: Department of Chemistry and Biochemistry, Messiah College, One College Avenue, Mechanicsburg, Pennsylvania 17055, United States

[c]Current address: Center of Environmental Science and New Energy Technology, Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen 518055, China

[d]Department of Computer Science and Engineering, University of Minnesota, 200 Union Street Southeast, Minneapolis, Minnesota 55455, United States

[e]Department of Chemistry—Ångström, The Theoretical Chemistry Programme, and Uppsala Center for Computational Chemistry—UC_3, Uppsala University, Box 518, 751 20 Uppsala, Sweden

[1]W.J., and S.J.S. contributed equally to this work.

*To whom correspondence should be addressed.

Table of Contents

S1. Raw Data Generation

Main group diatomic molecules were selected for the training set based on the availability of experimental reference data from the CRC handbook (bond dissociation energies)[1,2] and NIST (equilibrium bond distances and first vibrational constants).[3]

All training data was generated using CASSCF[4] and CASPT2[5,6] in MOLCAS 8.2[7] using the ANO-RCC-VTZP.[8] Cholesky decomposition[9] with the default threshold of 10^{-4} a.u. was used to simplify the calculation and storage of two-electron integrals. Spin was chosen to match the experimental ground state. Spatial symmetry was not employed, i.e., all calculations were in C_1. Potential energy curves (PECs) were calculated in two sets of single-point calculations. All calculations began at the experimental equilibrium bond distance ($r_{e,exp}$) using the MOLCAS "GssOrb" guess orbitals as input orbitals. Following an initial CASSCF calculation at $r_{e,exp}$, single-point CASSCF/CASPT2 calculations proceeded on a loop over a list of increasing or decreasing internuclear distances, with each distance using the MOLCAS "JobIph" binary file from the

previous calculation as input. The distance interval between data points was made small (0.004 Å) near $r_{e,exp}$ and gradually increased to 1.000 Å at large distances.

Active spaces were selected systematically such that every permitted combination of the total number of active orbitals and total number of active electrons was considered. The number of active electrons was allowed to be 2, 4, 6, 8, or 10 for even-electron systems and 1, 3, 5, 7, or 9 for odd-electron systems, or up to the total number of electrons in the molecule if that number was less than 10. The number of active orbitals was allowed to be any integer value above half the number of electrons and less than or equal to 10. For a given active space size, the specific orbitals were chosen so that the given number of electrons would be active without any manual reordering of input orbitals (i.e., through the use of the MOLCAS "ALTER" keyword). A consequence of this approach is that orbitals were selected based on their proximity to the HOMO and LUMO rather than properties such as their binding character or atomic orbital contributions. We chose this way of selecting active spaces for simplicity at this initial stage, but in future work we intend to expand the training set to include more variety within a given active space size.

The use of a post-MCSCF method such as CASPT2 was necessitated by the use of experimental reference data. We selected CASPT2 as it is the most popular post-MCSCF method, but our protocol could just as easily be used to train with other methods such as NEVPT2[10] or MC-PDFT.[11] For CASPT2 we used the default IPEA shift[12] of 0.25 a.u. to correct the zeroth-order Hamiltonian, and also used an imaginary shift[13] of 0.2 a.u. to minimize intruder states. While in principle our protocol could work with other settings for the IPEA shift if one were so inclined, we found that lower values of the imaginary shift led to significant increases in negative results due to discontinuous PECs.

For each active space of the diatomic molecules investigated, potential energy curves were obtained with CASPT2 energies unless CASSCF/CASPT2 calculations fail to converge (Figure S1). The bond dissociative/spectroscopic properties (i.e., the bond dissociation energy (D_e), the equilibrium bond length (r_e), the vibrational constants including the harmonic (ω_e)) from the computed potential energy curves were calculated using VIBROT module in MOLCAS[14]. The module solves the ro-vibrational Schrödinger equation numerically by fitting a potential energy curve using cubic splines. To obtain accurate the properties, the number of grid points and the internuclear distance range for the numerical solution were set to 1,000 and 1.0 to 10.0 Angstroms, respectively.

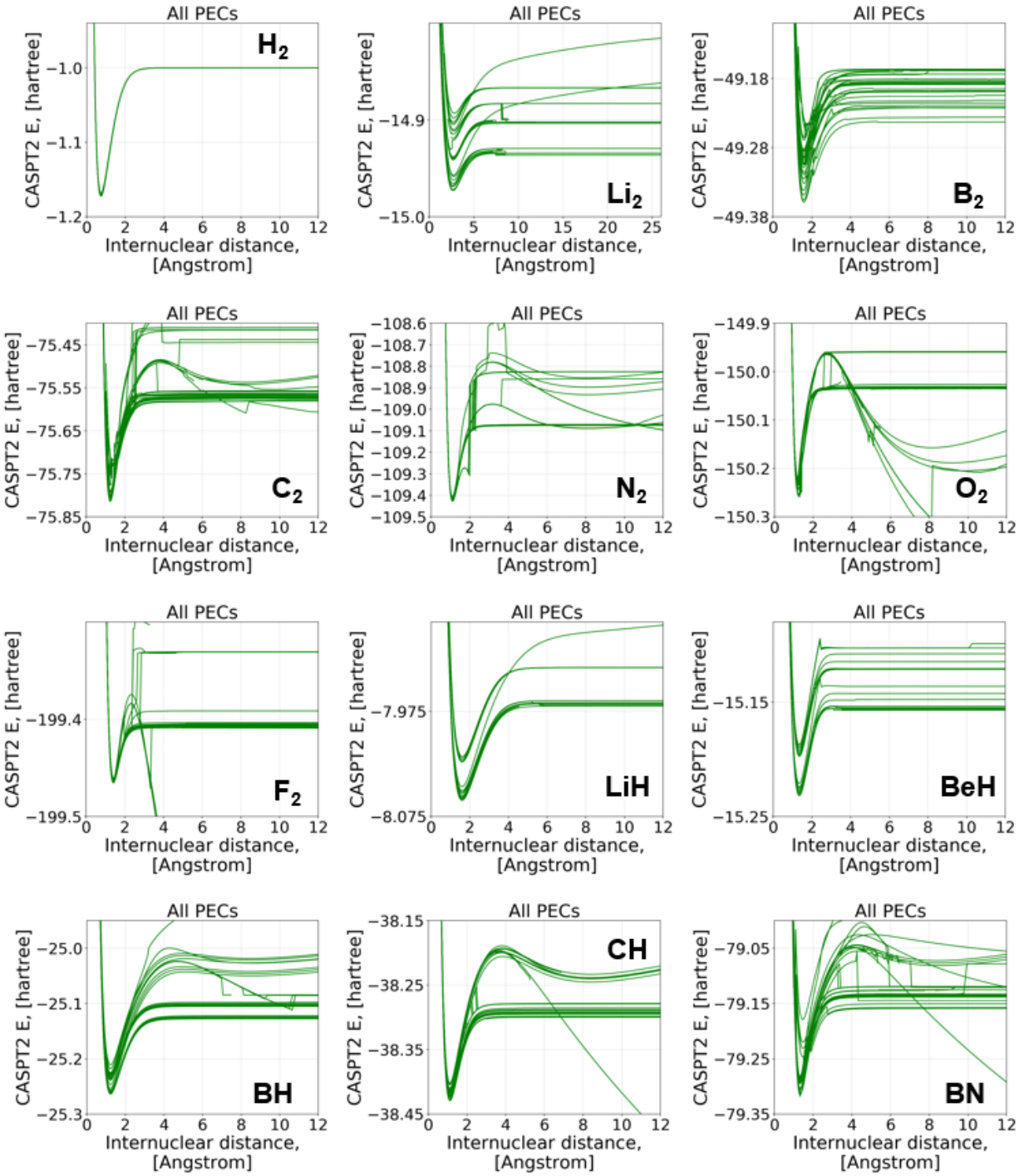

Figure S1-1. All potential energy curves using CASSCF/CASPT2 energies for homonuclear diatomic molecules, hydrides, and BN.

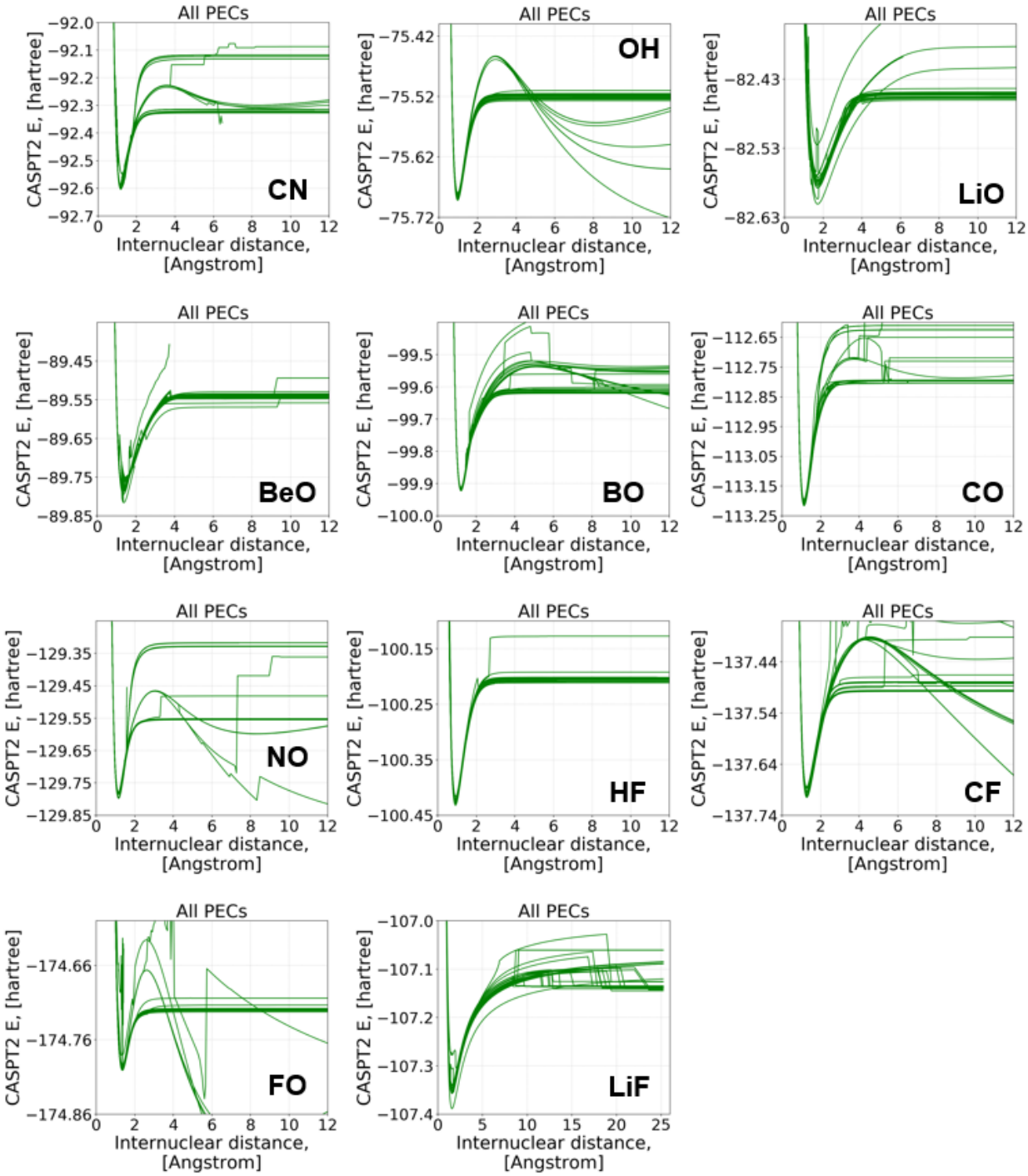

Figure S1-2. All potential energy curves using CASSCF/CASPT2 energies for oxides, fluorides, and CN.

S2. Featurization of Raw Data

Predictive variables (i.e., features) include the numbers of active electrons and orbitals, the internuclear distance (in Ångstroms), occupation numbers, and molecular orbital (MO) coefficients. Only MO coefficients related to 1s, 2s, 2p, 3s, and 3p atomic orbitals are extracted from CASSCF calculation results in order to exclude insignificant information and reduce the computational cost of training the ML models. MO coefficients are set to zero for MOs where the occupation number is zero in order to ignore the virtual orbitals and insignificant orbitals regarding orbital occupancy.

S3. Automated Labeling Procedure

To select a reference potential energy curve (PEC) data for each system among simulated PECs obtained through CASSCF/CASPT2 calculations, the Hulburt-Hirschfelder (HH) potential function was adopted (equations below).[15,16] Among the various complex potential functions for diatomic molecules, the Hulburt-Hirschfelder potential is helpful because it does not require additional high-level of calculations, only experimental data such as bond dissociation energy, equilibrium bond length, vibrational constants that are available for diatomic systems of our work.

$$V_{HH} = D_e \left[\left\{ 1 - e^{-\beta(r-r_e)} \right\}^2 + \{1 + b\beta(r-r_e)\} c\beta^3 (r-r_e)^3 e^{-2\beta(r-r_e)} \right]$$

$$\beta = \frac{\omega_e}{2r_e (B_e D_e)^{\frac{1}{2}}}$$

$$a_0 = \frac{\omega_e^2}{4B_e}$$

$$a_1 = -1 - \frac{\alpha_e \omega_e}{6B_e^2}$$

$$a_2 = \frac{5}{4}a_1^2 - \frac{2}{3}\frac{\omega_e x_e}{B_e}$$

$$c = 1 + a_1\left(\frac{D_e}{a_0}\right)^{\frac{1}{2}}$$

$$b = 2 - \frac{\left(\frac{7}{12} - \frac{D_e a_2}{a_0}\right)}{c}$$

$$K = \beta(r - r_e)$$

where D_e is the energy of dissociation, r is internuclear distance, r_e is the equilibrium bond length, ω_e is the harmonic vibrational constant, $\omega_e x_e$ is the first anharmonicity constant (note that the symbol $\omega_e x_e$ is a single constant, not a product), α_e is the first term rotational constant (also known as the vibration-rotation coupling constant), B_e is the rotational constant in equilibrium position, a_n is the Dunham's coefficients, and b, c are the Hulburt-Hirschfelder constants.

To compare PECs, PECs needs to be shifted along y axis (i.e., energy) since the multiconfigurational calculations with different active spaces could result in different absolute energies even though the overall shape of the PECs are similar (See Figure S2, an example of BeH). By comparing with the HH potential or reference PECs, simulated PECs are shifted to minimize a sum of median absolute errors between energies of two PECs at each internuclear distance.

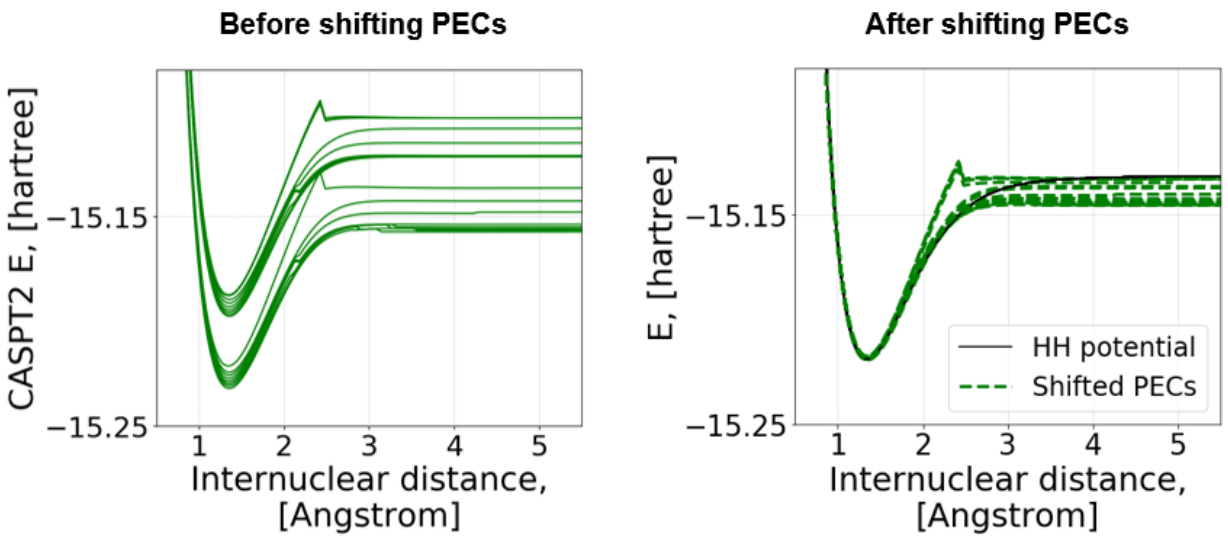

Figure S2. Comparison of original potential energy curves and shifted curves for BeH

For selecting a reference PEC that is the most similar PEC to the corresponding Hulburt-Hirschfelder (HH) PEC, deviation area was calculated. To do this, curves of the HH PEC and one of CASPT2 PECs were redefined as two different curves: upper and lower bound curves. After that, each curve was fitted separately based on the B-spline method using interpolate function in the open-source Python library SciPy.[17] In the range from 0.65*r_e to 5.0*r_e, the area bound by the fitted upper and lower curves was computed numerically. As shown in Figure S3, the selected reference PECs are well matched with corresponding HH PECs except for BeO and LiF (Figure S4). In the case of BeO, calculated bond dissociation energies via CASSCF/CASPT2 are much larger than the experimental value. For LiF, most cases with different active space resulted in a large discontinuity at large separation (i.e., larger than 10 Angstrom). The errors in BeO are likely due to dissociation to the wrong state. Our calculations are spin-constrained, meaning that the singlet spin of the BeO molecule is preserved throughout the entire dissociation. For most diatomic systems this does not pose a problem, but BeO dissociates to a singlet Be atom and a triplet O atom, which would be a triplet overall.[18] The errors in our

calculated dissociation energies with respect to experiment can largely be explained by the energy difference between the ground-state ^3P O atom and the excited-state ^1D. For LiF, most PECs dissociate to Li^+ and F^-, and at large distances abruptly transition to neutral Li and F, which introduces large discontinuities in the PEC. For both systems, most data points in all PECs are describing states other than the states of interest, and so BeO and LiF were both excluded from the ML protocol development.

Table S1. Comparison of Experimental and Simulation Data of Bond Dissociative Properties for Reference PECs with the Best Active Space Selection

System	Active space	D_e [kcal/mol]			r_e [Å]			ω_e [cm^{-1}]		
		cal.	exp.	error	cal.	exp.	error	cal.	exp.	error
H_2	(2, 4)	107.4	109.5	-2.1	0.758	0.741	0.017	4389	4401	-12
Li_2	(4,10)	23.8	24.2	-0.4	2.688	2.673	0.015	348	351	-3
B_2	(6,10)	66.5	69.9	-3.5	1.608	1.590	0.018	1076	1060	16
C_2	(8, 7)	151.2	149.5	1.7	1.249	1.243	0.006	1852	1855	-3
N_2	(10,10)	261.2	228.3	32.9	1.104	1.098	0.006	2328	2359	-31
O_2	(8, 7)	122.3	120.5	1.8	1.213	1.208	0.005	1571	1580	-9
F_2	(6, 6)	36.7	38.3	-1.7	1.423	1.412	0.011	901	917	-16
LiH	(4,10)	56.4	58.0	-1.6	1.605	1.595	0.010	1397	1405	-8
BeH	(5, 4)	51.4	54.9	-3.5	1.352	1.343	0.009	2040	2061	-21
BH	(6, 6)	84.6	85.0	-0.4	1.230	1.232	-0.002	2398	2367	31
CH	(7, 6)	81.5	84.0	-2.5	1.120	1.120	0.000	2834	2861	-27
OH	(7, 6)	106.7	107.2	-0.4	0.975	0.970	0.005	3720	3738	-18
HF	(8, 7)	140.4	141.1	-0.8	0.925	0.917	0.008	4081	4138	-57
BN	(6,10)	94.8	91.6	3.2	1.330	1.325	0.005	1515	1515	0
CN	(7, 6)	179.3	181.3	-2.0	1.173	1.172	0.001	2063	2069	-6
LiO	(9, 8)	80.3	81.7	-1.4	1.716	1.688	0.028	793	815	-22
BeO	(10,7)	172.3	105.6	66.6	1.334	1.331	0.003	19527	1457	18070
BO	(7, 8)	193.1	195.2	-2.1	1.212	1.205	0.007	1881	1885	-4
CO	(10, 9)	257.5	259.5	-2.0	1.134	1.128	0.006	2147	2170	-23
NO	(9, 7)	144.7	152.8	-8.1	1.159	1.151	0.008	1870	1904	-34
FO	(9, 6)	51.2	53.2	-2.0	1.360	1.354	0.006	1043	1053	-10
LiF	(10,6)	136.4	138.3	-2.0	1.765	1.564	0.201	4615	911	3704
CF	(5,10)	128.1	123.8	4.3	1.279	1.272	0.007	1435	1308	127

*D_e: bond dissociation energy, r_e: equilibrium bond length, ω_e: vibrational constant

All of the errors for bond dissociation energy (D_e) are larger than the chemical accuracy of 1 kcal/mol, indicating that chemical accuracy cannot be used to identify which active space selections would be good enough among available data. Errors larger than chemical accuracy for bond dissociation energy of diatomic molecules are not rare even using multiconfiguration calculations with a larger basis set than the one we used.[19] In particular, N_2 showed the largest error and it is known that a very large basis set is needed to obtain accurate bond dissociation energy for this triple-bonded system.[19] Similarly, the errors for vibrational frequency show a large variation (~30 cm^{-1}) that is beyond the spectroscopic accuracy (i.e, ±1 cm^{-1}).[20] However, many of the errors for equilibrium bond length are smaller than 0.01 Å.

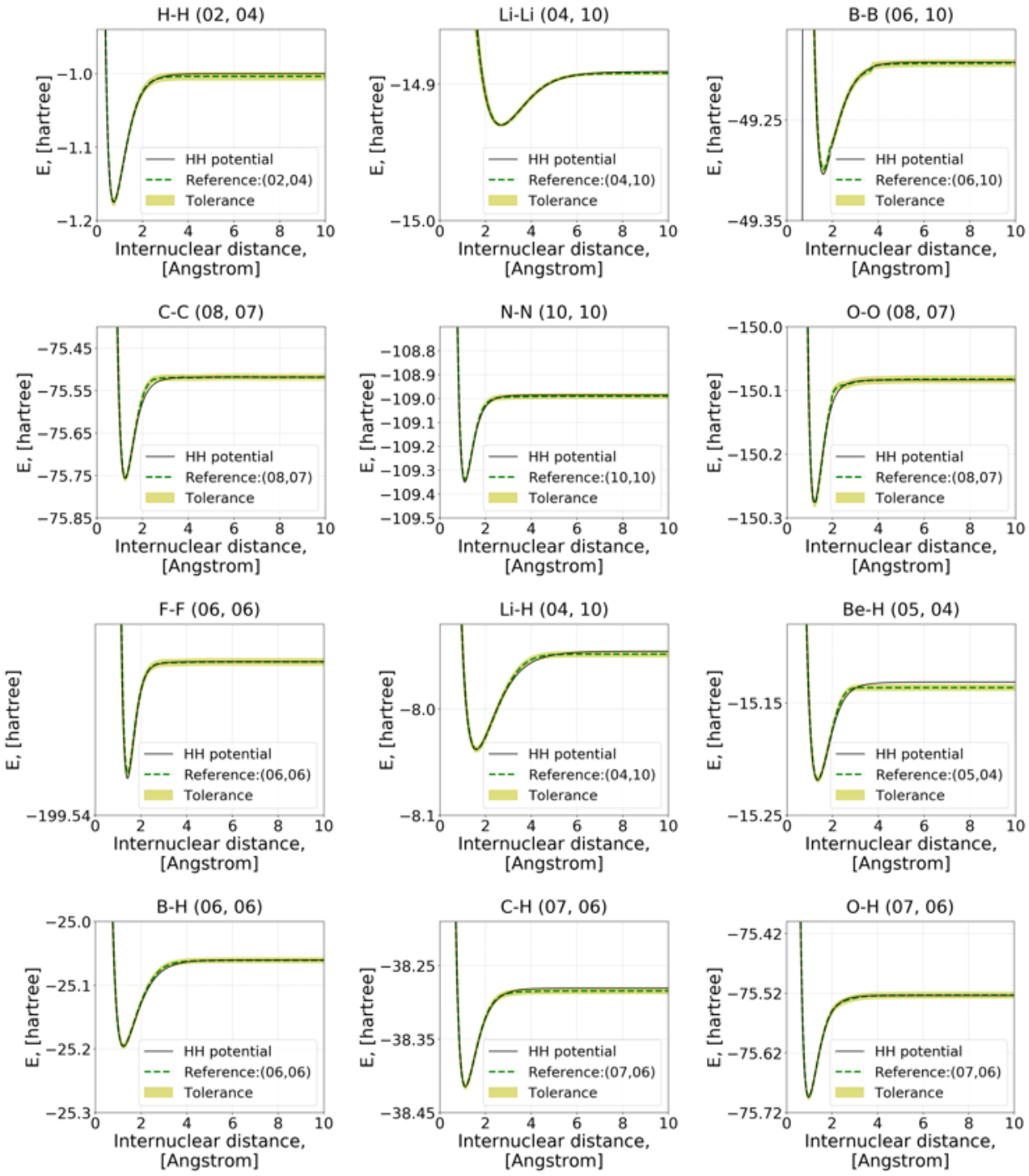

Figure S3-1. Reference potential energy curves for diatomic molecules that are the most similar to the corresponding Hulburt-Hirschfelder potential energy curve

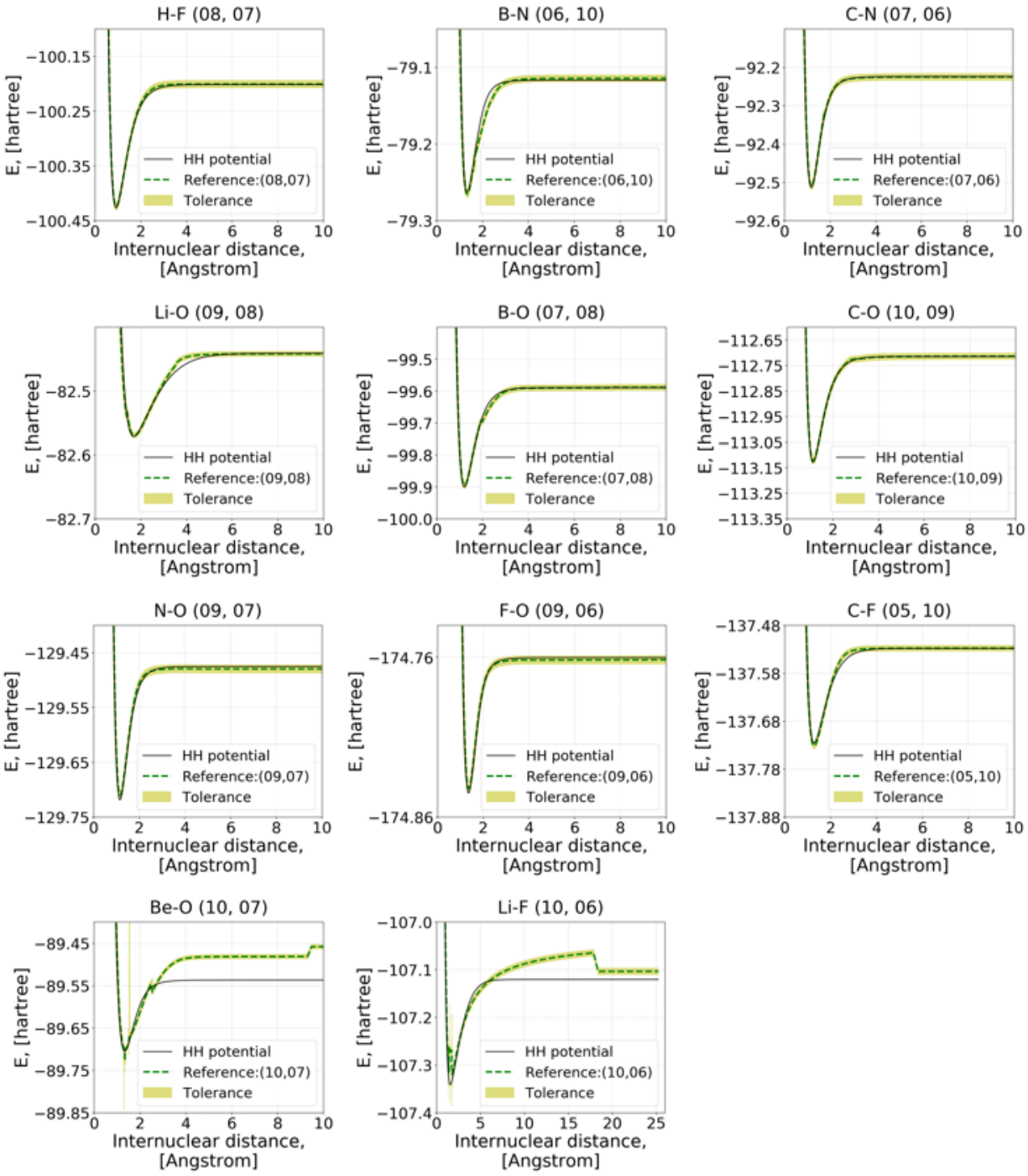

Figure S3-2. Reference potential energy curves for diatomic molecules that are the most similar to the corresponding Hulburt-Hirschfelder potential energy curve

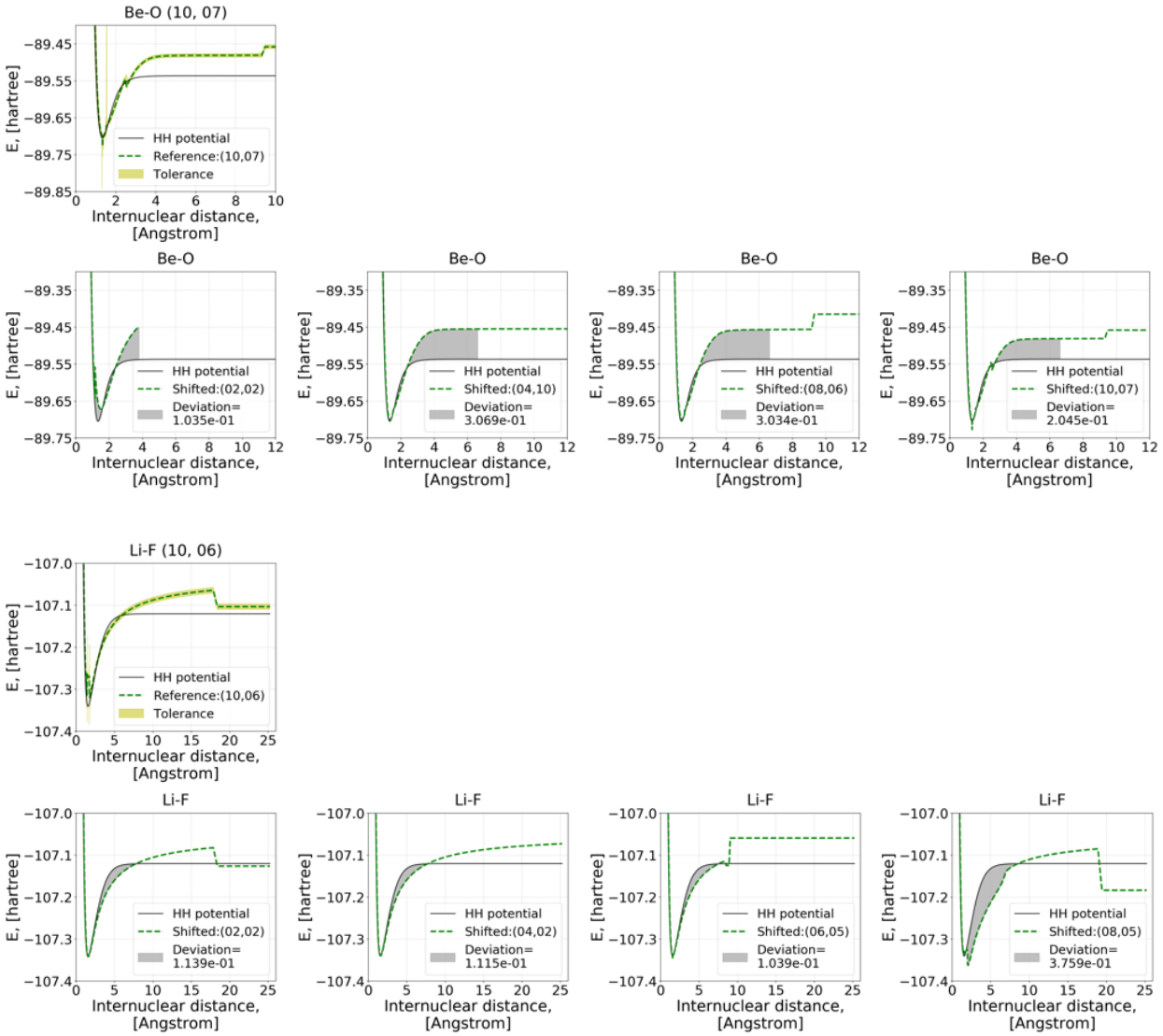

Figure S4. Representative potential energy curves for BeO and LiF. There is no similar potential energy curve compared to the corresponding Hulburt-Hirschfelder potential energy curve.

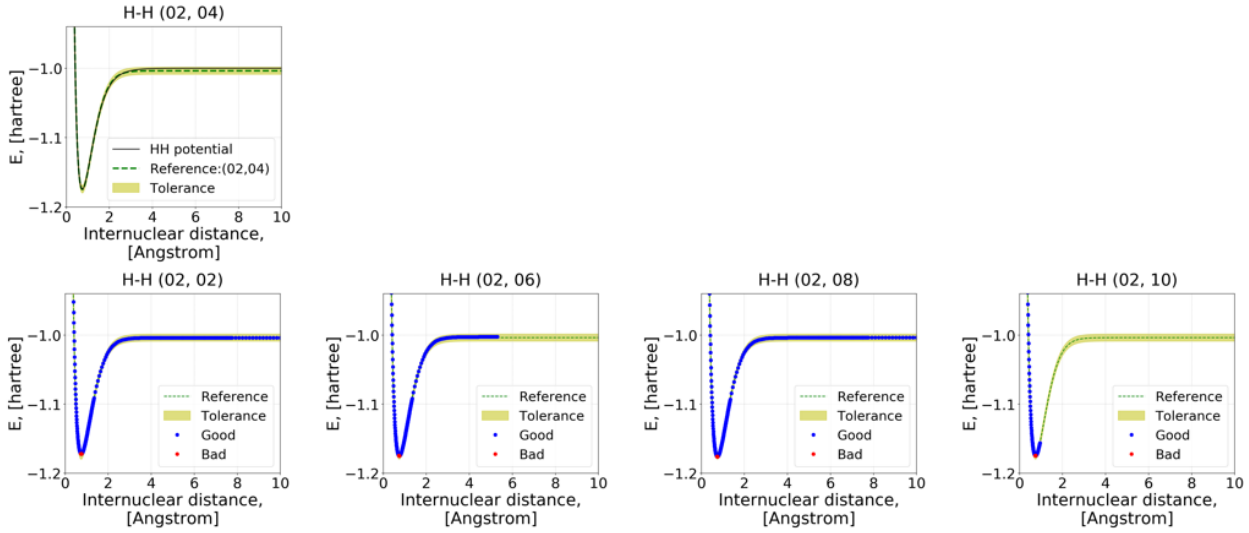

Figure S5. Representative potential energy curves for H_2 that show imbalanced data points. A portion of the bad-labeled data points are extremely small, and the data points only exist near the equilibrium bond length.

Assigning a good or bad label to each data point is based on comparison between a test PEC of interest and the corresponding reference PEC with respect to energy and its derivative. We have set two different criteria for the labeling: First, to assign a good label to a data point, energy of the data point of a target system should be within the energy tolerance of 3% on dimensionless PEC space that is generated by dividing the internuclear distance (i.e., x-axis) and the energy (i.e., y-axis) by the corresponding equilibrium bond length and bond dissociation energy, respectively. We used the dimensionless PEC space because the original x/y axes have different units, so the 3% energy tolerance could not capture similar trend of the PEC shapes at very short internuclear distance where slopes of the PECs are very large. To compute upper and lower E bounds of the given reference PEC at arbitrary distances for the comparison, two equidistant curves (i.e., parallel curves) with respect to the reference PEC were obtained via a fitted energy and its derivate of the reference PEC using the B-spline method (Python library

SciPy). Second, derivatives of energy were compared between the reference and test PECs. The derivatives were calculated from fitted PEC lines obtained via the B-spline method (Python library SciPy) on the dimensionless PEC space produced by dividing x axis and (y axis-y_min) by the corresponding equilibrium bond length and bond dissociation energy, respectively. The difference between E derivates for each PEC with a smaller derivate value was used to determine whether a given data point is labeled as good or bad as below. The smaller derivate was considered because E derivative tolerance needs to be larger when both slopes are large to capture the overall variation trend of PEC. Too small derivative is ignored and changed to 0.05.

$$labeled\ as\ good\ if\ \frac{\text{abs(derivative of the reference PEC} - \text{derivative of the test PEC)}}{\min\ (\text{derivative of the reference PEC},\ \text{derivative of the test PEC})} \leq 1.0$$

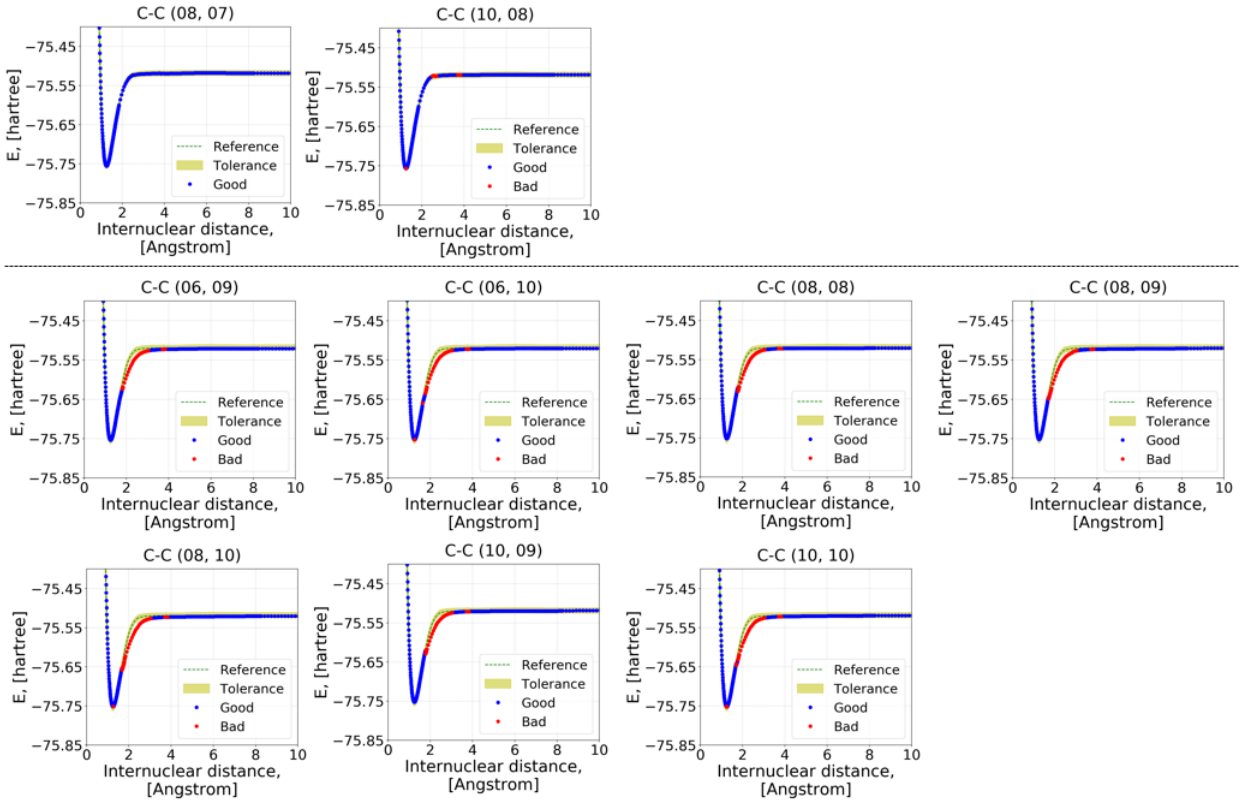

Figure S6. Confusing potential energy curves for C_2. Both good and bad PECs have similar errors in dissociative properties, but only can be distinguished only by different curvatures compared to the reference PEC.

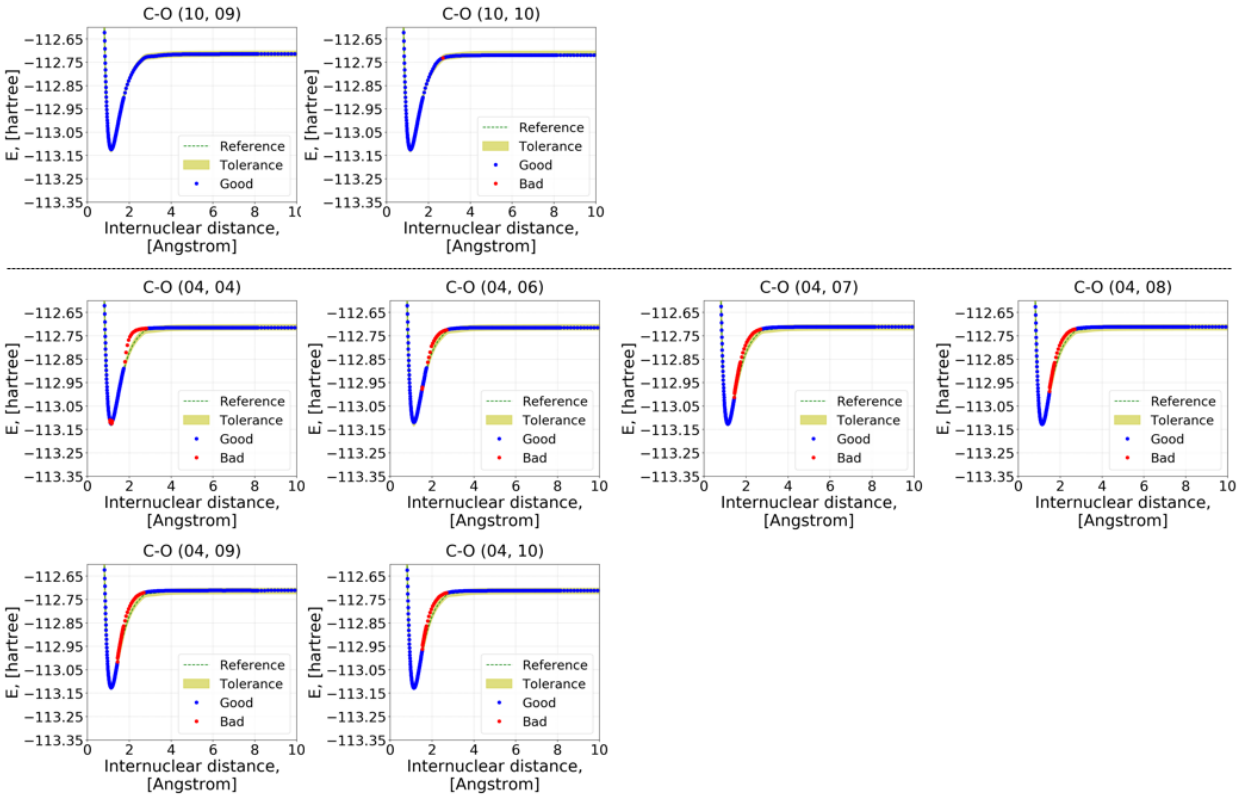

Figure S7. Confusing potential energy curves for CO. Both good and bad PECs have similar errors in dissociative properties, but only can be distinguished only by different curvatures compared to the reference PEC.

Table S2. Errors of dissociative properties for C_2 and CO.

Diatomic molecule	Active space	PEC label	Absolute error			Relative error		
			D_e [eV]	r_e [Å]	ω_e [cm^{-1}]	D_e [%]	r_e [%]	ω_e [%]
C_2	(8, 7)	1	0.08	0.0067	3.0	1.23	0.54	0.16
	(10, 8)	1	0.01	0.0037	0.3	0.15	0.30	0.02
	(6, 9)	0	0.15	0.0070	9.5	2.31	0.56	0.51
	(6,10)	0	0.17	0.0073	25.4	2.62	0.59	1.37
	(8, 8)	0	0.12	0.0084	16.0	1.85	0.68	0.86
	(8, 9)	0	0.16	0.0084	10.5	2.47	0.68	0.57
	(8,10)	0	0.19	0.0076	137.0	2.93	0.61	7.39
	(10, 9)	0	0.09	0.0057	10.3	1.39	0.46	0.56
	(10,10)	0	0.13	0.0162	113.7	2.01	1.30	6.13
CO	(10, 9)	1	0.09	0.0061	23.2	0.80	0.54	1.07
	(10,10)	1	0.23	0.0063	22.4	2.04	0.56	1.03
	(4, 4)	0	0.07	0.0047	281.8	0.62	0.42	12.99
	(4, 6)	0	0.25	0.0038	10.7	2.22	0.34	0.49
	(4, 7)	0	0.04	0.0055	7.6	0.36	0.49	0.35
	(4, 8)	0	0.08	0.0057	10.6	0.71	0.51	0.49
	(4, 9)	0	0.05	0.0059	6.3	0.44	0.52	0.29
	(4,10)	0	0.06	0.0058	10.1	0.53	0.51	0.47

*D_e: bond dissociation energy, r_e: equilibrium bond length, ω_e: vibrational constant

Table S3. Number of data points for the diatomic molecules used in this work.

No.	Diatomic molecule	Spin multiplicity	Total number of data points	Number of good labeled points	Number of bad labeled points	% of good data points
1	H_2	1	1746	1712	34	98.05
2	Li_2	1	6797	5093	1704	74.93
3	B_2	3	8754	6262	2492	71.53
4	C_2	1	8897	5192	3705	58.36
5	N_2	1	7674	5472	2202	71.31
6	O_2	3	8427	4400	4027	52.21
7	F_2	1	8718	7006	1712	80.36
8	LiH	1	4581	3936	645	85.92
9	BeH	2	5293	3788	1505	71.57
10	BH	1	6992	5195	1797	74.30
11	CH	2	6985	5459	1526	78.15
12	BN	3	8552	4803	3749	56.16
13	CN	2	7691	4881	2810	63.46
14	OH	2	7810	6004	1806	76.88
15	LiO	2	8800	7175	1625	81.53

16	BeO	1	8541	4357	4184	51.01
17	BO	2	7853	5450	2403	69.40
18	CO	1	9370	5568	3802	59.42
19	NO	2	7076	4583	2493	64.77
20	HF	1	7361	6873	488	93.37
21	CF	2	7734	6096	1638	78.82
22	FO	2	6720	5211	1509	77.54
23	LiF	1	10095	7892	2203	78.18

S4. Development of XGBoost (eXtreme Gradient Boosting) Models

The open source gradient boosting decision tree Python library XGboost[21] was used to build and train the classification ML models for this work. XGBoost is known to be powerful for practical ML problems in the *Kaggle competitions*,[22,23] and it is appropriate for training a large number of data points since it supports parallelization of training procedure. It is also easier to optimize hyperparameters in XGBoost than in artificial neural networks, which enables automation of the hyperparameter optimization procedure. Hyperparameter tuning was performed using Hyperopt,[24] a Bayesian optimization tool in Python with 10-fold cross-validation. The explored hyperparameter space was set as listed in Table S4, and 20 cycles were conducted for the hyperparameter optimization.

For both of the training and evaluations of ML models, accuracy is adopted as a metric, meaning that the same number of good and bad data points were sampled with the maximum available number of data points for each system randomly for each run. In general, for an imbalanced data set (i.e., different number of data points for each class), the area under the curve (AUC) of the receiver operating curve (ROC) is used as the evaluation metric. However, we did not use the AUC because it measures binary classifier performance across all possible decision thresholds,[25] not for a specific threshold such as 50% in this work. In addition, accuracy is easier to interpret than the AUC. All of ML prediction results in Figures 3 and 4 were obtained by

averaging results from 10 different ML models with different random seeds that changed the shuffling/sampling of training/test data and hyperparameters of the ML models.

Table S4. Hyperparameter search space.

No.	Hyperparameter	Search space
1	Number of trees (n_estimator)	From 100 to 1000 in intervals of 10
2	Boosting learning rate (learning_rate)	1e-4, 1e-3, 1e-2, 1e-1, 1e0, 2e-4, 2e-3, 2e-2, 2e-1, 2e0, 3e-4, 3e-3, 3e-2, 3e-1, 3e0, 5e-4, 5e-3, 5e-2, 5e-1, 5e0
3	Minimum sum of instance weight needed in a child (min_child_weight)	0.1, 0.5, 1, 2, 3, 4, 5, 6. 7, 8, 9, 10
4	Maximum tree depth (max_depth)	From 5 to 50 in intervals of 1

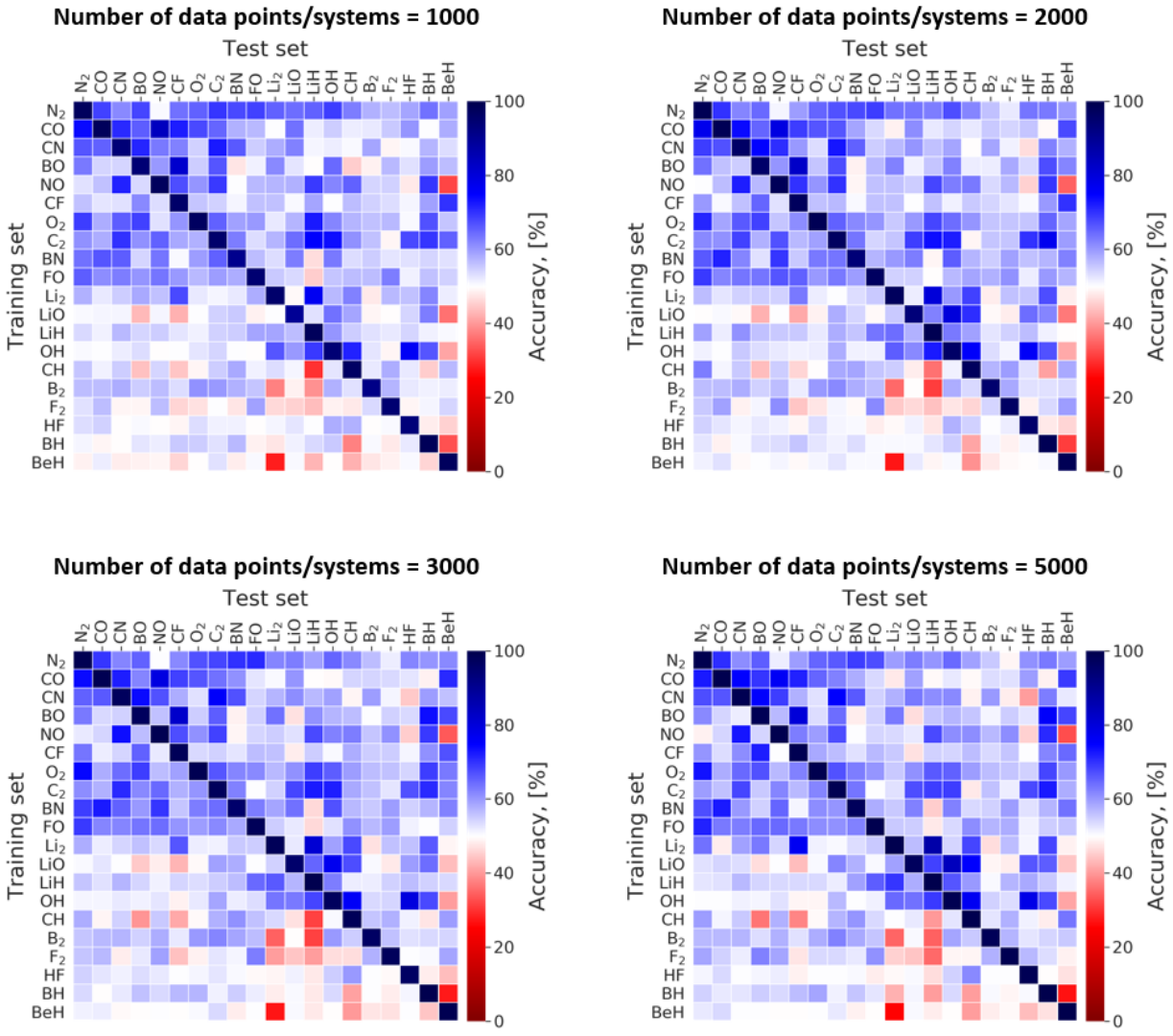

Figure S8. Comparison of ML model prediction performances when the models are trained on the same numbers of training data points (i.e., 1000, 2000, 3000, 5000) per a diatomic molecule and then predicted on all the diatomic systems we investigated.

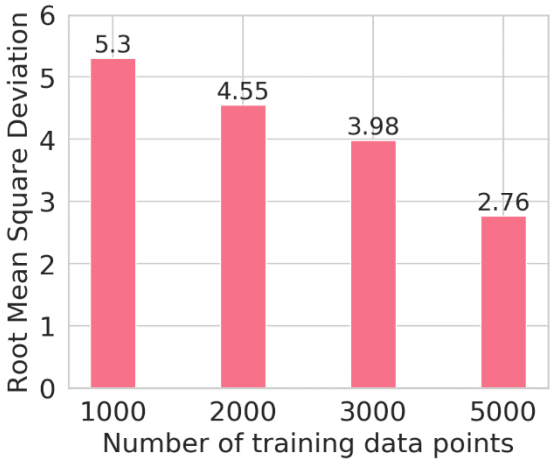

Figure S9. Average root-mean-square deviation between heat maps generated using different number of training data points and the heat map produced with all available training data points.

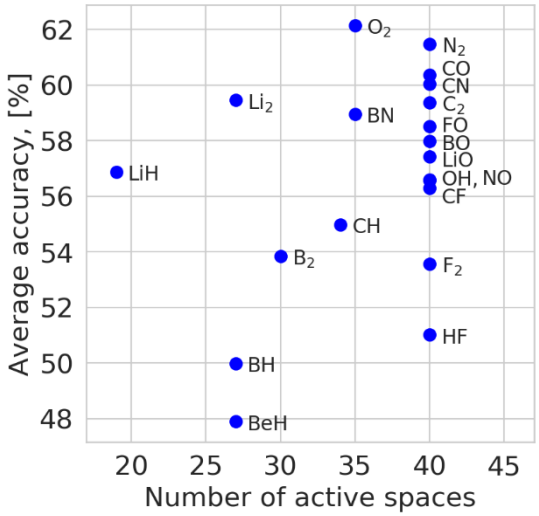

Figure S10. Average prediction accuracy of ML models trained on single diatomic system over other 19 diatomic systems versus the number of possible active spaces limited to the maximum size of 10.

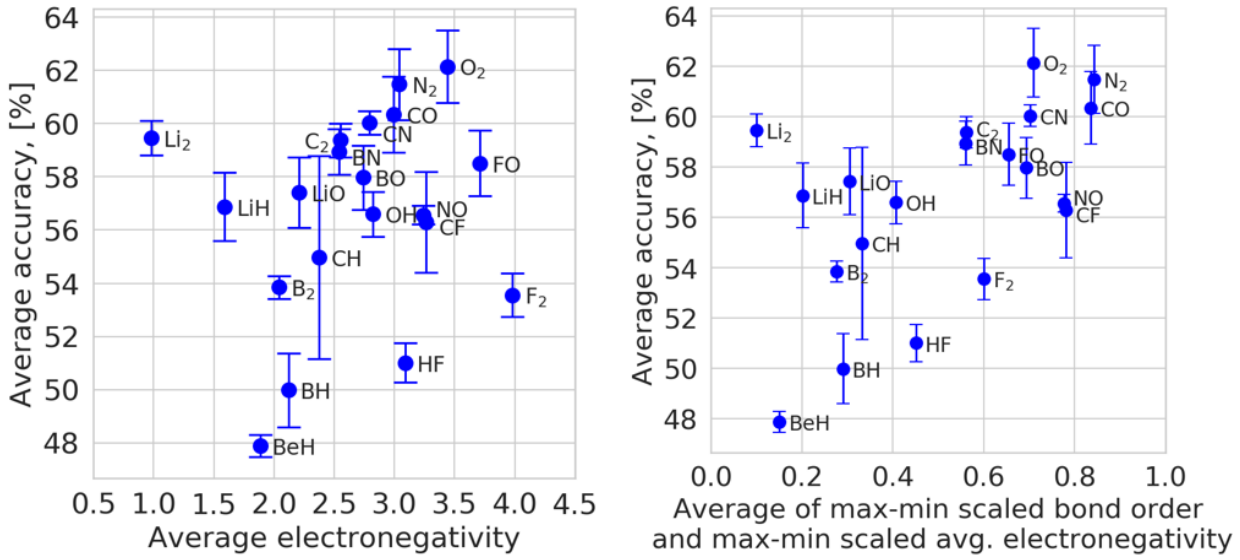

Figure S11. Average prediction accuracy of ML model trained on single diatomic system over other 19 diatomic systems versus (a) average electronegativity and (b) new metric obtained by averaging max-min rescaled bond order and average electronegativity.

Table S5. Top 3 correlated diatomic systems for a target system.

No.	Target system	Best correlated system	2nd best correlated system	3rd best correlated system
1	Li_2	LiH	OH	BO
2	B_2	F_2	CN	NO
3	C_2	CN	N_2	NO
4	N_2	O_2	FO	CO
5	O_2	N_2	CO	C_2
6	F_2	BN	NO	CH
7	LiH	Li_2	O_2	C_2
8	BeH	BO	O_2	CH
9	BH	BO	C_2	NO
10	CH	OH	LiO	Li_2
11	OH	LiO	C_2	LiH
12	HF	OH	CO	C_2
13	BN	N_2	CN	O_2
14	CN	CO	NO	C_2
15	LiO	CN	BN	C_2
16	BO	CN	CO	CF
17	CO	BN	N_2	CN
18	NO	CO	CN	FO
19	FO	N_2	LiH	O_2
20	CF	Li_2	BO	CO

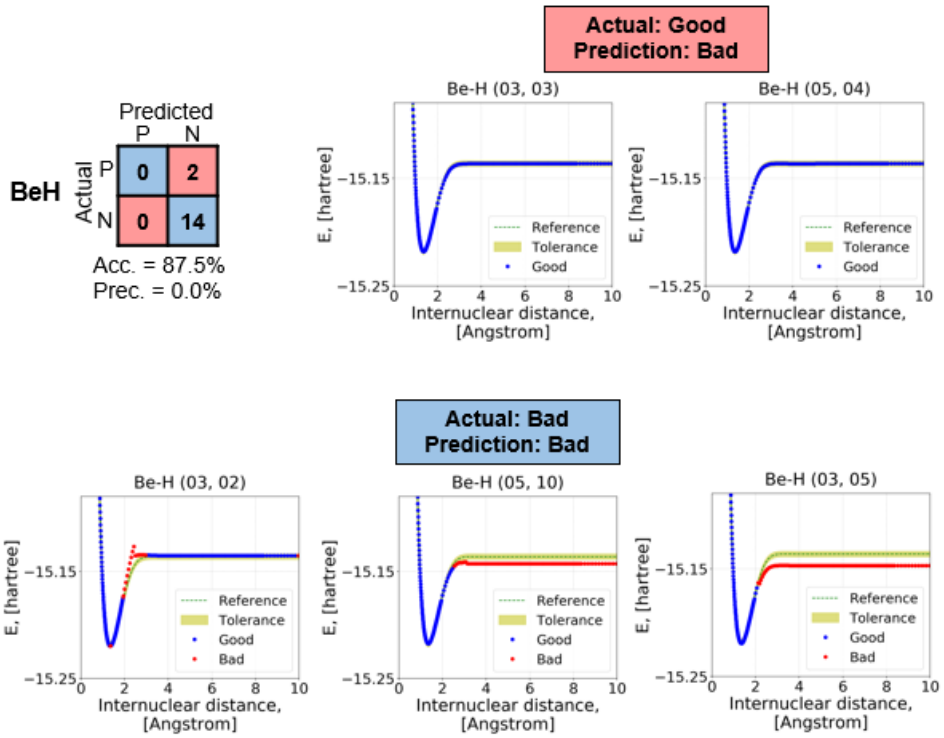

Figure S12. Representative potential energy curves for each case of the confusion matrixes for BeH.

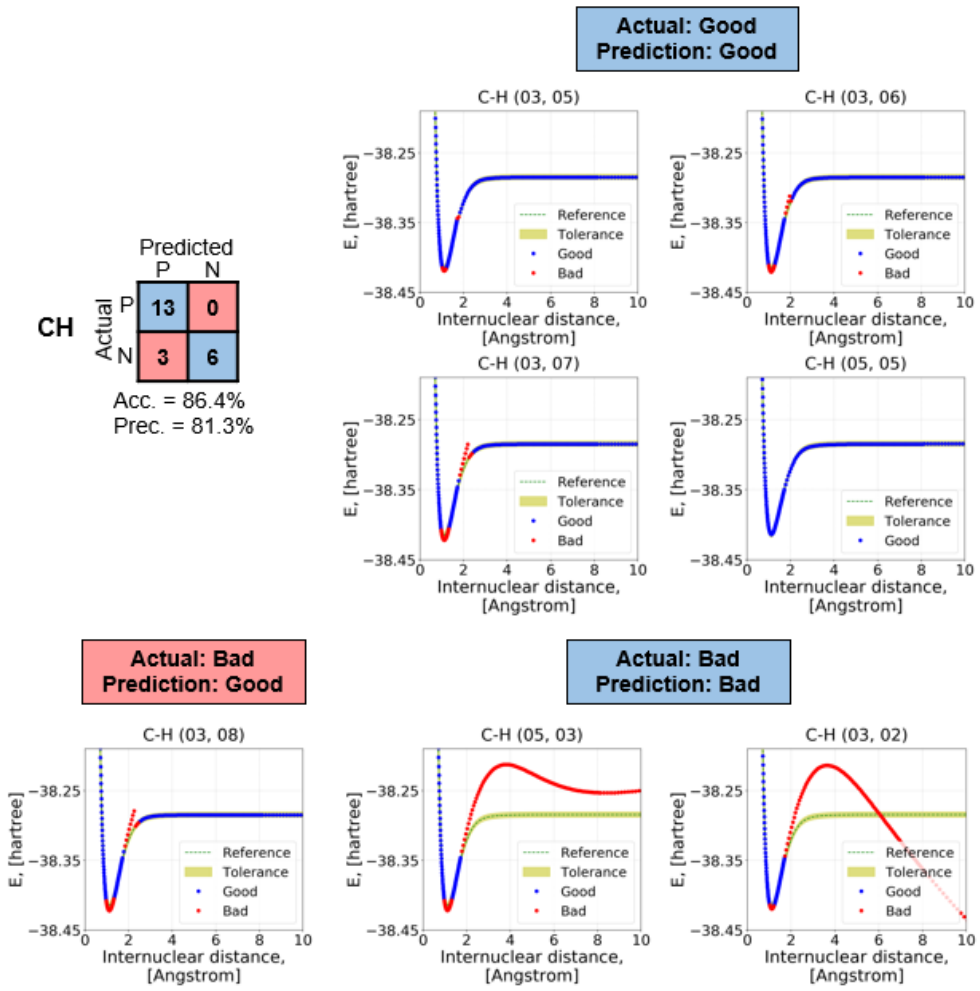

Figure S13. Representative potential energy curves for each case of the confusion matrixes for CH.

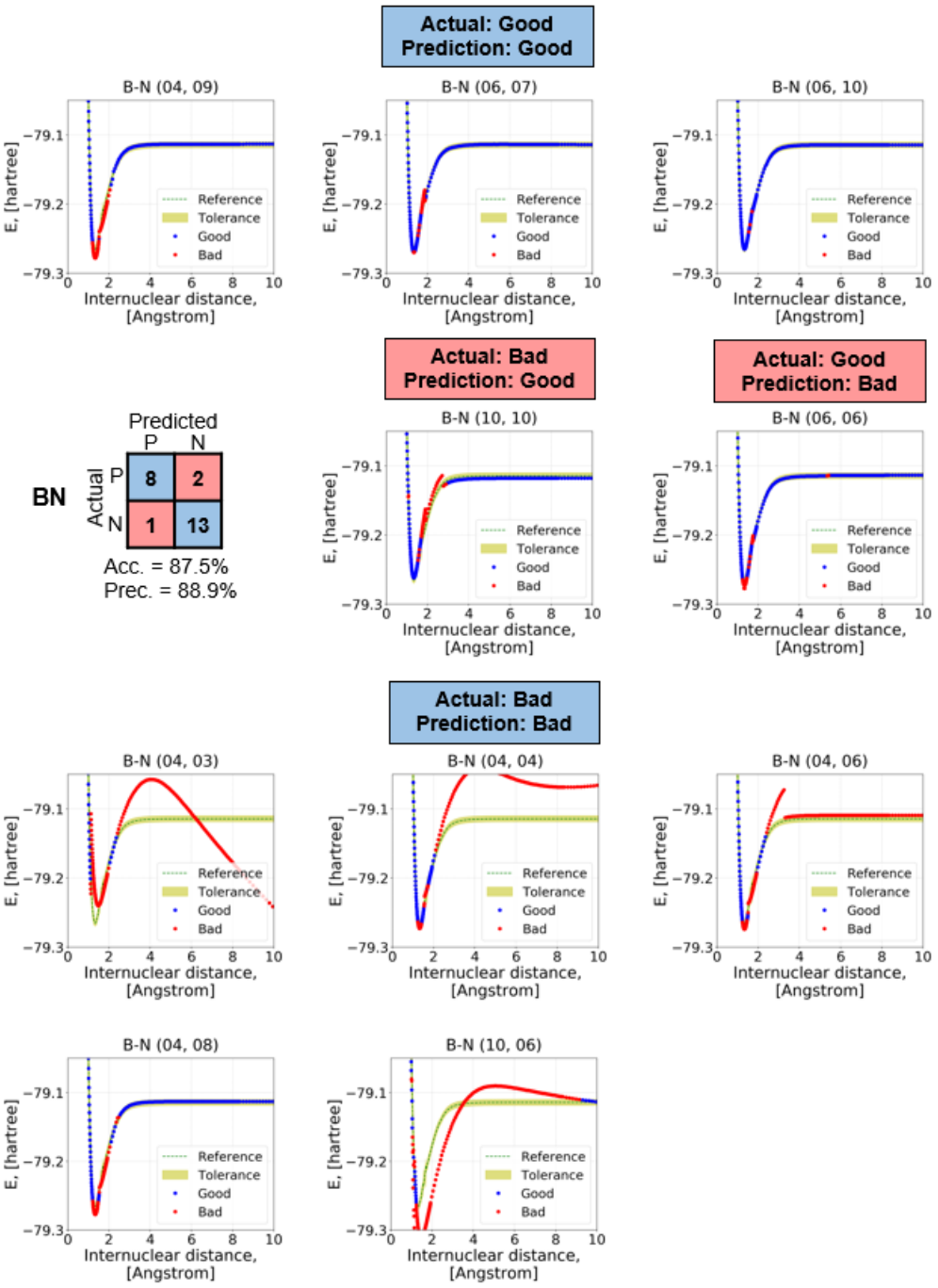

Figure S14. Representative potential energy curves for each case of the confusion matrixes for BN.

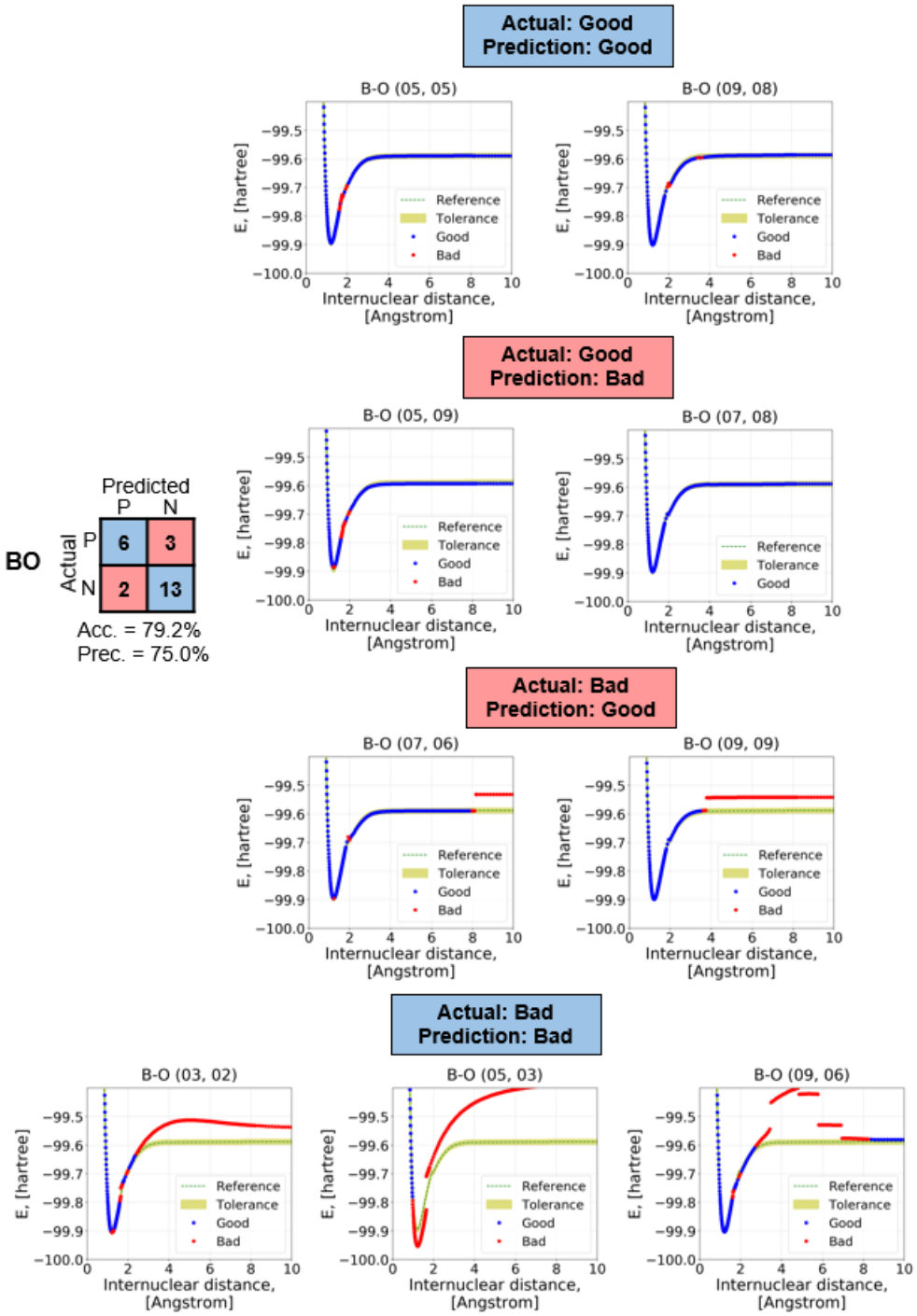

Figure S15. Representative potential energy curves for each case of the confusion matrixes for BO.

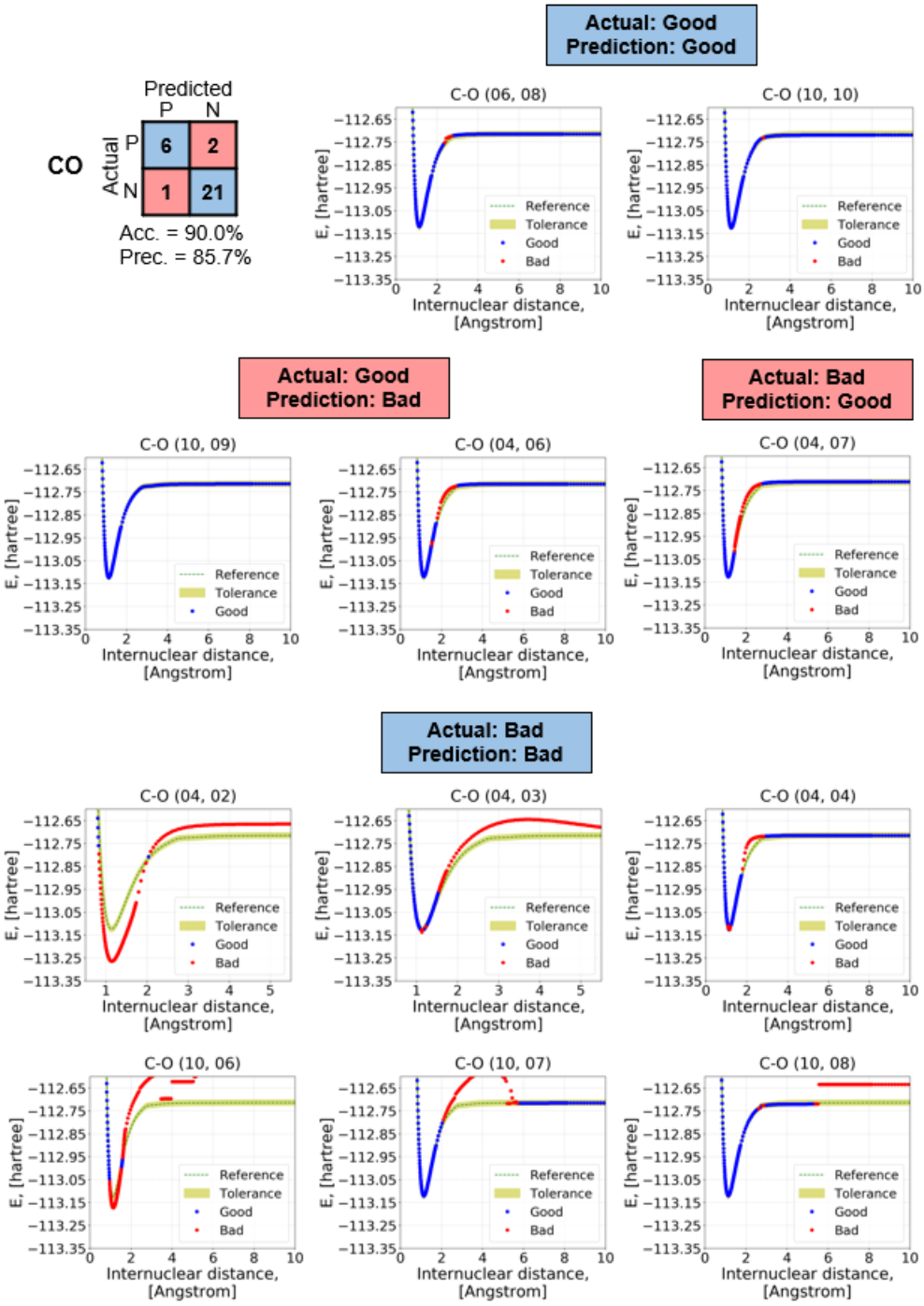

Figure S16. Representative potential energy curves for each case of the confusion matrixes for CO.

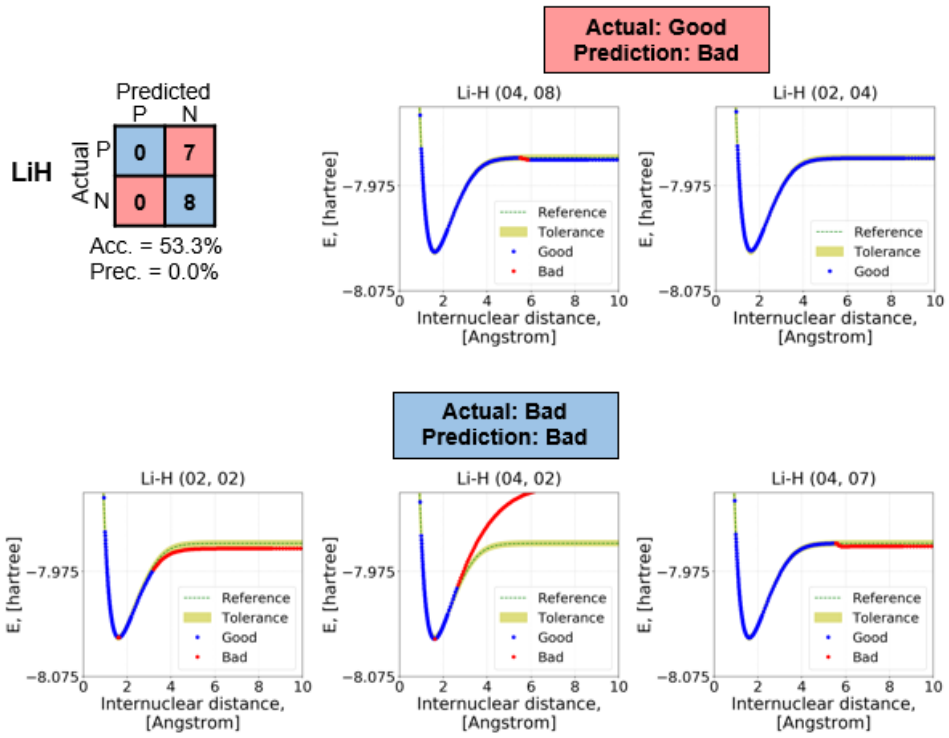

Figure S17. Representative potential energy curves for each case of the confusion matrixes for LiH.

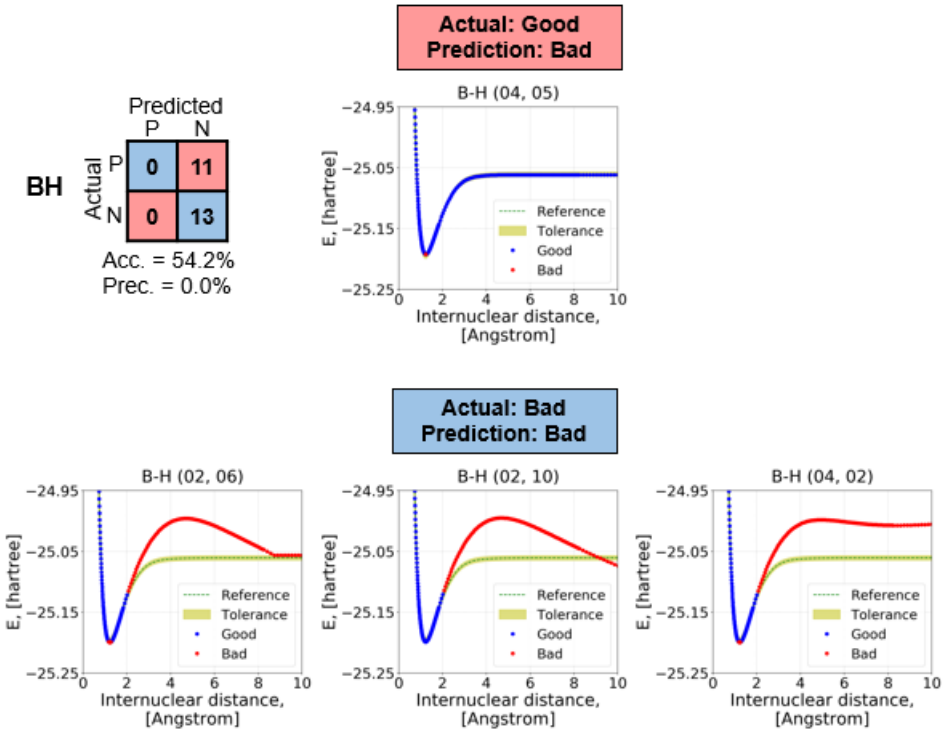

Figure S18. Representative potential energy curves for each case of the confusion matrixes for BH.

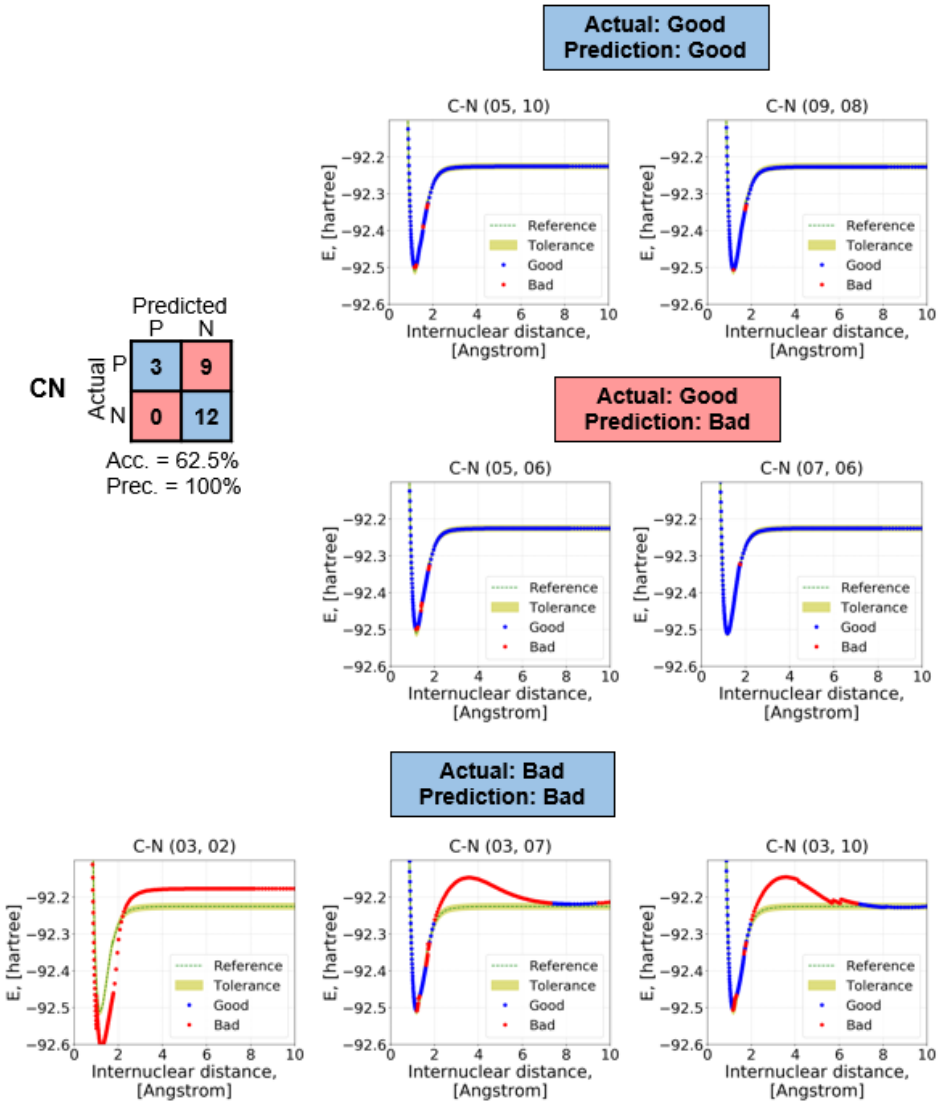

Figure S19. Representative potential energy curves for each case of the confusion matrixes for CN.

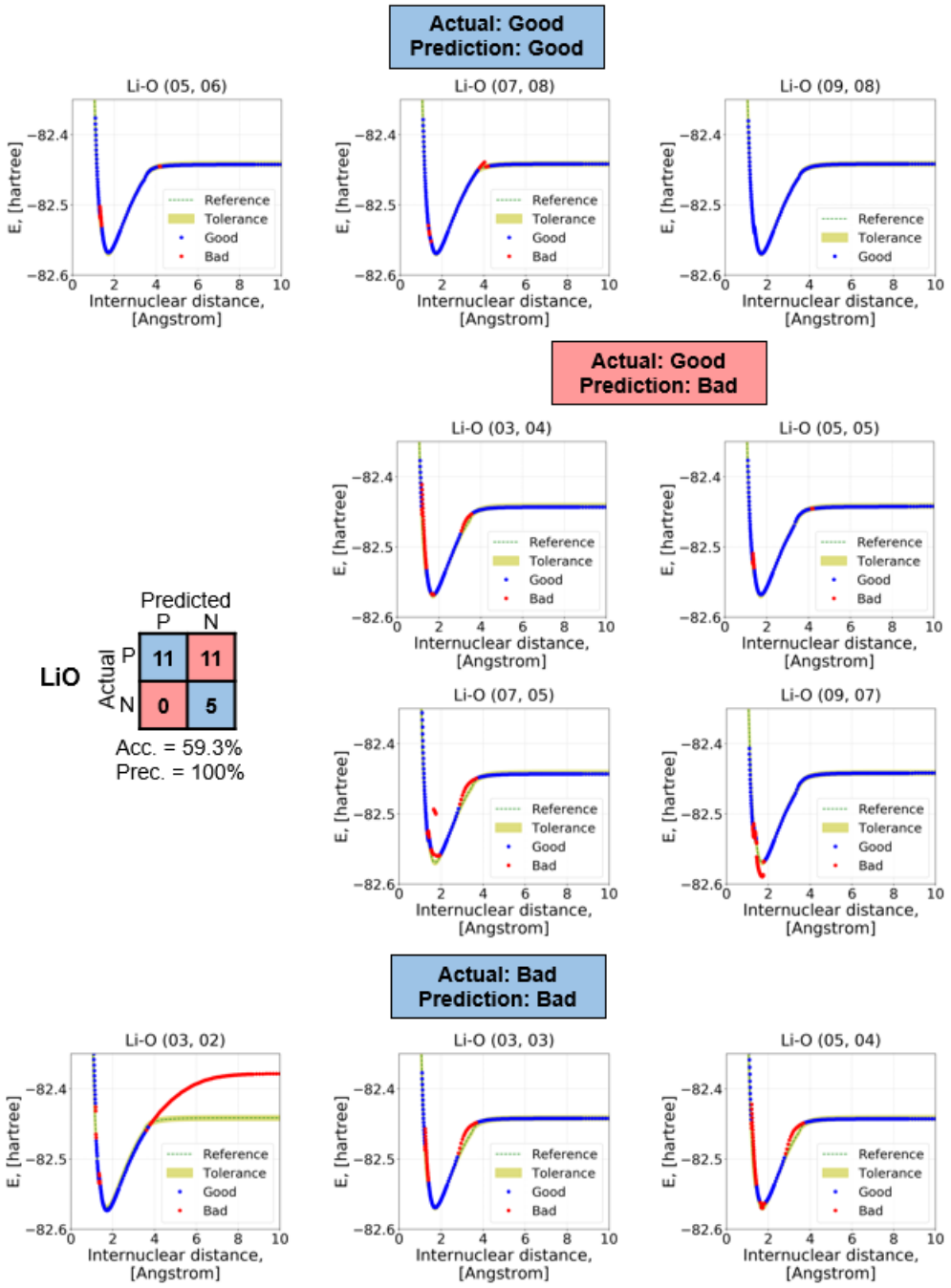

Figure S20. Representative potential energy curves for each case of the confusion matrixes for LiO.

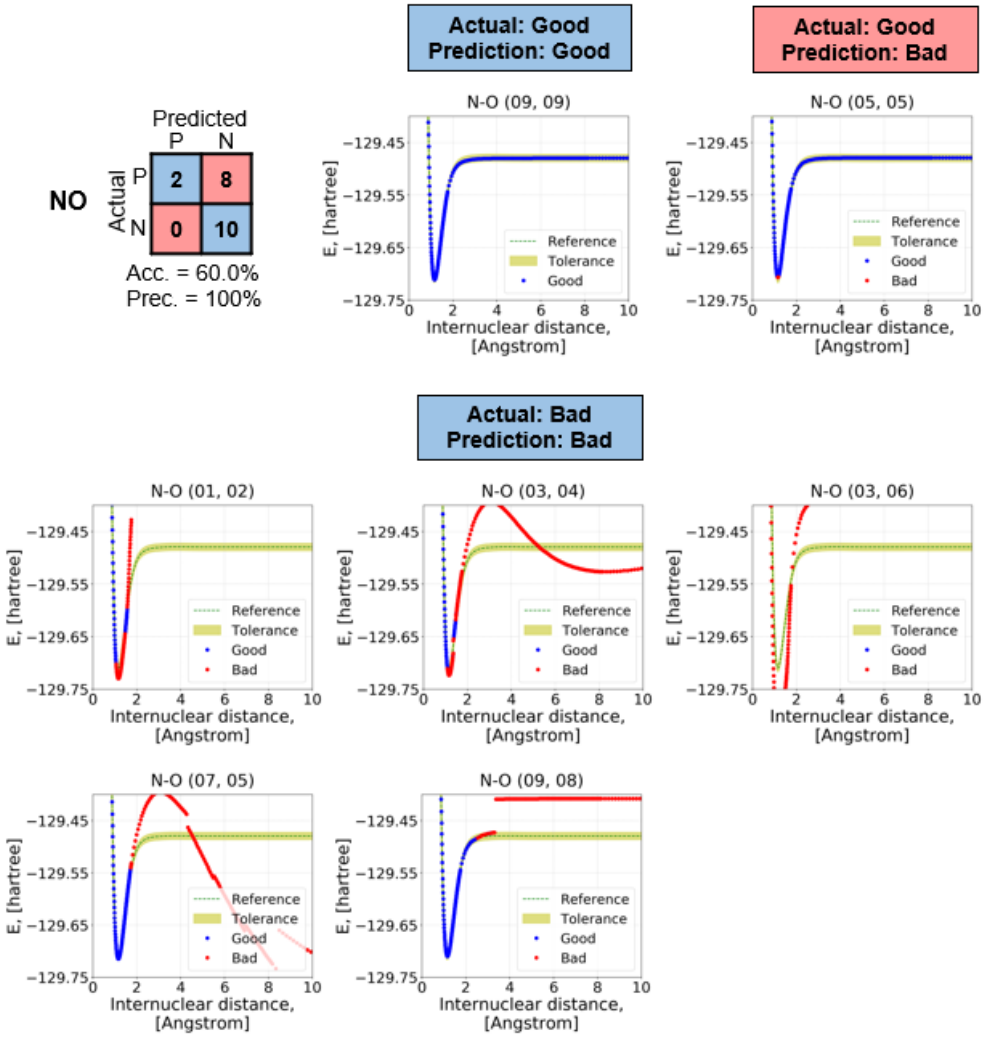

Figure S21. Representative potential energy curves for each case of the confusion matrixes for NO.

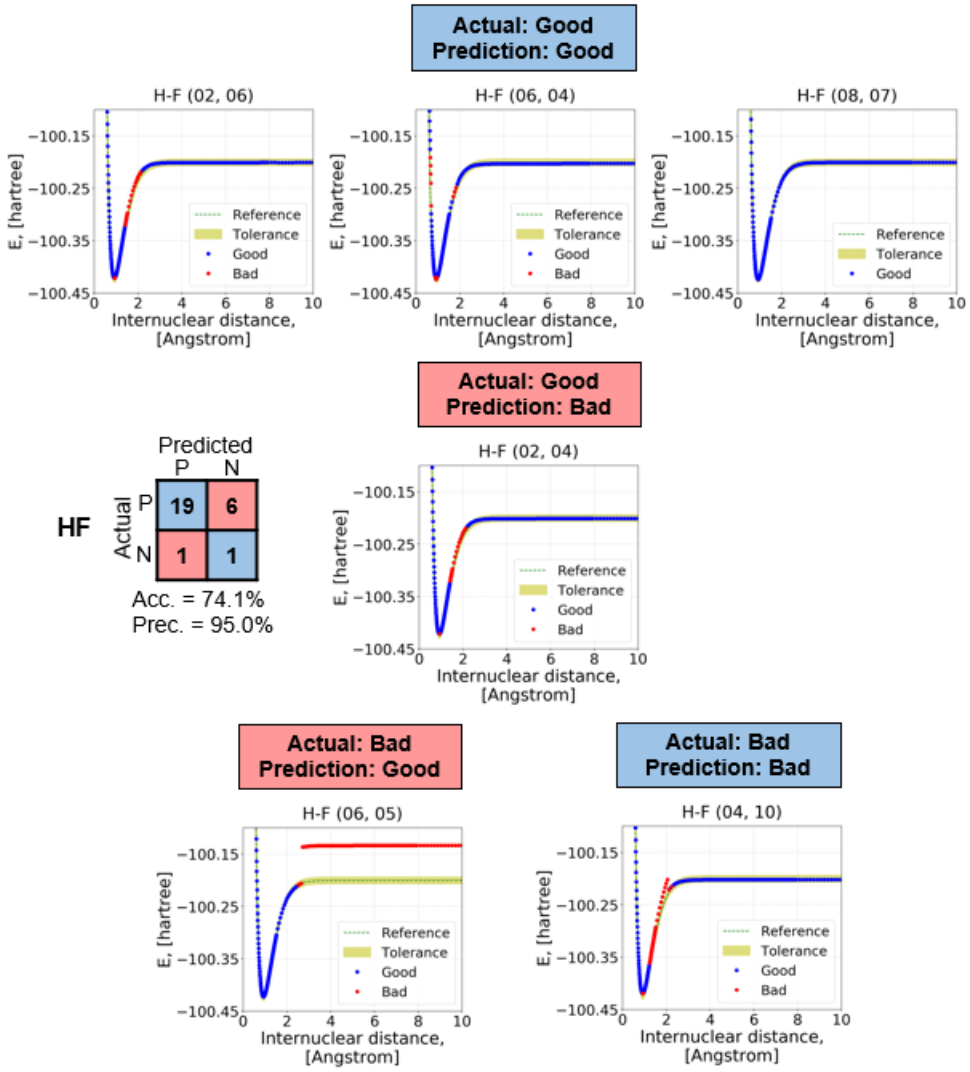

Figure S22. Representative potential energy curves for each case of the confusion matrixes for HF.

Table S6. Comparison of Top 3 good active space selections with the smallest number of configurations between those identified via the automated labeling procedure and those predicted via the ML protocol. The numbers with the underline indicate a bad active space identified via the automated labeling.

Number of good active spaces matched	System	Automated labeling		ML protocol	
		Good active space	Number of configurations	Good active space	Number of configurations
3	**CH**	(3, 5), (3, 6), (5, 5)	40, 70, 75	(3, 5), (3, 6), (5, 5)	40, 70, 75
2	**HF**	(4, 3), (2, 4), (6, 4)	6, 10, 10	(4, 3), (6, 4), (8, 5)	6, 10, 15
	BN	(6, 6), (6, 7), (4, 9)	189, 588, 630	(6, 7), (4, 9), (6, 8)	588, 630, 1512
	BO	(5, 5), (5, 8), (5, 9)	75, 1008, 1890	(5, 5), (5, 8), (9, 8)	75, 1008, 2352
	CO	(4, 6), (6, 8), (6, 9)	105, 1176, 2520	(6, 8), (6, 9), (6,10)	1176, 2520, 4950
	CF	(5, 5), (7, 7), (5, 8)	75, 784, 1008	(7, 7), (5, 8), (5, 9)	784, 1008, 1890
1	**OH**	(3, 5), (7, 5), (9, 6)	40, 40, 70	(3, 5), (5, 5), (3, 7)	40, 75, 112
	FO	(3, 5), (3, 6), (9, 6)	40, 70, 70	(9, 6), (5, 6), <u>(7, 6)</u>	70, 210, <u>210</u>
0	**CN**	(5, 6), (7, 6), (5, 7), (9, 7)	210, 210, 490, 490	(9, 8), (5,10), (9, 9)	2352, 3300, 8820
	LiO	(3, 4), (3, 5), (7, 5)	20, 40, 40	(5, 6), (5, 7), (7, 7)	210, 490, 784
	NO	(5, 5), (5, 6), (7, 6)	75, 210, 210	(9, 9), (9,10)	8820, 27720
N/A	**LiH**	(2, 4), (2, 5), (2, 6)	10, 15, 21	N/A	N/A
	BeH	(3, 3), (5, 4)	8, 20	N/A	N/A
	BH	(4, 5), (4, 6), (6, 6)	50, 105, 175	N/A	N/A

References

(1) Lipscomb, J. D.; Andersson, K. K.; Miinck, E.; Kent, T. A.; Hooper, A. B. Resolution of Multiple Heme Centers of Hydroxylamine Oxidoreductase from Nitrosomonas. 2. Mossbauer Spectroscopy. *Biochemistry* **1982**, *21*, 3973–3976.

(2) Luo, Y. R. In *Comprehensive Handbook of Chemical Bond Energies*, 1st ed.; CRC Press: Boca Raton, FL, 2007.

(3) Computational Chemistry Comparison and Benchmark DataBase (NIST Web Site). https://cccbdb.nist.gov/exp2x.asp?casno=7782447&charge=0 (accessed Jan 27, 2019), Experimental data for O_2 (Oxygen diatomic).

(4) Roos, B. O. The Complete Active Space Self-Consistent Field Method and Its Applications in Electronic Structure Calculations. In *Ab Initio Methods in Quantum Chemistry - II*; Lawley, K. P., Ed.; Wiley: New York, 2007; Vol. 69, pp 399–445.

(5) Andersson, K.; Malmqvist, P. Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-Order Perturbation Theory with a CASSCF Reference Function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

(6) Andersson, K.; Malmqvist, P. Å.; Roos, B. O. Second-Order Perturbation Theory with a Complete Active Space Self-Consistent Field Reference Function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(7) Aquilante, F.; Autschbach, J.; Carlson, R. K.; Chibotaru, L. F.; Delcey, M. G.; De Vico, L.; Fdez. Galván, I.; Ferré, N.; Frutos, L. M.; Gagliardi, L.; et al. Molcas 8: New Capabilities for Multiconfigurational Quantum Chemical Calculations across the Periodic Table. *J. Comput. Chem.* **2016**, *37*, 506–541.

(8) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. Density Matrix Averaged Atomic Natural

Orbital (ANO) Basis Sets for Correlated Molecular Wave Functions. *Theor. Chim. Acta* **1990**, *77*, 291–306.

(9) Aquilante, F.; Bondo Pedersen, T.; Sánchez De Merás, A.; Koch, H. Fast Noniterative Orbital Localization for Large Molecules. *J. Chem. Phys.* **2006**, *125*, 174101.

(10) Angeli, C.; Cimiraglia, R.; Evangelisti, S.; Leininger, T.; Malrieu, J. P. Introduction of N-Electron Valence States for Multireference Perturbation Theory. *J. Chem. Phys.* **2001**, *114*, 10252.

(11) Li Manni, G.; Carlson, R. K.; Luo, S.; Ma, D.; Olsen, J.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2014**, *10*, 3669–3680.

(12) Ghigo, G.; Roos, B. O.; Malmqvist, P. Å. A Modified Definition of the Zeroth-Order Hamiltonian in Multiconfigurational Perturbation Theory (CASPT2). *Chem. Phys. Lett.* **2004**, *396*, 142–149.

(13) Forsberg, N.; Malmqvist, P. Å. Multiconfiguration Perturbation Theory with Imaginary Level Shift. *Chem. Phys. Lett.* **1997**, *274*, 196–204.

(14) Veryazov, V.; Widmark, P.-O.; Serrano-Andrés, L.; Lindh, R.; Roos, B. O. 2MOLCAS as a Development Platform for Quantum Chemistry Software. *Int. J. Quantum Chem.* **2004**, *100*, 626–635.

(15) Hulburt, H. M.; Hirschfelder, J. O. Potential Energy Functions for Diatomic Molecules. *J. Chem. Phys.* **1941**, *9*, 61–69.

(16) Araújo, J. P.; Alves, M. D.; da Silva, R. S.; Ballester, M. Y. A Comparative Study of Analytic Representations of Potential Energy Curves for O_2, N_2, and SO in Their Ground Electronic States. *J. Mol. Model.* **2019**, *25*, 198.

(17) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; Walt, S. J.; Brett, M; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**.

(18) B. Narayan. In *Fundamentals of Spectroscopy*; Allied Publishers, 2011.

(19) O. Roos, B.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. Main Group Atoms and Dimers Studied with a New Relativistic ANO Basis Set. *J. Phys. Chem. A* **2003**, *108*, 2851–2858.

(20) Peterson, K. A.; Feller, D.; Dixon, D. A. Chemical Accuracy in Ab Initio Thermochemistry and Spectroscopy: Current Strategies and Future Challenges. *Theor. Chem. Acc.* **2012**, *131*, 1079.

(21) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; ACM: New York, NY, USA, 2016; pp 785–794.

(22) Nielsen, D. Tree Boosting With XGBoost - Why Does XGBoost Win "Every" Machine Learning Competition? 2016. NTNU Open Web Site. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2433761 (accessed Jul 25, 2019)

(23) Kaggle competitions. https://www.kaggle.com/competitions (accessed Jul 25, 2019).

(24) Bergstra, J.; Komer, B.; Eliasmith, C.; Yamins, D.; Cox, D. D. Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization. *Comput. Sci. Discov.*

2015, *8*, 14008.

(25) Lobo, J. M.; Jiménez-Valverde, A.; Real, R. AUC: A Misleading Measure of the Performance of Predictive Distribution Models. *Glob. Ecol. Biogeogr.* **2008**, *17*, 145–151.