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

Authors: 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.

Abstract

Predicting and understanding the chemical bond is one of the major challenges of computational quantum chemistry. Kohn−Sham density functional theory (KS-DFT) is the most common method, but approximate density functionals may not be able to describe systems where multiple electronic configurations are equally important. Multiconfigurational wave functions, on the other hand, can provide a detailed understanding of the electronic structure and chemical bond of such systems. In the complete-active-space self-consistent field (CASSCF) method one performs a full configuration interaction calculation in an active space consisting of active electrons and active orbitals. However, CASSCF and its variants require the selection of these active spaces. This choice is not black-box; it requires significant experience and testing by the user, and thus active space methods are not considered particularly user-friendly and are employed only by a minority of quantum chemists. Our goal is to popularize these methods by making it easier to make good active space choices. We present a machine learning protocol that performs an automated selection of active spaces for chemical bond dissociation calculations of main group diatomic molecules. The protocol shows high prediction performance for a given target system as long as a properly correlated system is chosen for training. Good active spaces are correctly predicted with a considerably better success rate than random guess (larger than 80% precision for most systems studied). Our automated machine learning protocol shows that a "black-box" mode is possible for facilitating and accelerating the large-scale calculations on multireference systems where single-reference methods such as KS-DFT cannot be applied.

■ Introduction

A wide range of advancements have been provided by chemistry over the last several decades, especially through materials discovery,[1–3] but many of the most important discoveries have

benefited from an irreplicable degree of luck.[4] Meanwhile, chemistry is faced with increasing challenges such as renewable energy production and storage, and these challenges are only growing in urgency.[2,4–7] It has therefore been proposed that automated processes for materials discovery could enable large-scale systematic exploration of chemical space without requiring extensive effort by human researchers at each step.[1–4,6–8] Such automation requires advances in human-computer interfacing, robotic synthesis, and artificial intelligence-driven theory.[1,2,5–8]

With regards to theoretical developments, efforts are ongoing to use machine learning (ML) to enhance computational chemistry,[3,9–22] including to predict the results of many calculations without having to perform more than a few explicitly,[1,2,23–31] or to obtain high-level results with inexpensive methods.[23,30–42] In most cases a significant degree of computational effort is required to obtain necessary training data, and it has been repeatedly noted in the literature that one of the bottle-necks in progress towards generally applicable or automated machine learning is the insufficiency of current databases.[2,4,5,23,30,34,43] While large set of computational results have been compiled, they have often been performed under different conditions and for different applications, which limits their use in more general applications.[2,4,5] The immediate solution, generating large sets of consistent data, would require the performance of many calculations in a systematic (preferably automated) fashion.[3,30] In the long term, hopes of efficiently exploring extremely large sections of chemical space depend on the ability to automatically set up artificial intelligence protocols on the fly,[1,4,13] including automatically generating additional training data as needed.[13,44] It has also been noted that a robust program would need the ability to consider various levels of theory and find appropriate balances of cost and accuracy.[5,8,43]

Among electronic structure theory methods, Kohn-Sham density functional theory (KS-DFT)[45] is popular in a variety of contexts due to its simplicity of use, relatively good accuracy,

and low cost in comparison to wave function theory.[4] For machine learning, it is common to make KS-DFT results the predictive target,[23,26,28,31–35,42,46,47] although increasingly there have also been papers focused on single-reference wave function theory such as coupled-cluster.[30,33,36–41] However, both standard coupled-cluster methods and approximate KS-DFT functionals have weaknesses in multireference cases,[5,35,48–51] and the many different density functionals can provide different results.[4,51–53] Previous studies that focused solely on high-throughput screening for a specific application using KS-DFT or semi-empirical methods have relied on the outcome having a high tolerance for error so long as the rankings are unaffected,[4,5] but such rankings are application-specific, and the error tolerance does not necessarily extend to the general case. For autonomous machine learning protocols to be as robust as is desired, there will inevitably be some cases that require training using higher level multireference wave function theory[8] such as complete active space self-consistent field theory (CASSCF).[54]

In CASSCF, the wave function is a linear combination of multiple electronic configurations, the relative weights of which are determined through variational optimization. Because the computational expense scales with the number of configurations and is typically considered unaffordable with more than 1 billion configurations,[55] an active space is selected and all configurations possible for a specified number of electrons within the "active" orbitals are considered (within spin and spatial symmetry constrains, if applied by the user). Molecular orbitals lower in energy than the active orbitals are "inactive" and are doubly occupied in all configurations included in the wave function. Similarly, orbitals higher in energy than the active orbitals are "virtual" and are unoccupied in all considered configurations. (In this work we follow the convention of labeling active spaces by size as (n,N), where n is the number of active electrons and N is the number of active orbitals.)

Because the CASSCF active space is selected by the user, there is abundant opportunity for error and inefficiency arising from subjective human opinion and inexperience, and expertise is necessary for the identification of the most important orbitals. This human element has been cited as a significant obstacle to new users and to the automation of the method for large-scale applications.[8,56–58] There have been many attempts to establish schemes for active space selection,[57–64] but none have yet achieved widespread adoption, and it has been argued that no system of rules can generally apply to all systems of interest (let alone be put into a computer code).[56] Accordingly, automation of active space methods may require on-the-fly solutions, such as the AutoCAS approach of Stein and Reiher.[65–68] AutoCAS attempts to remove the uncertainty of active space selection via a multi-stage approach that uses the approximate CI-solver of density matrix renormalization group (DMRG)[69–71] with very large active spaces and then increases the level of accuracy with smaller active spaces selected based on orbital entropy information. This strategy may prove to be beneficial to researchers seeking assistance with active space selection in specific cases, but the preceding DMRG calculations introduce additional computational overhead for each specific case,[68] thus adding considerable expense if used in a large-scale machine learning training protocol or autonomous materials discovery lab. Solutions have been proposed for avoiding the active space problem by selecting specific configurations,[72–78] but these approaches could face similar overhead problems when scaled to large numbers of systems.

In this work, we propose the use of machine learning to predict if active spaces for systems outside of the ML training set are good or bad. Our model classifies a test active space as good or bad based on a limited number of CASSCF calculations. Transferability between different chemical systems is a primary goal, and while we note that Miller and coworkers[38,39] have worked on predicting energies from single reference methods in a transferable manner, to the best of our

knowledge no one else has used machine learning for active space selection in multireference wave function calculations. In this initial offering we focus on obtaining accurate descriptions of potential energy curves (PECs) for bond dissociations of main group diatomic molecules with reference to experiment. We demonstrate that when we train on a limited subset of molecules that are highly correlated to a target molecule we can predict good active space selections for molecules outside of the training set with a significantly better success rate than random guess. We begin with a summary of the machine learning strategy, including how to generate features and labels automatically, followed by a discussion of our results thus far: In the conclusion we address plans for further development. Additional technical details are presented in the Supporting Information.

Automated Active Space Selection Using a Machine Learning Model

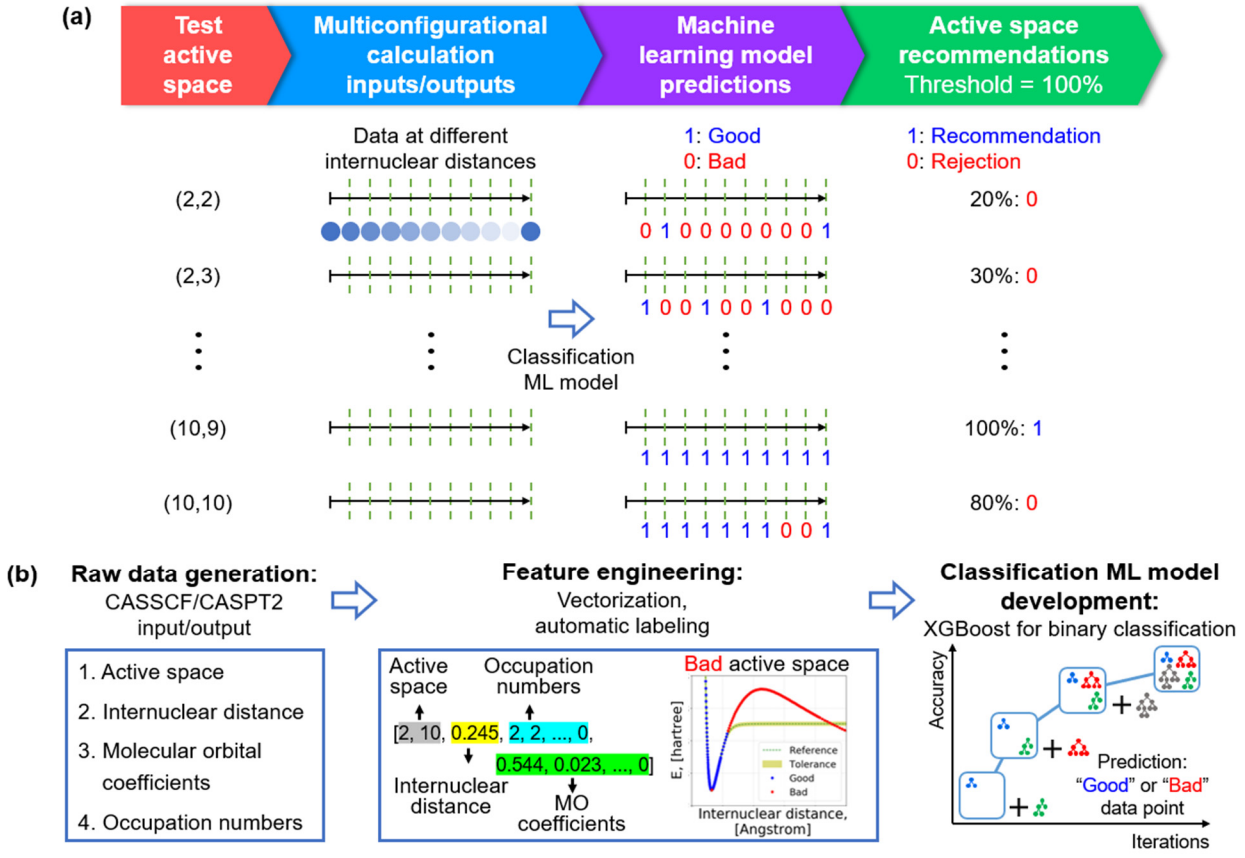

Figure 1. (a) Schematic of the automated active space selection protocol. CASSCF calculation inputs/outputs at several different internuclear distances are needed for each test active space. A given CASSCF data point at different internuclear distances is classified as good or bad as part of a PEC. If all of the tested data points of a given active space are predicted to be good, the active space is predicted to be good. (b) Development of the supervised ML classification mode using XGBoost, a gradient boosting decision-tree based algorithm. Input parameters and outputs of multiconfigurational calculations are used as raw data and converted to vectorized training features. Label codes (indicating good or bad) are generated via automated protocol.

For automated active space selection in calculations of bond dissociations, we propose an automated protocol using a classification machine learning (ML) model (Figure 1a). The protocol extracts a data point from a CASSCF input/output at a specific internuclear distance and classifies it as either good or bad. A good data point should be one of the data points in a desirable PEC that yields relatively good dissociative/spectroscopic properties such as the bond dissociation energy, equilibrium bond length, and vibrational constants. A variety of data points at different distances are predicted for each test active space. The overall active space is predicted to be good if 100%

of the predicted individual data points are good. The 100% threshold was adopted to be conservative and to avoid ambiguity that might come from a user-specific parameter. Note that one does not need to perform many calculations consecutively until confirming that a desirable (i.e., a smooth and continuous) shape of a PEC is obtained. Also, there is no requirement of additional computational cost on top of CASSCF calculations such as second-order perturbation theory correction (CASPT2) to obtain more quantitatively accurate results, e.g., for comparison with experimental dissociative/spectroscopic properties data. While CASPT2 calculations are required to generate labels for training the supervised ML model (Figure 1b), once the training of the ML model is completed the CASPT2 calculations are no longer needed. Additionally, the automated ML protocol can identify which active spaces would be worthwhile choices even for the cases where simulated PECs do not perfectly agree with the experimental data, possibly due to the limit of the theory or an insufficient basis set size.

In this automated active space selection protocol, the internal workings of the ML model (i.e., generation of features and refining of ML hyperparameters) are also automated (Figure 1b). Raw data were generated with CASSCF/CASPT2 using MOLCAS 8.2.[55] For simplicity, the dissociation of neutral, main group diatomic molecules was investigated, and the raw data were generated for only ground-state spin states with active spaces up to (10,10). Additionally, molecular orbital (MO) ordering was not manually altered from the initial guess orbitals. Features are extracted from both the inputs and outputs of the CASSCF/CASPT2 calculations and converted to a fixed-length vector of 345 input features. The input parameters include the number of active electrons, the number of active orbitals, the internuclear distance, and the CASSCF output results such as occupation numbers and MO coefficients.

In order to train and evaluate the ML model, two different types of labels were determined via a systematic procedure. A label code for each data point is required (i.e., 1 for good or 0 for bad with respect to whether the given data point contributes to obtaining a good PEC), and a label code for each active space (i.e., 1 for recommendation or 0 for rejection with regard to whether the active space selection enables us to have a good PEC) is generated to evaluate performances of our protocol for the final recommendation of good active spaces. The data point label codes are generated from the vectorized features in an automated way without any manual checking of the multiconfigurational calculation outputs. Active spaces are classified into either good or bad based on the proportion of good-labeled data points for each active space (with 90% being the threshold in this work as described in the next subsection). Finally, the gradient boosting decision tree-based algorithm XGBoost (eXtreme Gradient Boosting)[79,80] was selected to develop the classification ML model. Further details regarding the production and featurization of the raw data, and hyperparameter optimization of the ML model are available in Supporting Information.

Automated Labeling of Features and Active Spaces

A key objective of this work is to differentiate good active spaces from bad ones, which requires defining what good active space selections are. In addition, in our scheme the classification ML model should be capable of identifying whether the given features are correlated to either the good or bad active space selections. In order to obtain accurate label codes, we devised an automated labeling procedure (Figure 2). The Hulburt-Hirschfelder (HH) potential function[81,82] was adopted as the baseline data for comparison with our simulation data. The HH potential was chosen because it allows derivation of an accurate shape of a PEC based on experimental dissociative/spectroscopic constants such as the bond dissociation energy (D_e), the equilibrium

bond length (r_e), the vibrational harmonic and anharmonic constants (ω_e and $\omega_e x_e$), the vibration-rotation coupling (α_e), and the rotational constant (B_e). PECs of different active spaces are likely to have slightly different absolute energy values, so the test PECs were shifted with respect to the HH PEC. After that, the degree of the discrepancies between test PECs and the HH PEC was measured by computing the area between the PECs (referred to as a "deviation area"). Details of the HH potential functions and related parameters, and the PEC shifting algorithms are available in the Supporting Information.

The PEC that exhibits the minimum deviation area was set as a reference PEC. Here, we assumed that the selected reference PEC of a specific active space is the best standard data among the generated raw data with different active spaces. Although it could be possible to obtain more accurate results with a larger basis set and larger active spaces, our aim is to select good active spaces among available data that we generated at an affordable computational cost. In the future the same procedure will be used to explore larger active spaces and basis sets. Using the extracted reference PEC data and a specified tolerance of deviation from the reference data, label codes were assigned to each data point at different interatomic distances by shifting test PECs along with the reference one. Details of the automated labeling are available in the Supporting Information. Additionally, label codes for active space selection (i.e. for the whole PEC data points) are determined by checking whether the percentage of good data points in the distance range from 0.5 to 10.0 Angstroms is within a threshold of 90% based on 50 nearly equidistant samples.

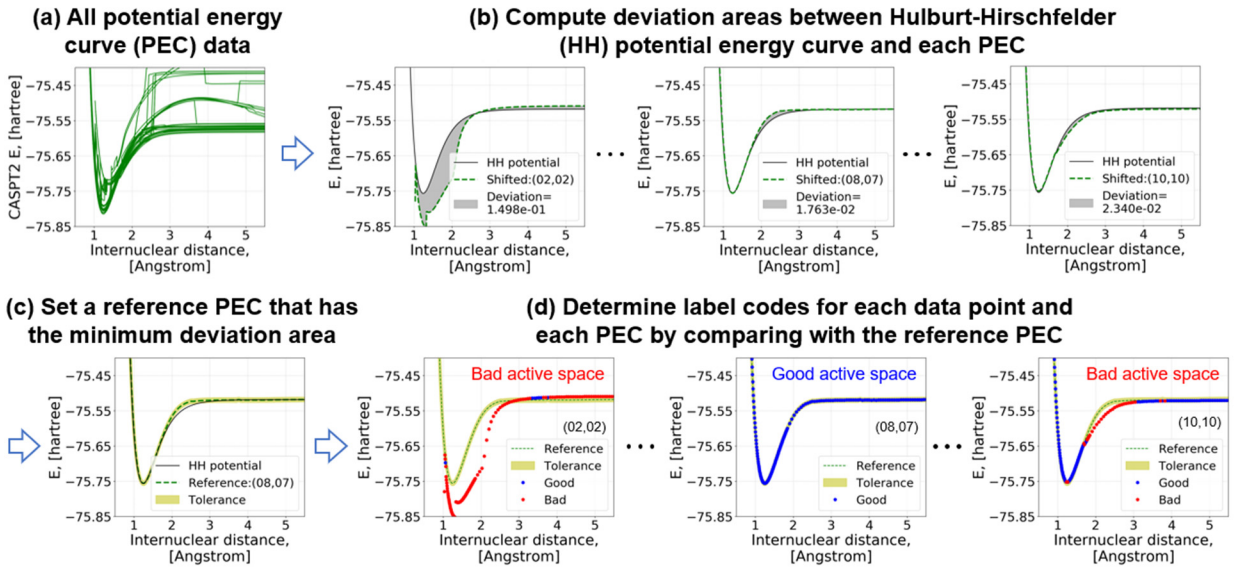

Figure 2.Automated procedure of the label code generation (using diatomic carbon as an example). (a) Many different PECs are generated by the different active spaces. (b) The HH PEC is prepared based on experimental bond dissociative properties. Testing PECs are shifted with respect to the HH PEC, and then the deviation areas are computed. (c) The PEC showing the smallest deviation area is selected as a reference PEC. (d) A label code for each data point at different internuclear distances is determined. A list of good active space selections is made based on the percentage of good data points (here a somewhat conservative value, 90%, is used).

■ Results and Discussion

Generation of Label Codes for Identifying Good Active Space Selection

A total of 23 main group diatomic molecules were selected based on availability of experimental data for bond dissociation: H_2, Li_2, B_2, C_2, N_2, O_2, F_2, LiH, BeH, BH, CH, OH, HF, BN, CN, LiO, BeO, BO, CO, NO, FO, LiF, and CF. Active spaces where CASSCF/CASPT2 calculations were not converging within the range of 0.5 Å to 10.0 Å were excluded from the evaluation of the final active space recommendation, but any converged individual data points were labeled and used for training the ML model. The best active space selections and corresponding errors of bond dissociative properties are listed in Table S1. It was found that CASSCF/CASPT2 calculations described incorrect states for BeO and LiF (Figure S4 and surrounding discussion), so they were excluded from further investigation. In contrast, H_2 was also

excluded because the multiconfigurational calculations described the PECs very well with all active space. There were only a few bad labeled data points (i.e, only 34 points, 1.95% of entire data for H_2), resulting in severely imbalanced data that could not be used for training and testing an ML model (Figure S5).

Using the HH potential for selecting the reference PECs is advantageous in several respects. First of all, considering different and system-dependent orders of errors for each property, using the HH potential enables us to avoid a tricky problem of deciding what threshold values for each property and each system should be used. Second, it allows us to determine how much deviation from desirable PEC shape are acceptable when there are few perfectly smooth and continuous PECs. For the majority of PECs computed using CASSCF/CASPT2 discontinuities, cusp and/or abrupt jump appear at near the equilibrium distance, the region where the potential energy starts to flatten out, or at large separations (Figure S1). These abnormalities in the PECs are pervasive for all multiconfigurational methods except full CI as the dominant electron configurations varies considerably due to significant change of system geometry.[60,83] Third, PECs having similar errors for dissociative properties but with different curvatures can be easily distinguished (e.g., C_2 and CO in Figures S6 and S7, and Table S2). Distinguishing between the different curves is important because including wrongly labeled data points in a training data set could lead to a reduction in ML model performance.

Prediction Performances of Classification Machine Learning Models

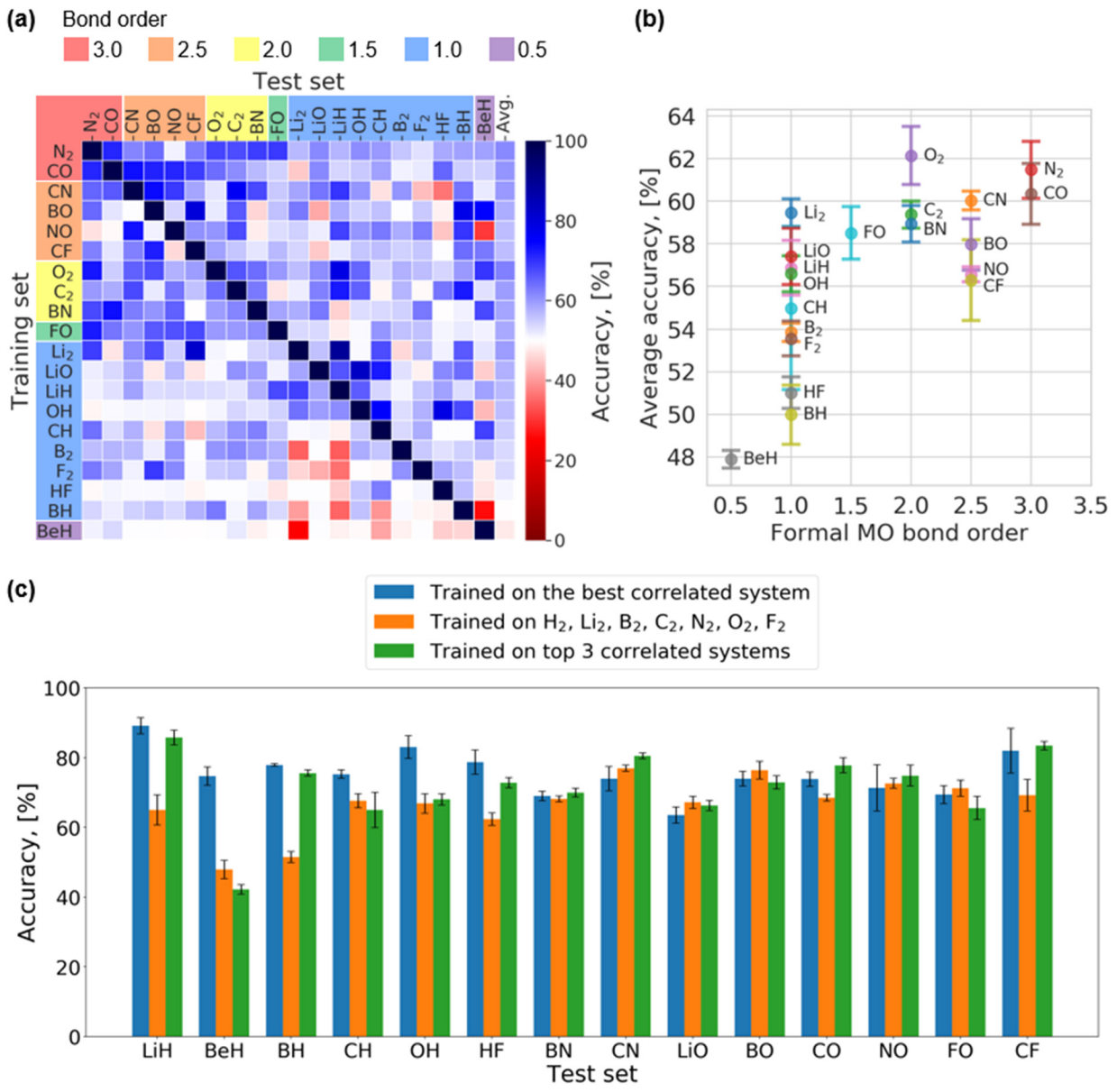

Figure 3. Prediction performances of the classification machine learning models trained on different single systems or combinations of systems. (a) Heat map of prediction performances of ML models that were trained on a single diatomic system and tested for each system. The heat map was constructed by averaging 10 separate calculations for extracting a general trend taking into account the stochastic nature of training ML models. The heat map is arranged by formal molecular bond order and average accuracy for a given training set. (b) Correlations between average accuracy of ML models trained on single diatomic system and the formal MO bond order. The average accuracy is defined as the average of accuracies for predictions on other 19 systems except the system used for the training of the ML model that trained on a single diatomic systems (c) Comparison of the ML classifiers trained on different training options: (1) the best correlated system, (2) all available homonuclear systems, and (3) the top 3 correlated systems.

As a first step towards developing a transferable ML model across various diatomic systems, we trained classification ML models on each single diatomic system and then predicted for each diatomic system. The prediction accuracies of each ML model on the main group diatomic

systems are plotted in the format of a heatmap as shown in Figure 3a. Since the hyperparameter tuning and shuffling/sampling of training/test data points were performed stochastically with different random seeds at each time, the performance of the developed ML models trained on even the same diatomic system could be varied to some extent for each run. For better statistics of the ML performances, 10 separate ML model training and predictions were conducted to generate 10 heat maps, and average accuracies for each cell in the heatmaps were used for Figure 3a. The average of the standard deviation for each cell of the 10 heatmaps is 2.9%, and the maximum standard deviation is 14.4% for the case where the ML model was trained on CH and tested on LiH. Most importantly, the ML model can predict well for systems that are not in the training set. For instance, the ML model trained on Li_2 data can achieve the largest prediction accuracy (89.2%) among all paired combinations of dissimilar diatomic molecules as training/test sets. This indicates that there are correlations between two different diatomic systems depending on which pair of diatomic systems are used for the training and test sets, respectively. Here we refer to a pair of systems as (properly) correlated if the molecule used for training that leads to an accuracy larger than 50% (i.e, a random guess) for the target molecule. Note that the heatmap is not symmetric, which means that the degree of correlation between a pair of diatomic molecules is asymmetric. For example, the ML models trained on Li_2 show an average prediction accuracy of 81.9% for CF, while those trained on CF exhibited 54.0% prediction accuracy for Li_2.

The varying degrees of correlation between different diatomic molecules might come from (1) data inconsistency such as different numbers of data points and explored active spaces for different diatomic molecules, and/or (2) different chemistry in the dissociation of the molecules. To test the data inconsistency hypothesis, heatmaps of the ML model accuracies were generated with the same numbers of training data points for each system (i.e., 1000, 2000, 3000, 5000. Figure

S8). Regardless of the number of data points used for the training, the trends of the correlations between different systems is maintained (Figures S8 and S9). The number of explored active spaces can also be excluded as the critical factor, as there are several diatomic systems such as CH, F_2, and HF that have a relatively large number of possible active spaces but a relatively low average accuracy (Figure S10). Therefore, the data inconsistency hypothesis is discarded.

To test the chemistry-based hypothesis, we examined correlations between the average accuracy from single-system training and simple bonding characteristics such as the formal molecular orbital bond order and average electronegativity (i.e., average of electronegativities of the constituent atoms, Figure 3b and Figure S11a). This investigation was motivated by the work by Shaik and coworkers that shows the bond character is correlated with the electronegativities of bonded atoms.[84,85] We observe training on a diatomic molecule having larger bond order generally results in a higher average accuracy of the ML model (Figure 3b). This can be interpreted as features for the systems with higher bond order possessing more transferable information on bond dissociation such as similar correlated MOs in terms of sigma and pi bonds. In contrast, training on a diatomic system with low bond order (i.e., 0.5, and 1.0 except for diatomic molecules containing Li) does not predict well for other systems, and it may be that training systems having only sigma bond character makes them less transferrable to the other systems. For example, when an ML model is trained on BeH, the data points of BeH include information on the bond breakage only with a bond order of 0.5, and the ML model acts like a random guess (i.e, showing nearly 50% accuracy) for other systems except Li_2. On the other hand, ML models trained on either BO or O_2 are able to predict BeH relatively well (i.e, 74.7% and 69.7%, for BO and O_2, respectively).

In order to develop a simple descriptor for identifying the correlation between the diatomic systems, the MO bond order and average electronegativity of the systems were max-min scaled

separately, and then averaged to produce a new metric (Figure S11b). Although the new metric may show better linear correlation with average accuracy, there are outliers such as Li_2, LiH, and LiO. While we identified correlations between bond characteristics such as bond order and average electronegativity, we cannot designate which system(s) are the most correlated and therefore define an effective single-system training set for predicting a given target system. Instead, we turn to multi-system training sets to improve transferability.

Figure 3c compares the accuracy of ML models trained on (1) the best correlated system, (2) all available homonuclear systems (i.e., H_2, Li_2, B_2, C_2, N_2, O_2, and F_2), and (3) top 3 correlated systems. The best correlated and top 3 correlated systems are listed in Table S5. For predicting H-containing systems, the homonuclear systems training set has much lower prediction performances than training with the top one or three correlated systems. H-containing diatomic systems have low bond order and no pi bonding character, so it is possible that the extra information from the more diverse system set is leading to the degradation in performance. Instead, for H-containing systems, only the best correlated system can be used to develop a good-performing ML model. Conversely, diatomic molecules containing a N or O atom have a larger bond order (and therefore both sigma and pi bonds) and cover more chemical space on dissociation of the molecules, so the ML performances using the best correlated single system and the homonuclear systems are similar. Using the top 3 correlated systems for training improves ML model performance only for predicting CN, CO and CF, and the degree of the improvement is not significant. Additionally, predicting BeH, using top 3 correlated systems for training decreases the ML model performance considerably compared to the training on the best system since BeH has almost no properly correlated systems. This indicates that developing a good-performing ML model is possible only when the correlated system(s) is carefully chosen for a given target diatomic system.

Recommendation of Good Active Space Selections

Good active spaces can be predicted using the developed classification ML models and a specific number of CASSCF data points. To test how many data points are required, different sample sizes of 10, 30, or 50 points per active space were employed with the ML models trained on the best correlated system for each heteronuclear system (Figure 4a). The sampling was done by selecting data points nearly equidistant ranging from 0.5 to 10.0 Angstroms. Except for HF and BN, the accuracies of the recommendation for the heteronuclear systems are largely independent of sample size, i.e., only 10 CASSCF data points per test active space are sufficient to determine whether the given active space will be good.

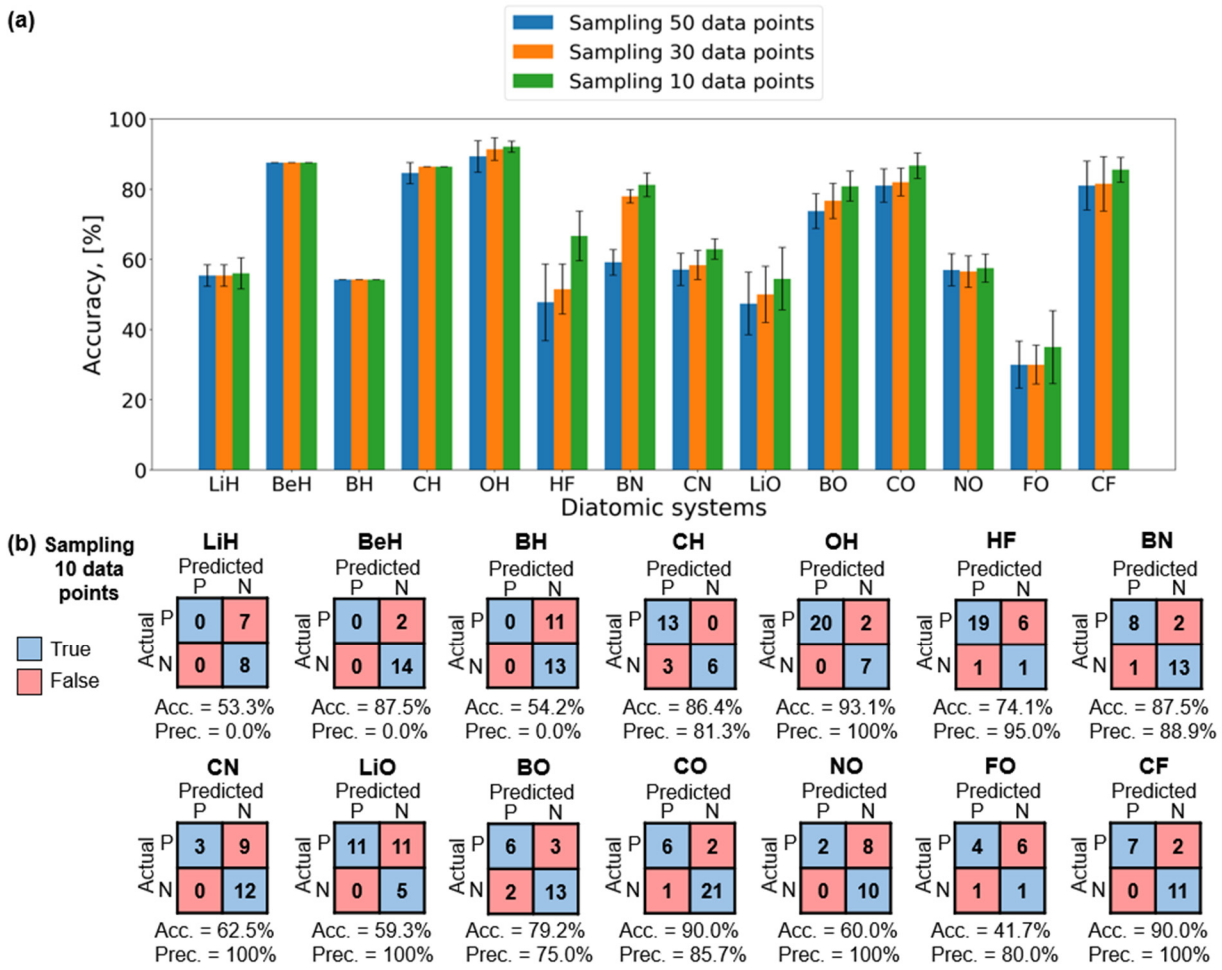

Figure 4. (a) Recommendation accuracy comparison for selecting good active spaces with different data point sample sizes. Recommendation accuracies were calculated by averaging 10 different recommendations using 10 different ML models trained separately. (b) Confusion matrixes of the recommendations for heteronuclear diatomic molecules based on ensemble of the 10 different ML models (a 50% vote is considered a "recommendation"). P represents a positive and N indicates a negative.

The minimum and maximum prediction performances of the ML models are 63.5% and 89.1% (for LiO and LiH as the target system, respectively), and average accuracy is 75.4%, as shown in Figure 3c. While overall the ML classifiers are not perfect, the recommendation accuracy is surprisingly high: ≥80% for BeH, CH, OH, BN, BO, CO, and CF. However, accuracies of active space recommendation for LiH, BH, HF, CN, LiO, NO, and FO are relatively low despite the ML models for these systems exhibiting higher prediction performance (89.1%, 77.9%, 78.7%, 73.9%, 63.5%, 71.3%, and 69.4%, respectively). These discrepancies could be attributed to different data sampling methods in terms of internuclear distances: denser sampling near the equilibrium

distance for developing/testing the ML models, and equidistant sampling for the final recommendation of active spaces.

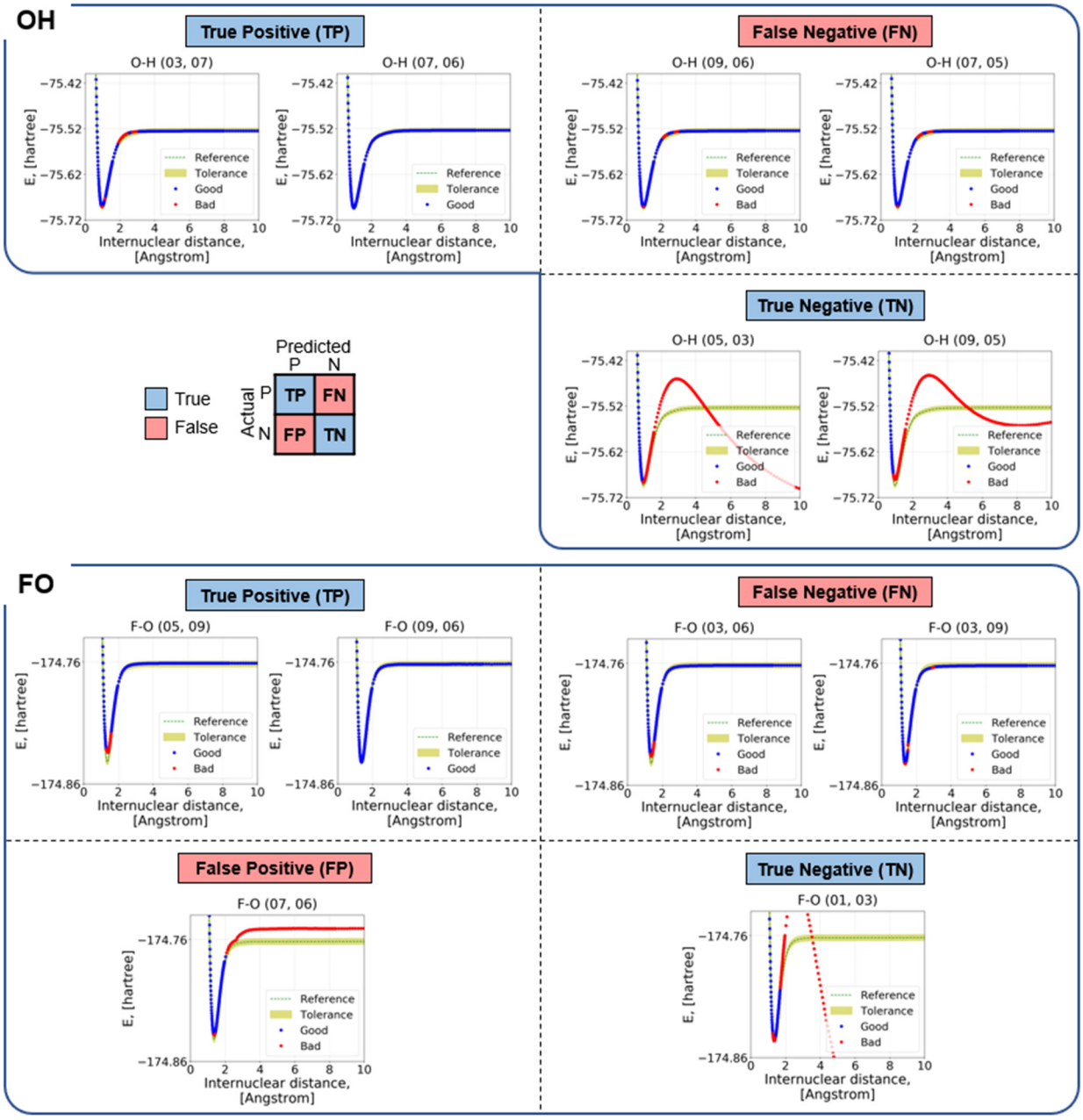

Figure 5. Representative PECs for each case of the confusion matrixes for OH (the best recommendation accuracy) and FO (the worst recommendation accuracy). Note that the good/bad labels of each data points and the actuals of the curves are assigned via the automated labeling procedure, not predicted based on an ML model.

For further analysis, confusion matrixes are produced for evaluating binary classification predictions of active spaces (Figure 4b). Each active space is automatically labelled as "actually good" or "actually bad" and this label is compared to the ML prediction. "True positive" (TP) means both the automated labeler and the ML protocol indicate a good active space, "true negative" (TN) means both the labeler and the ML protocol indicate a bad active space, "false positive" (FN) means the active space was labeled bad but the ML protocol predicted it would be good, and "false negative" (FN) means the active space was labeled as good but the ML protocol predicted it would be bad. The ML protocol predicts bad active spaces well, i.e., TN rather than FP. This is because, for the bad active spaces, PECs are considerably deviated from the corresponding reference PEC for most of the diatomic systems we investigated. For example, TN PECs for OH and FO (Figure 5) have distinct features such as of a hump or unconverging energies at large separations. Likewise, there are various abnormal curve shapes observed for bad active spaces of other diatomic systems, such as discontinuities at short or long internuclear distances that could be easily discriminated from the reference curves. (More examples of bad active spaces are available in the Supporting Information, Figures S12~S22.) The low recommendation accuracies of some systems (LiH, BH, HF, CN, LiO, NO, and FO) arose not from failure to identify bad active spaces but from overzealously labeling some good active spaces as bad (FN rather than TP). One explanation is that the PECs in question have similar shapes that are not easily differentiated using the ML models. Due to the small differences in PEC shape resulting in small differences between features (i.e., occupation numbers and MO coefficients), the ML model cannot be trained well with the features.

Despite the largely fluctuating recommendation accuracies depending on a target diatomic system, the most beneficial aspect of the ML protocol is that the protocol shows precision values

greater than 80%, excluding LiH, BH, and BeH. (Here "precision" = TP / (TP + FP).) This means that predicted good active spaces by the ML protocol are most likely actual good active spaces that can be used to perform multiconfigurational calculations. Additionally, the ML protocol successfully identified at least one of the three smallest good active spaces for the majority of the target systems (Table S6), which is beneficial for minimizing computational cost.

■ Conclusion

In order to eliminate subjective human intervention in the active space selection for multiconfigurational calculations, we propose an automated protocol based on a supervised ML classifier. The ML protocol development included automated labeling of features and hyperparameter tuning, and we demonstrated that a high-performance can be achieved only when the properly correlated diatomic system(s) is chosen for the training. Our ML protocol can correctly predict many good active spaces and most of the bad ones and does not require any chemical intuition in choosing an appropriate active space, enabling a "black-box" mode for multiconfigurational calculations for large-scale screenings. To the best of our knowledge we are the first to demonstrate that machine learning can be applied to active space selection in multiconfigurational methods.

For future work, we are pursuing several directions. The current ML protocol could be made more fully automated in terms of selection of properly correlated system(s) or data for a given target systems. This could be achieved by developing simple chemical descriptors or secondary ML models that measure a degree of similarity between systems/data. We also intend to expand the chemical space, beginning with diatomic molecules containing 3rd, and 4th period elements and in a subsequent phase expanding to more general and complex systems. Such

expansions will move beyond cases where reliable reference data exists, which will require upgrading the automated labeling scheme with unsupervised ML algorithms that can be used to evaluate given data and generate label codes. We also plan to eliminate the need for complex and time-consuming CASSCF calculations by building a regression ML model to generate MO input data. Lastly, we will improve the specificity of the active space recommendations with regards to the specific orbitals that should be included in the active space, and efforts to develop more interpretable MO features that could replace MO occupation numbers and MO coefficients are underway. In the long term, we expect the ML technique presented here to help those performing multireference calculations with selecting appropriate active spaces, thereby facilitating the use of this wave function theory model that is especially appropriate in various strongly correlated systems.

■ **Acknowledgments**

This work was performed at the University of Minnesota and was supported as part of the Nanoporous Materials Genome Center, funded by the U.S. Department of Energy, Office of Basic Energy Sciences, under Award DE-FG02-17ER16362, as part of the Computational Chemical Sciences Program. Computer resources were provided by the Minnesota Supercomputing Institute at the University of Minnesota. R.L. acknowledges the Swedish research council (grant 2016-03398) and the Olle Engkvist foundation (grant 18-2006).

■ References

(1) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282–291.

(2) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; et al. Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation. *Nat. Rev. Mater.* **2018**, *3* (5), 5–20.

(3) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12* (3), 191–201.

(4) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45* (1), 195–216.

(5) Alberi, K.; Nardelli, M. B.; Zakutayev, A.; Mitas, L.; Curtarolo, S.; Jain, A.; Fornari, M.; Marzari, N.; Takeuchi, I.; Green, M. L.; et al. The 2019 Materials by Design Roadmap. *J. Phys. D. Appl. Phys.* **2019**, *52* (1), 013001.

(6) Wei, J.; De Luna, P.; Bengio, Y.; Aspuru-Guzik, A.; Sargent, E. Use Machine Learning to Find Energy Materials. *Nature* **2017**, *552* (7683), 23–25.

(7) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)Evolution. *ACS Cent. Sci.* **2018**, *4* (2), 144–152.

(8) Unsleber, J. P.; Reiher, M. The Exploration of Chemical Reaction Networks. **2019**, https://arxiv.org/abs/1906.10223v1.

(9) Häse, F.; Fdez. Galván, I.; Aspuru-Guzik, A.; Lindh, R.; Vacher, M. How Machine Learning Can Assist the Interpretation of Ab Initio Molecular Dynamics Simulations and Conceptual Understanding of Chemistry. *Chem. Sci.* **2019**, *10* (8), 2298–2307.

(10) Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-Throughput Screening of Bimetallic Catalysts Enabled by Machine Learning. *J. Mater. Chem. A* **2017**, *5* (46), 24131–24138.

(11) Tamayo-Mendoza, T.; Kreisbeck, C.; Lindh, R.; Aspuru-Guzik, A. Automatic Differentiation in Quantum Chemistry with Applications to Fully Variational Hartree-Fock. *ACS Cent. Sci.* **2018**, *4* (5), 559–566.

(12) Häse, F.; Valleau, S.; Pyzer-Knapp, E.; Aspuru-Guzik, A. Machine Learning Exciton Dynamics. *Chem. Sci.* **2016**, *7* (8), 5139–5147.

(13) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114* (9), 096405.

(14) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148* (24), 241722.

(15) Häse, F.; Kreisbeck, C.; Aspuru-Guzik, A. Machine Learning for Quantum Dynamics: Deep Learning of Excitation Energy Transfer Properties. *Chem. Sci.* **2017**, *8* (12), 8419–8426.

(16) Kim, S.; Jinich, A.; Aspuru-Guzik, A. MultiDK: A Multiple Descriptor Multiple Kernel Approach for Molecular Discovery and Its Application to Organic Flow Battery Electrolytes. *J. Chem. Inf. Model.* **2017**, *57* (4), 657–668.

(17) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M.

Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5* (1), 57–64.

(18) Moosavi, S. M.; Chidambaram, A.; Talirz, L.; Haranczyk, M.; Stylianou, K. C.; Smit, B. Capturing Chemical Intuition in Synthesis of Metal-Organic Frameworks. *Nat. Commun.* **2019**, *10* (1), 1–7.

(19) Panapitiya, G.; Avendano-Franco, G.; Ren, P.; Wen, X.; Li, Y.; Lewis, J. P. Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140* (50), 17508–17514.

(20) Chowdhury, A. J.; Yang, W.; Walker, E.; Mamun, O.; Heyden, A.; Terejanu, G. A. Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning. *J. Phys. Chem. C* **2018**, *122* (49), 28142–28150.

(21) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal-Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9* (16), 4562–4569.

(22) Bassman, L.; Rajak, P.; Kalia, R. K.; Nakano, A.; Sha, F.; Sun, J.; Singh, D. J.; Aykol, M.; Huck, P.; Persson, K.; et al. Active Learning for Accelerated Design of Layered Materials. *npj Comput. Mater.* **2018**, *4* (1), 1–9.

(23) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203.

(24) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A. A Mixed Quantum Chemistry/Machine Learning Approach for the Fast and Accurate Prediction of Biochemical Redox Potentials and Its Large-Scale Application to 315 000 Redox

Reactions. *ACS Cent. Sci.* **2019**, *5* (7), 1199–1210.

(25) Dral, P. O.; Barbatti, M.; Thiel, W. Nonadiabatic Excited-State Dynamics with Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9* (19), 5660–5663.

(26) Li, Y. P.; Han, K.; Grambow, C. A.; Green, W. H. Self-Evolving Machine: A Continuously Improving Model for Molecular Thermochemistry. *J. Phys. Chem. A* **2019**, *123* (10), 2142–2152.

(27) Novikov, I. S.; Suleimanov, Y. V.; Shapeev, A. V. Automated Calculation of Thermal Rate Coefficients Using Ring Polymer Molecular Dynamics and Machine-Learning Interatomic Potentials with Active Learning. *Phys. Chem. Chem. Phys.* **2018**, *20* (46), 29503–29512.

(28) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine Learning of Molecular Properties: Locality and Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241727.

(29) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less Is More: Sampling Chemical Space with Active Learning. *J. Chem. Phys.* **2018**, *148* (24), 241733.

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

(31) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K. R.; Anatole Von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

(32) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; Von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K. R. Assessment and Validation of Machine Learning

Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9* (8), 3404–3419.

(33) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11* (5), 2087–2096.

(34) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *34th Int. Conf. Mach. Learn. ICML 2017* **2017**, *3*, 2053–2070.

(35) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K. R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8* (1), 1-10.

(36) Margraf, J. T.; Reuter, K. Making the Coupled Cluster Correlation Energy Machine-Learnable. *J. Phys. Chem. A* **2018**, *122* (30), 6343–6348.

(37) Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10* (14), 4129–4135.

(38) Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14* (9), 4772–4779.

(39) Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A Universal Density Matrix Functional from Molecular Orbital-Based Machine Learning: Transferability across Organic Molecules. *J. Chem. Phys.* **2019**, *150* (13), 131103.

(40) McGibbon, R. T.; Taube, A. G.; Donchev, A. G.; Siva, K.; Hernández, F.; Hargus, C.; Law, K. H.; Klepeis, J. L.; Shaw, D. E. Improving the Accuracy of Møller-Plesset Perturbation Theory with Neural Networks. *J. Chem. Phys.* **2017**, *147* (16), 161725.

(41) Nudejima, T.; Ikabata, Y.; Seino, J.; Yoshikawa, T.; Nakai, H. Machine-Learned Electron Correlation Model Based on Correlation Energy Density at Complete Basis Set Limit. *J. Chem. Phys.* **2019**, *151* (2), 024104.

(42) Rupp, M.; Tkatchenko, A.; Müller, K. R.; Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.

(43) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38* (16), 1291–1307.

(44) Tran, K.; Ulissi, Z. W. Active Learning across Intermetallics to Guide Discovery of Electrocatalysts for CO2 Reduction and H2 Evolution. *Nat. Catal.* **2018**, *1* (9), 696–703.

(45) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140* (4A), A1133–A1138.

(46) Jasrasaria, D.; Pyzer-Knapp, E. O.; Rappoport, D.; Aspuru-Guzik, A. Space-Filling Curves as a Novel Crystal Structure Representation for Machine Learning Models. **2016**, http://arxiv.org/abs/1608.05747v1.

(47) Pyzer-Knapp, E. O.; Li, K.; Aspuru-Guzik, A. Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25* (41), 6495–6502.

(48) Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham Density Functional Theory Descending a Staircase. *J. Chem. Phys.* **2016**, *145* (13), 130901.

(49) Jiang, W.; Deyonker, N. J.; Wilson, A. K. Multireference Character for 3d Transition-Metal-Containing Molecules. *J. Chem. Theory Comput.* **2012**, *8* (2), 460–468.

(50) Wang, J.; Manivasagam, S.; Wilson, A. K. Multireference Character for 4d Transition

Metal-Containing Molecules. *J. Chem. Theory Comput.* **2015**, *11* (12), 5865–5872.

(51) Rugg, G.; Genest, A.; Rösch, N. DFT Variants for Mixed-Metal Oxides. Benchmarks Using Multi-Center Cluster Models. *J. Phys. Chem. A* **2018**, *122* (35), 7042–7050.

(52) Radoń, M. Revisiting the Role of Exact Exchange in DFT Spin-State Energetics of Transition Metal Complexes. *Phys. Chem. Chem. Phys.* **2014**, *16* (28), 14479–14488.

(53) Venturinelli Jannuzzi, S. A.; Phung, Q. M.; Domingo, A.; Formiga, A. L. B.; Pierloot, K. Spin State Energetics and Oxyl Character of Mn-Oxo Porphyrins by Multiconfigurational Ab Initio Calculations: Implications on Reactivity. *Inorg. Chem.* **2016**, *55* (11), 5168–5179.

(54) Roos, B. O. The Complete Active Space SCF Method in a Fock-matrix-based Super-CI Formulation. *Int. J. Quantum Chem.* **1980**, *18* (14 S), 175–189.

(55) 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* (5), 506–541.

(56) Pierloot, K. Nondynamic Correlation Effects in Transition Metal Coordination Compounds. In *Computational Organometallic Chemistry*; Cundari, T. R., Ed.; Marcel Dekker: New York, 2001; pp 123–158.

(57) Veryazov, V.; Malmqvist, P. Å.; Roos, B. O. How to Select Active Space for Multiconfigurational Quantum Chemistry? *Int. J. Quantum Chem.* **2011**, *111* (13), 3329–3338.

(58) Sayfutyarova, E. R.; Sun, Q.; Chan, G. K. L.; Knizia, G. Automated Construction of Molecular Active Spaces from Atomic Valence Orbitals. *J. Chem. Theory Comput.* **2017**,

13 (9), 4063–4078.

(59) Bofill, J. M.; Pulay, P. The Unrestricted Natural Orbital-Complete Active Space (UNO-CAS) Method: An Inexpensive Alternative to the Complete Active Space-Self-Consistent-Field (CAS-SCF) Method. *J. Chem. Phys.* **1989**, *90* (7), 3637–3646.

(60) Keller, S.; Boguslawski, K.; Janowski, T.; Reiher, M.; Pulay, P. Selection of Active Spaces for Multiconfigurational Wavefunctions. *J. Chem. Phys.* **2015**, *142* (24), 244104.

(61) Bao, J. J.; Dong, S. S.; Gagliardi, L.; Truhlar, D. G. Automatic Selection of an Active Space for Calculating Electronic Excitation Spectra by MS-CASPT2 or MC-PDFT. *J. Chem. Theory Comput.* **2018**, *14* (4), 2017–2025.

(62) Khedkar, A.; Roemelt, M. Active Space Selection Based on Natural Orbital Occupation Numbers from N-Electron Valence Perturbation Theory. *J. Chem. Theory Comput.* **2019**, *15* (6), 3522–3536.

(63) Sayfutyarova, E. R.; Hammes-Schiffer, S. Constructing Molecular π-Orbital Active Spaces for Multireference Calculations of Conjugated Systems. *J. Chem. Theory Comput.* **2019**, *15* (3), 1679–1689.

(64) Bao, J. J.; Truhlar, D. G. Automatic Active Space Selection for Calculating Electronic Excitation Energies Based on High-Spin Unrestricted Hartree–Fock Orbitals. *J. Chem. Theory Comput.* **2019**, *15* (10), 5308–5318.

(65) Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12* (4), 1760–1771.

(66) Stein, C. J.; Reiher, M. Automated Identification of Relevant Frontier Orbitals for Chemical Compounds and Processes. *Chimia (Aarau).* **2017**, *71* (4), 170–176.

(67) Stein, C. J.; Reiher, M. Measuring Multi-Configurational Character by Orbital

Entanglement. *Mol. Phys.* **2017**, *115* (17–18), 2110–2119.

(68) Stein, C. J.; Reiher, M. AutoCAS: A Program for Fully Automated Multiconfigurational Calculations. *J. Comput. Chem.* **2019**, *9999*, 1–11.

(69) White, S. R. Density Matrix Formulation for Quantum Renormalization Groups. *Phys. Rev. Lett.* **1992**, *69* (19), 2863–2866.

(70) White, S. R. Density-Matrix Algorithms for Quantum Renormalization Groups. *Phys. Rev. B* **1993**, *48* (14), 10345–10356.

(71) White, S. R.; Martin, R. L. Ab Initio Quantum Chemistry Using the Density Matrix Renormalization Group. *J. Chem. Phys.* **1999**, *110* (9), 4127–4130.

(72) Ivanic, J.; Ruedenberg, K. Identification of Deadwood in Configuration Spaces through General Direct Configuration Interaction. *Theor. Chem. Acc.* **2001**, *106* (5), 339–351.

(73) Ivanic, J.; Ruedenberg, K. Deadwood in Configuration Spaces. II. Singles + Doubles and Singles + Doubles + Triples + Quadruples Spaces. *Theor. Chem. Acc.* **2002**, *107* (4), 220–228.

(74) Evangelista, F. A. Adaptive Multiconfigurational Wave Functions. *J. Chem. Phys.* **2014**, *140* (12), 124114.

(75) Zimmerman, P. M. Incremental Full Configuration Interaction. *J. Chem. Phys.* **2017**, *146* (10), 104102.

(76) Eriksen, J. J.; Lipparini, F.; Gauss, J. Virtual Orbital Many-Body Expansions: A Possible Route towards the Full Configuration Interaction Limit. *J. Phys. Chem. Lett.* **2017**, *8* (18), 4633–4639.

(77) Eriksen, J. J.; Gauss, J. Many-Body Expanded Full Configuration Interaction. I. Weakly Correlated Regime. *J. Chem. Theory Comput.* **2018**, *14* (10), 5180–5191.

(78) Coe, J. P. Machine Learning Configuration Interaction. *J. Chem. Theory Comput.* **2018**, *14* (11), 5739–5749.

(79) 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.

(80) Nielsen, D. Tree Boosting With XGBoost Why Does XGBoost Win "Every" Machine Learning Competition? *Tree Boost. With XGBoost - Why Does XGBoost Win "Every" Mach. Learn. Compet.* **2016**.

(81) Hulburt, H. M.; Hirschfelder, J. O. Potential Energy Functions for Diatomic Molecules. *J. Chem. Phys.* **1941**, *9* (1), 61–69.

(82) Araújo, J. P.; Alves, M. D.; da Silva, R. S.; Ballester, M. Y. A Comparative Study of Analytic Representations of Potential Energy Curves for O2, N2, and SO in Their Ground Electronic States. *J. Mol. Model.* **2019**.

(83) Sanchez de Meras, A.; Lepetit, M.-B.; Malrieu, J.-P. Discontinuity of Valence CASSCF Wave Functions around Weakly Avoided Crossing between Valence Configurations. *Chem. Phys. Lett.* **1990**, *172* (2), 163–168.

(84) Shaik, S.; Danovich, D.; Wu, W.; Hiberty, P. C. Charge-Shift Bonding and Its Manifestations in Chemistry. *Nat. Chem.* **2009**, *1* (6), 443–449.

(85) Shaik, S.; Danovich, D.; Galbraith, J. M.; Braida, B.; Wu, W.; Hiberty, P. C. Charge-Shift Bonding: A New and Unique Form of Bonding. *Angew. Chemie Int. Ed.* **2019**.

Supporting Information

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

Authors: 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.

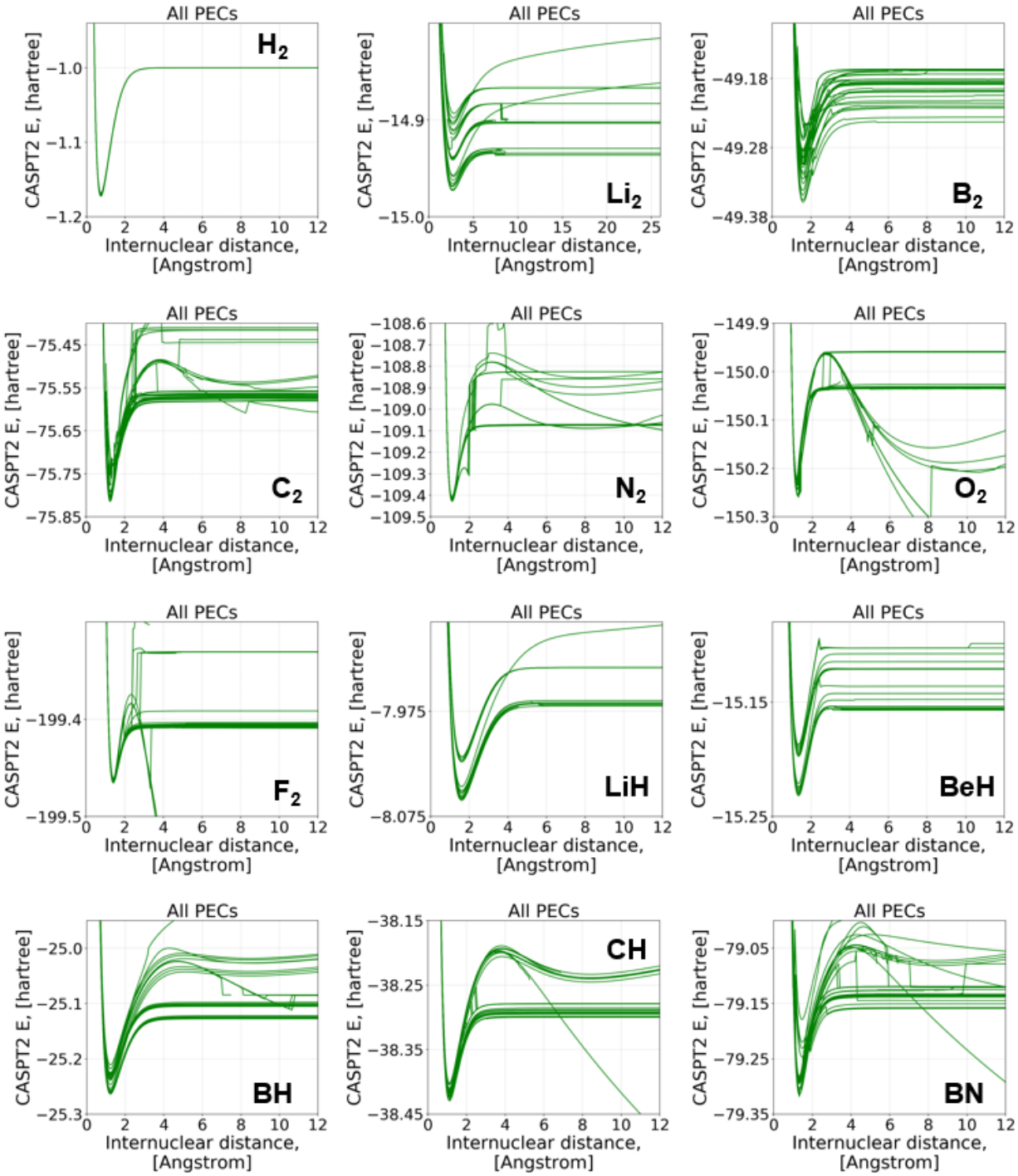

Figure S1-1. All potential energy curves using CASSCF/CASPT2 energies for homonuclear diatomic molecules, hydrides, and BN.

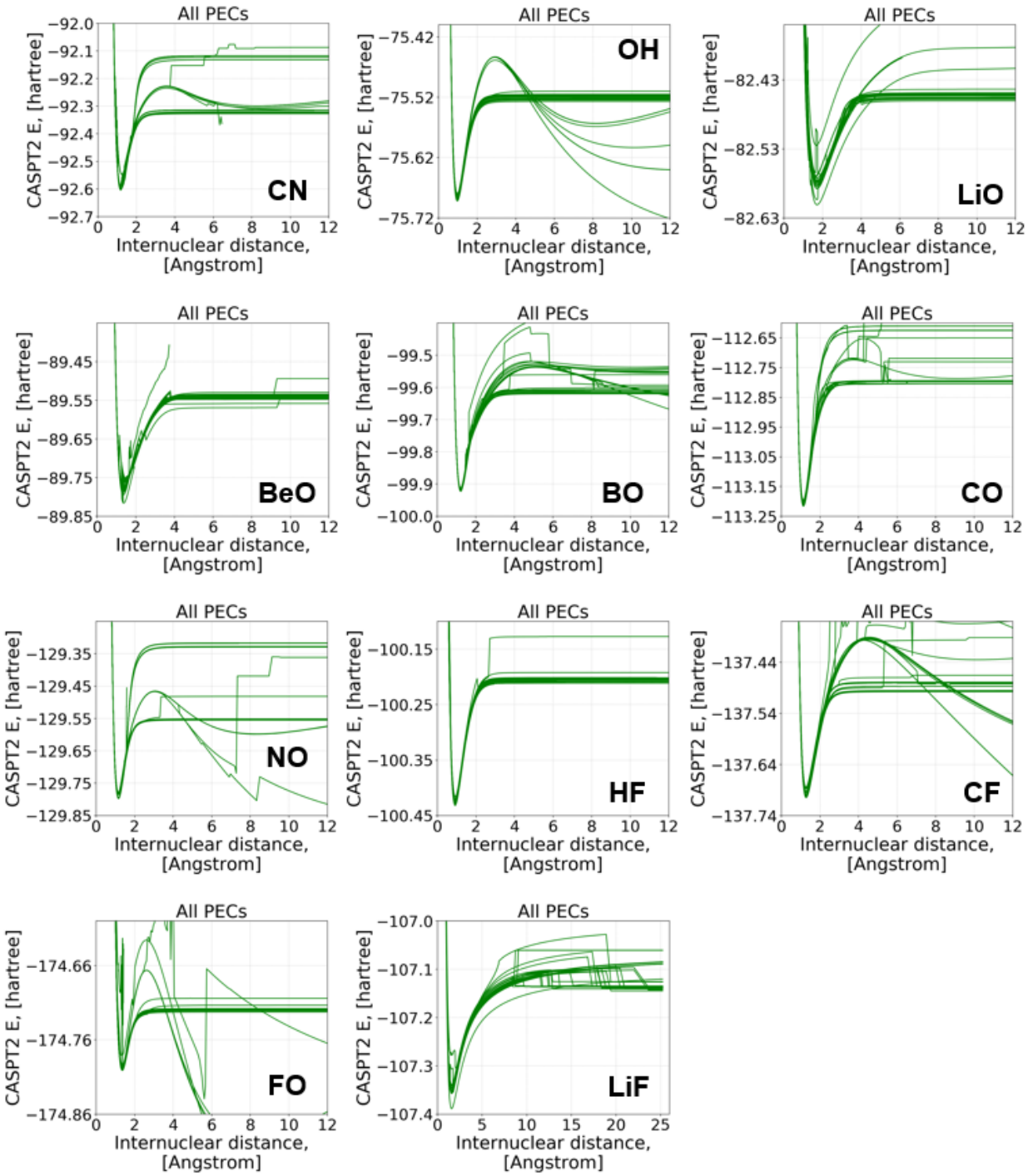

Figure S1-2. All potential energy curves using CASSCF/CASPT2 energies for oxides, fluorides, and CN.

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.

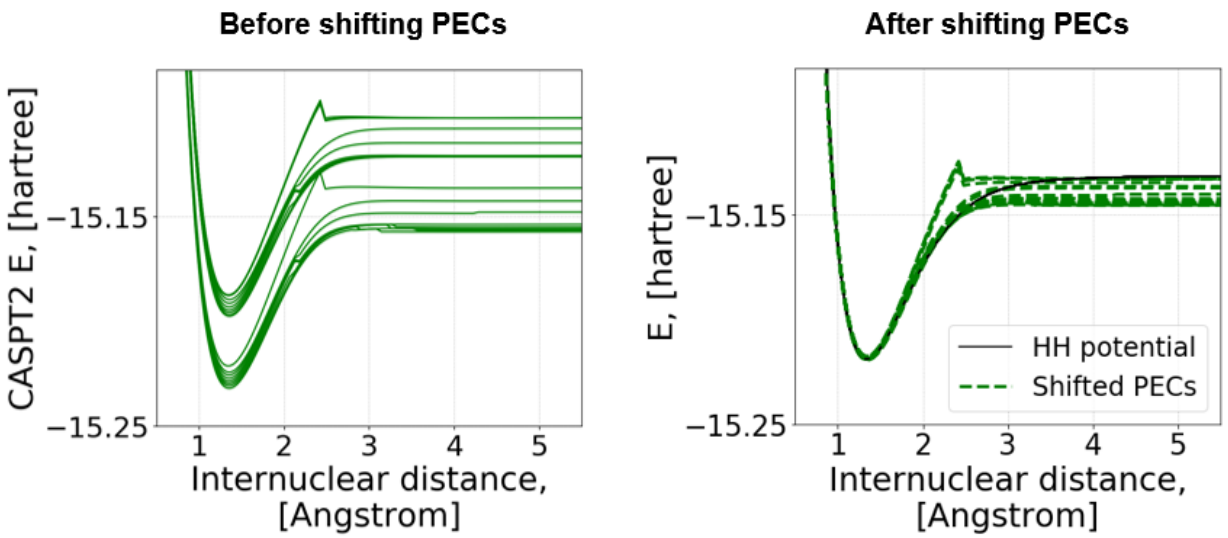

Figure S2. Comparison of original potential energy curves and shifted curves for BeH

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.

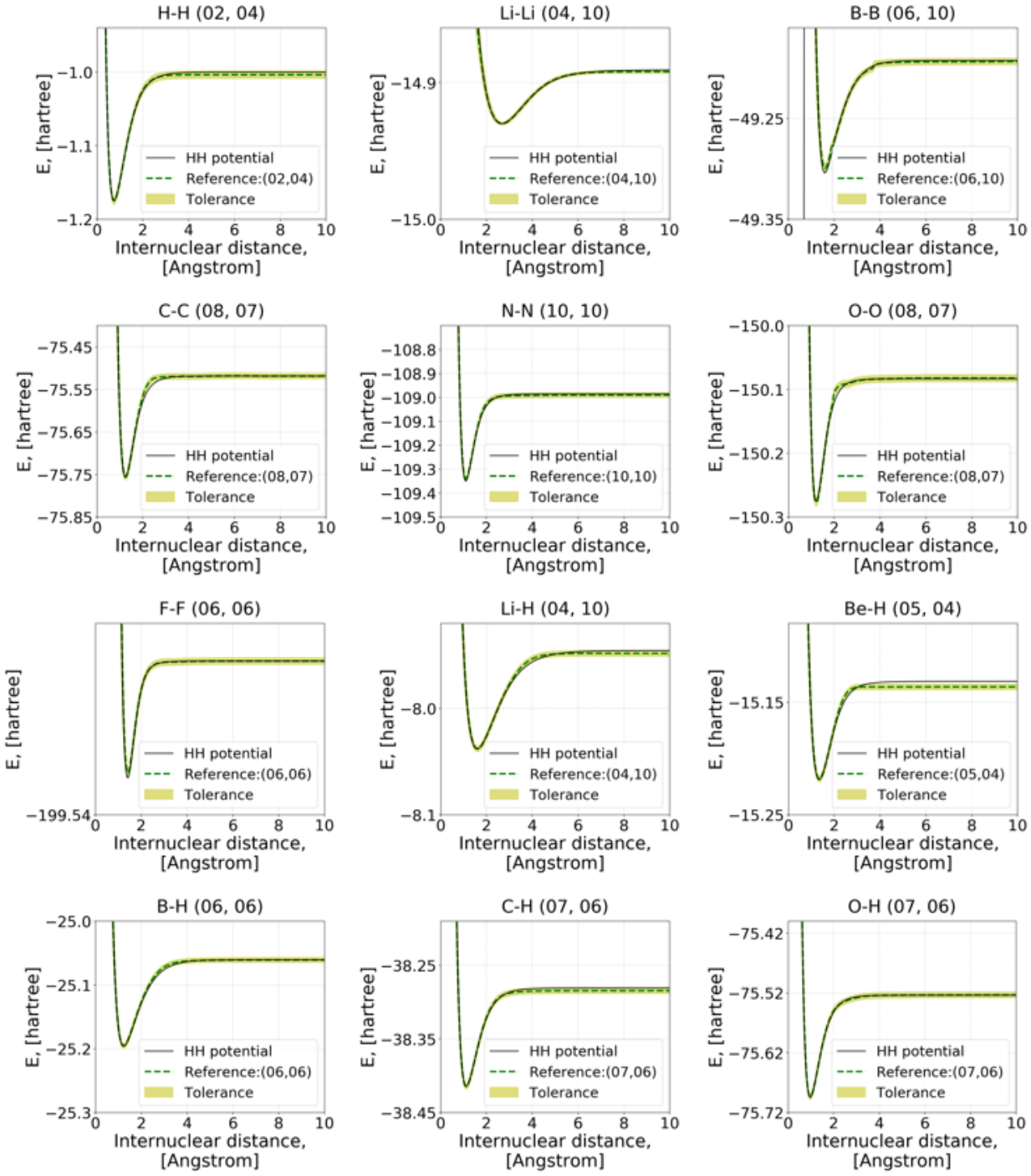

Figure S3-1. Reference potential energy curves for diatomic molecules that are the most similar to the corresponding Hulburt-Hirschfelder potential energy curve

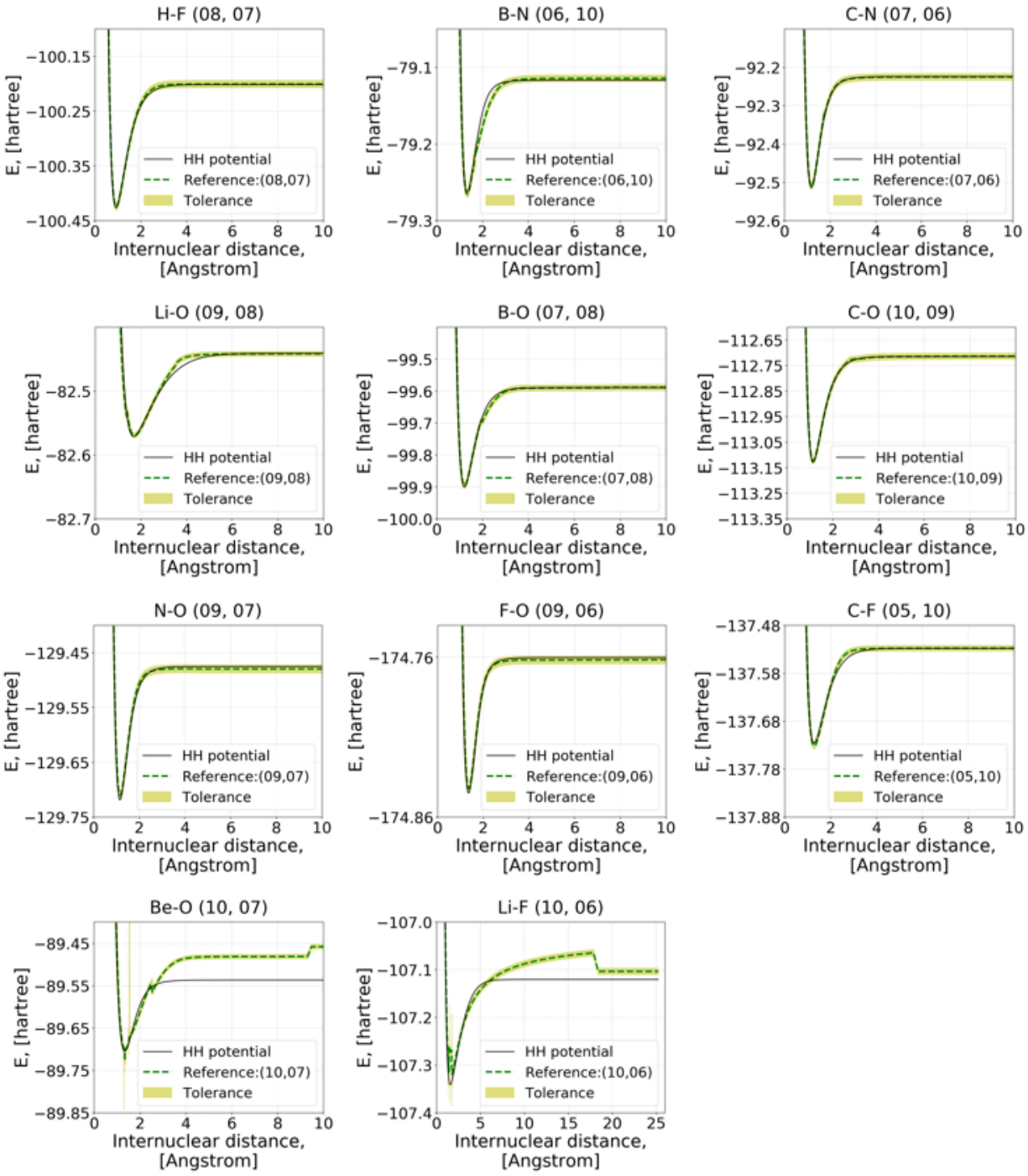

Figure S3-2. Reference potential energy curves for diatomic molecules that are the most similar to the corresponding Hulburt-Hirschfelder potential energy curve

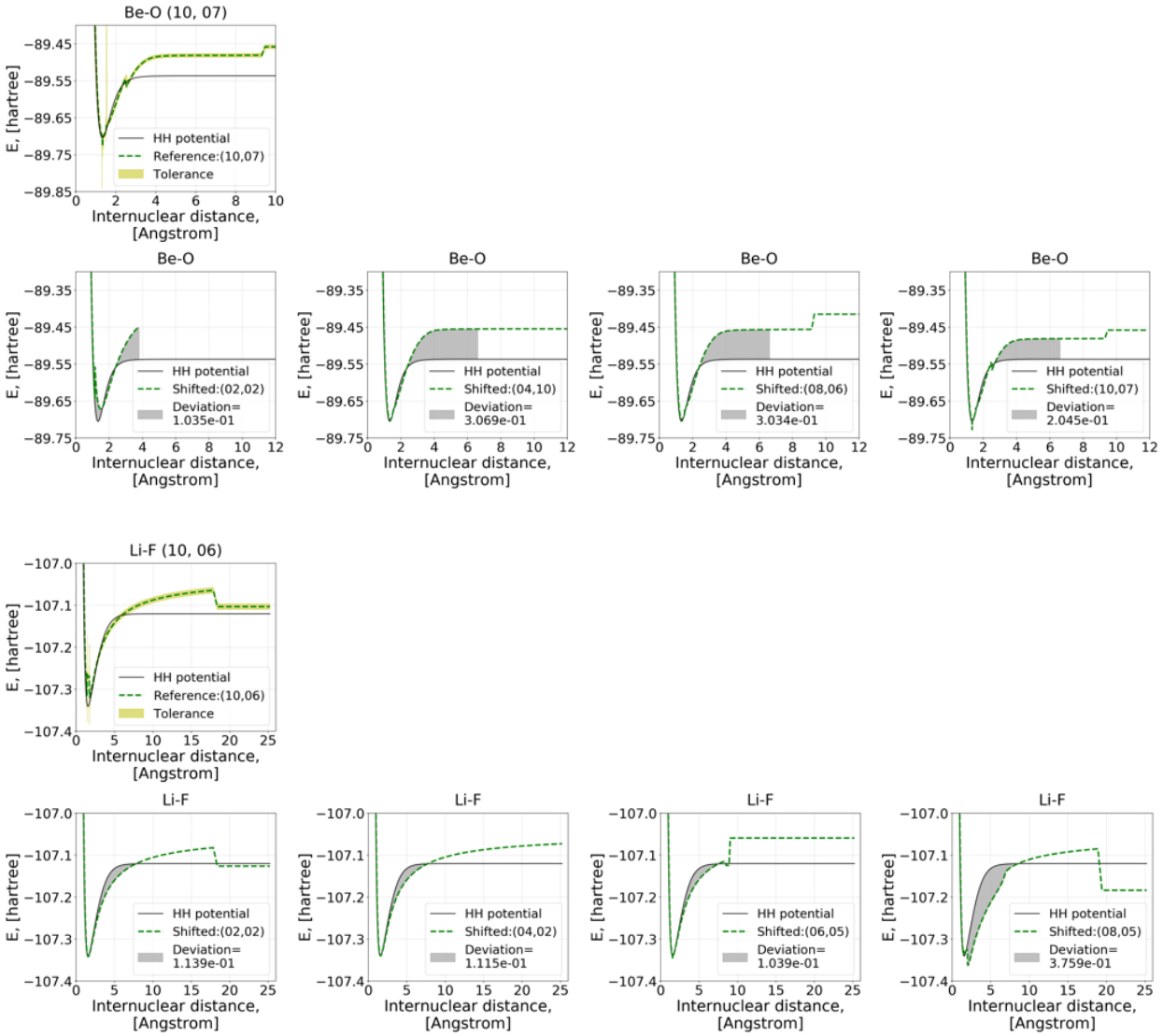

Figure S4. Representative potential energy curves for BeO and LiF. There is no similar potential energy curve compared to the corresponding Hulburt-Hirschfelder potential energy curve.

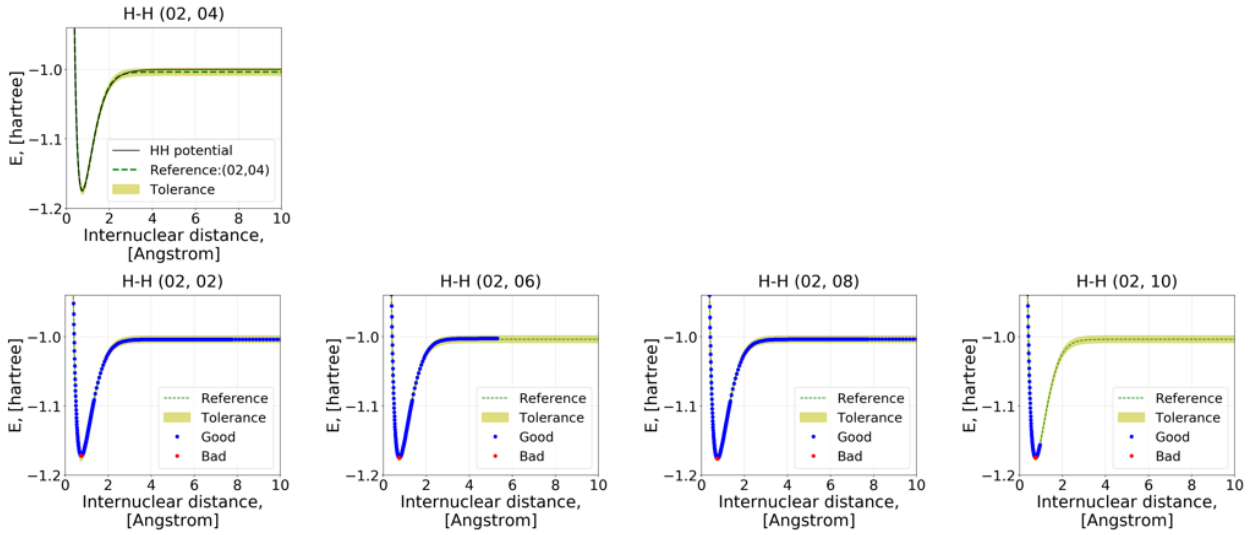

Figure S5. Representative potential energy curves for H_2 that show imbalanced data points. A portion of the bad-labeled data points are extremely small, and the data points only exist near the equilibrium bond length.

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.

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

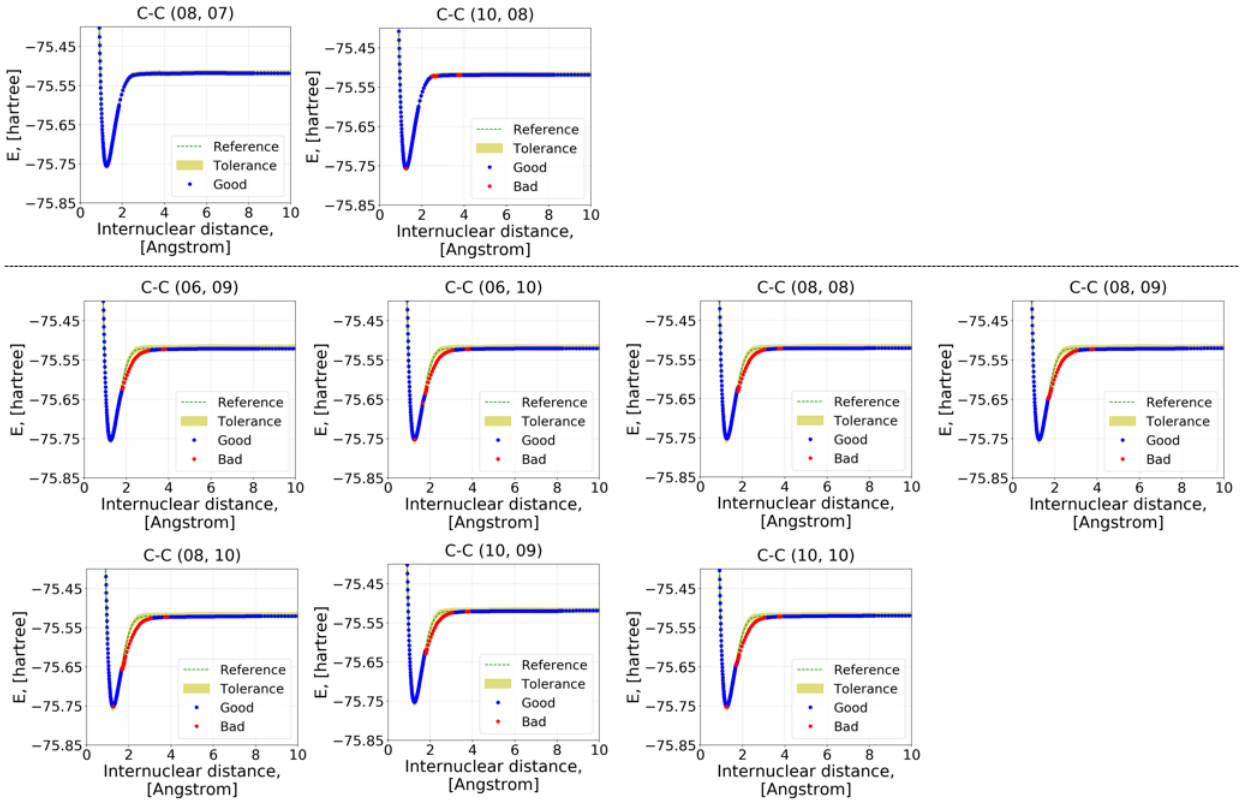

Figure S6. Confusing potential energy curves for C_2. Both good and bad PECs have similar errors in dissociative properties, but only can be distinguished only by different curvatures compared to the reference PEC.

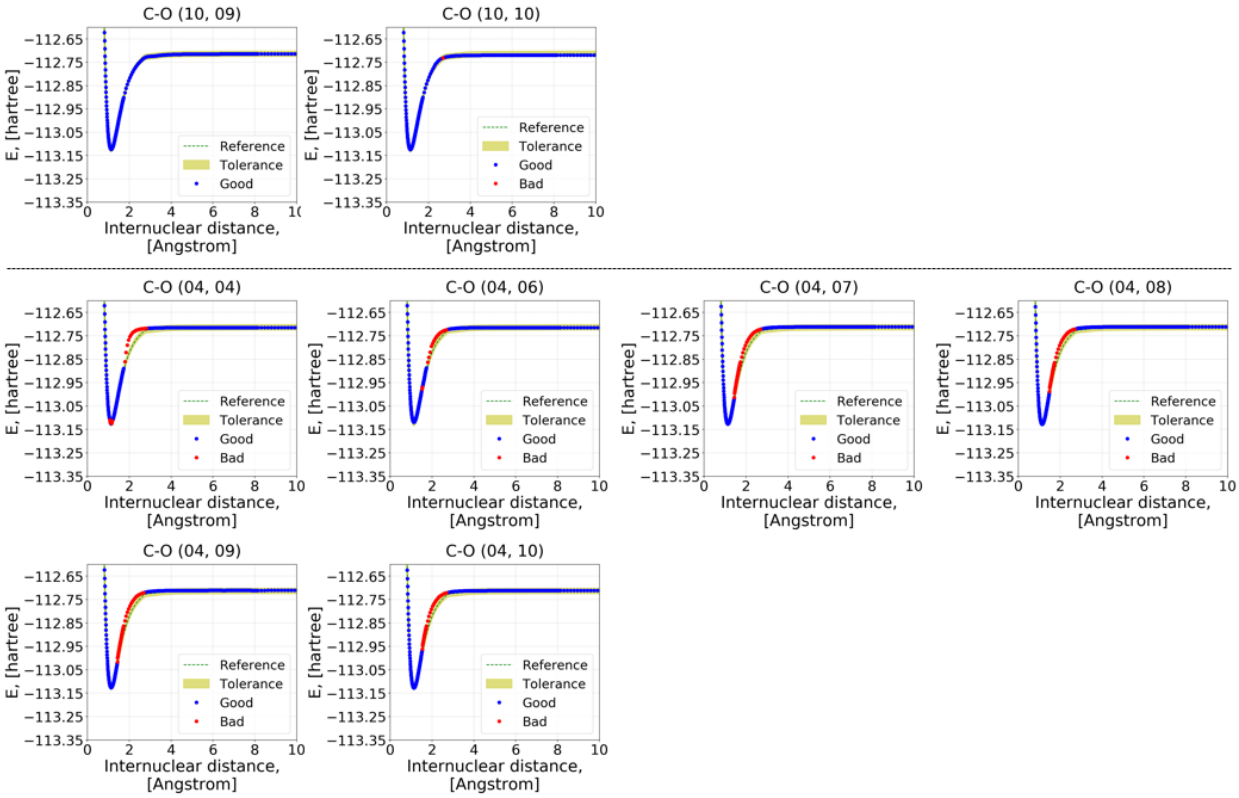

Figure S7. Confusing potential energy curves for CO. Both good and bad PECs have similar errors in dissociative properties, but only can be distinguished only by different curvatures compared to the reference PEC.

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

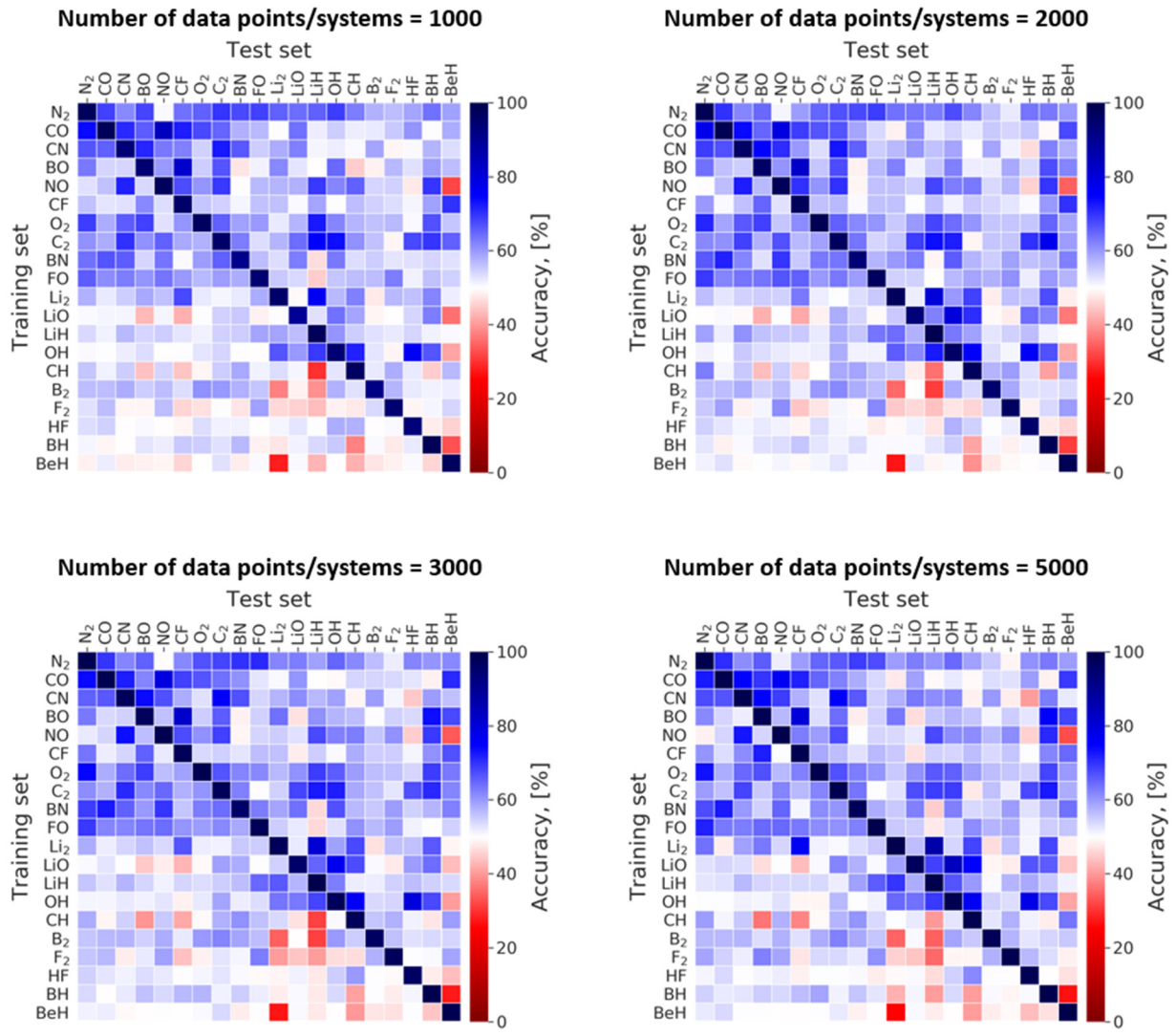

Figure S8. Comparison of ML model prediction performances when the models are trained on the same numbers of training data points (i.e., 1000, 2000, 3000, 5000) per a diatomic molecule and then predicted on all the diatomic systems we investigated.

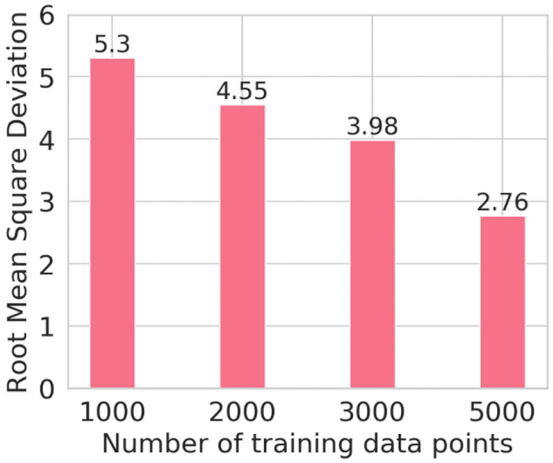

Figure S9. Average root-mean-square deviation between heat maps generated using different number of training data points and the heat map produced with all available training data points.

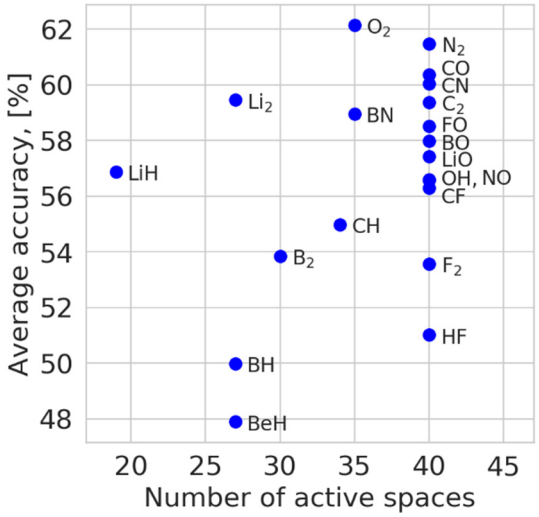

Figure S10. Average prediction accuracy of ML models trained on single diatomic system over other 19 diatomic systems versus the number of possible active spaces limited to the maximum size of 10.

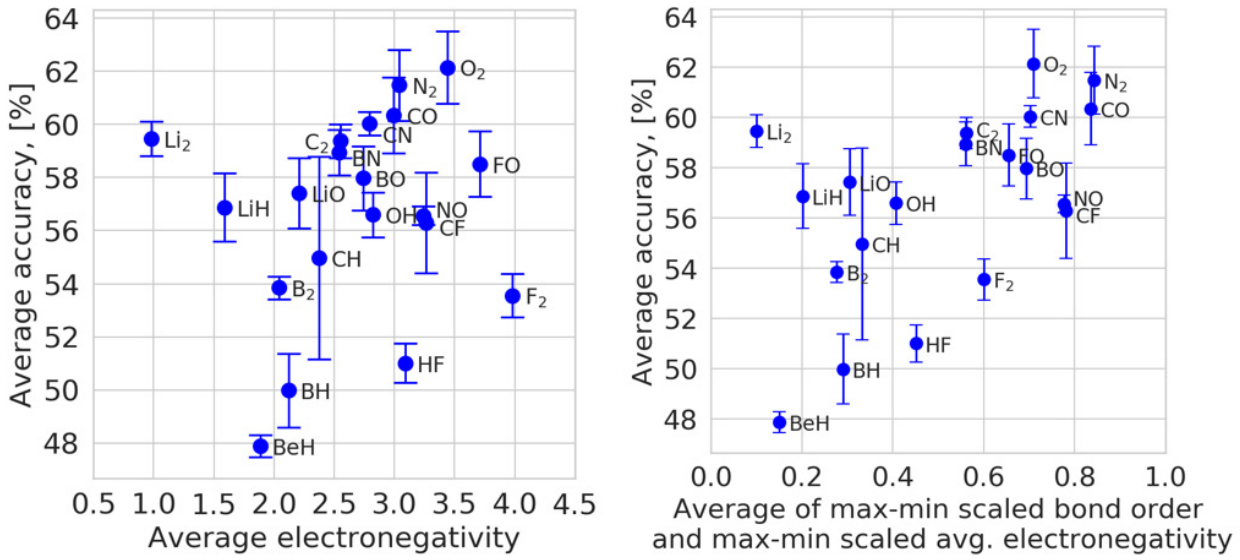

Figure S11. Average prediction accuracy of ML model trained on single diatomic system over other 19 diatomic systems versus (a) average electronegativity and (b) new metric obtained by averaging max-min rescaled bond order and average electronegativity.

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

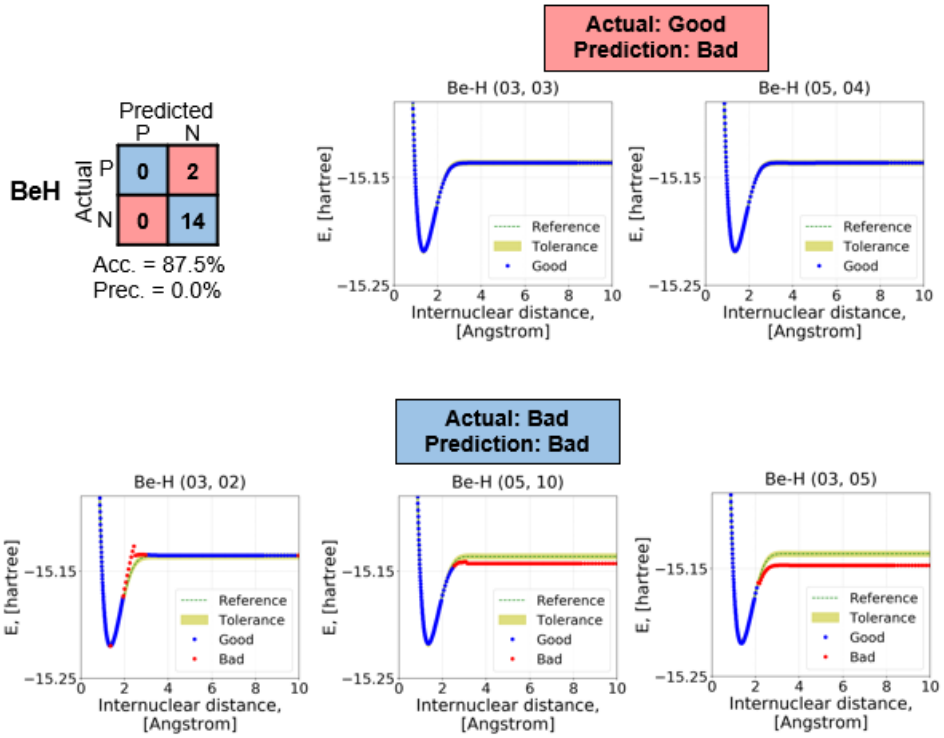

Figure S12. Representative potential energy curves for each case of the confusion matrixes for BeH.

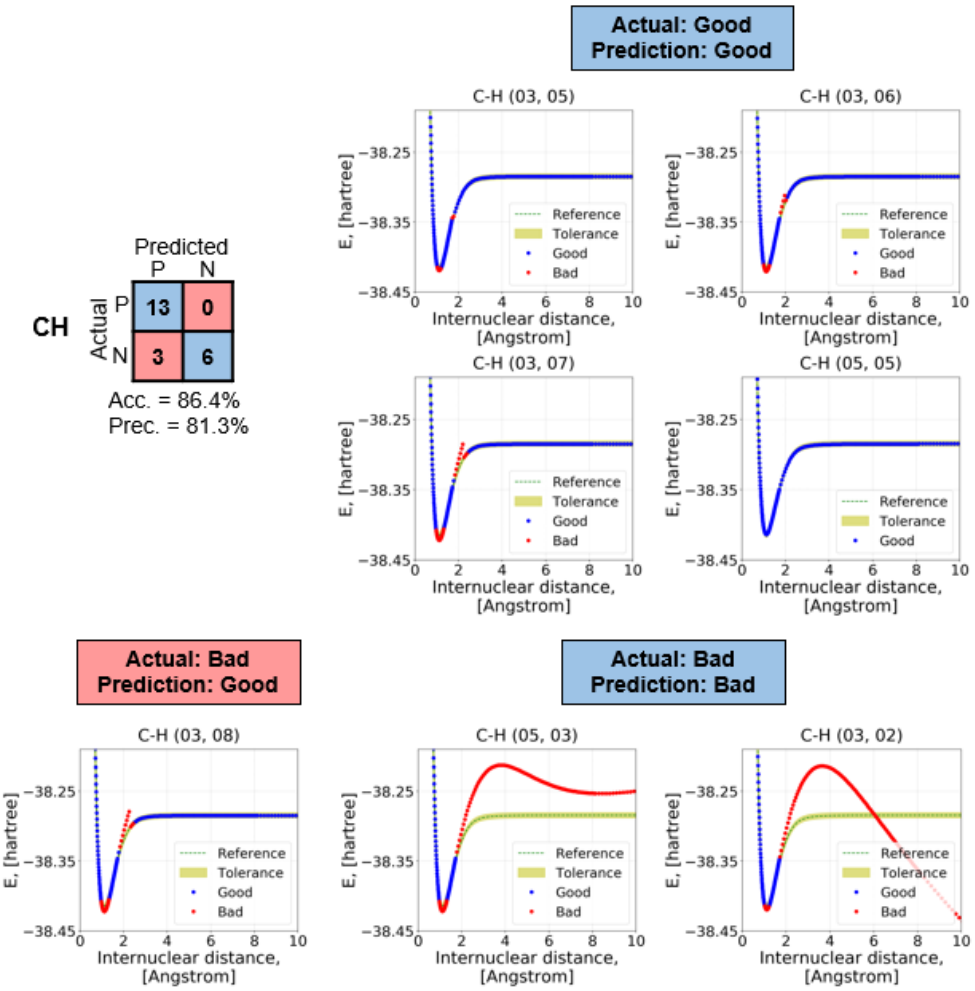

Figure S13. Representative potential energy curves for each case of the confusion matrixes for CH.

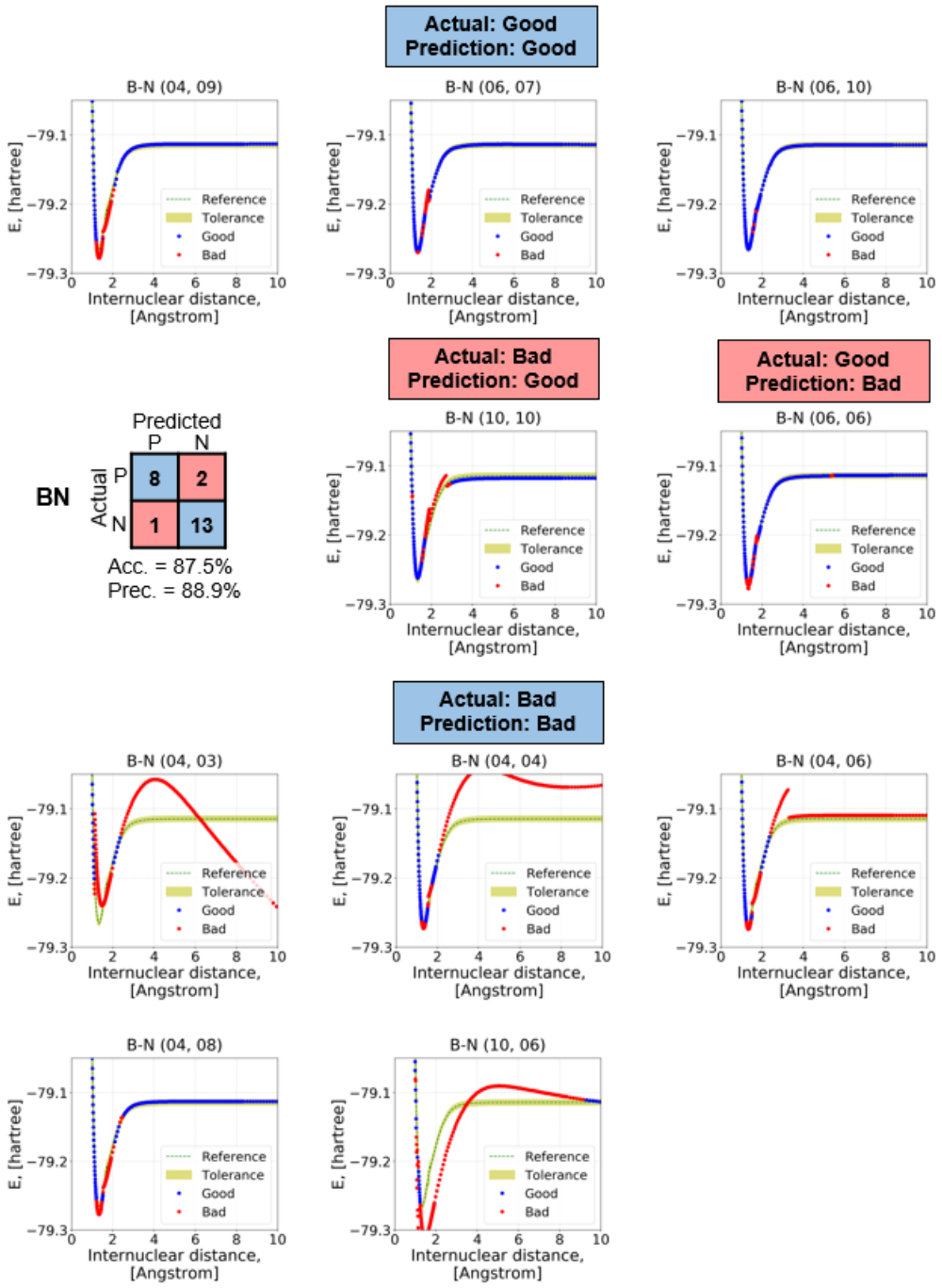

Figure S14. Representative potential energy curves for each case of the confusion matrixes for BN.

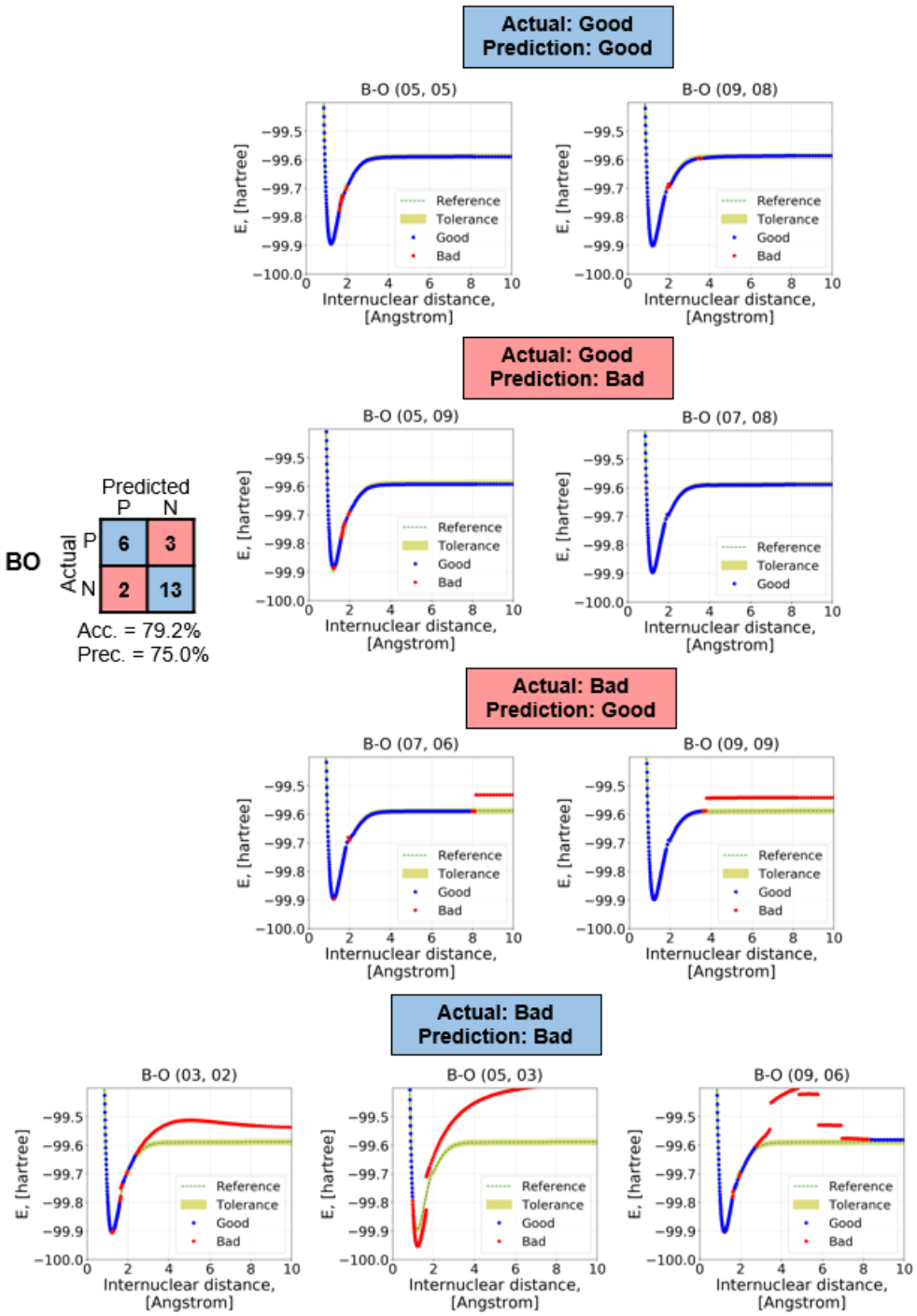

Figure S15. Representative potential energy curves for each case of the confusion matrixes for BO.

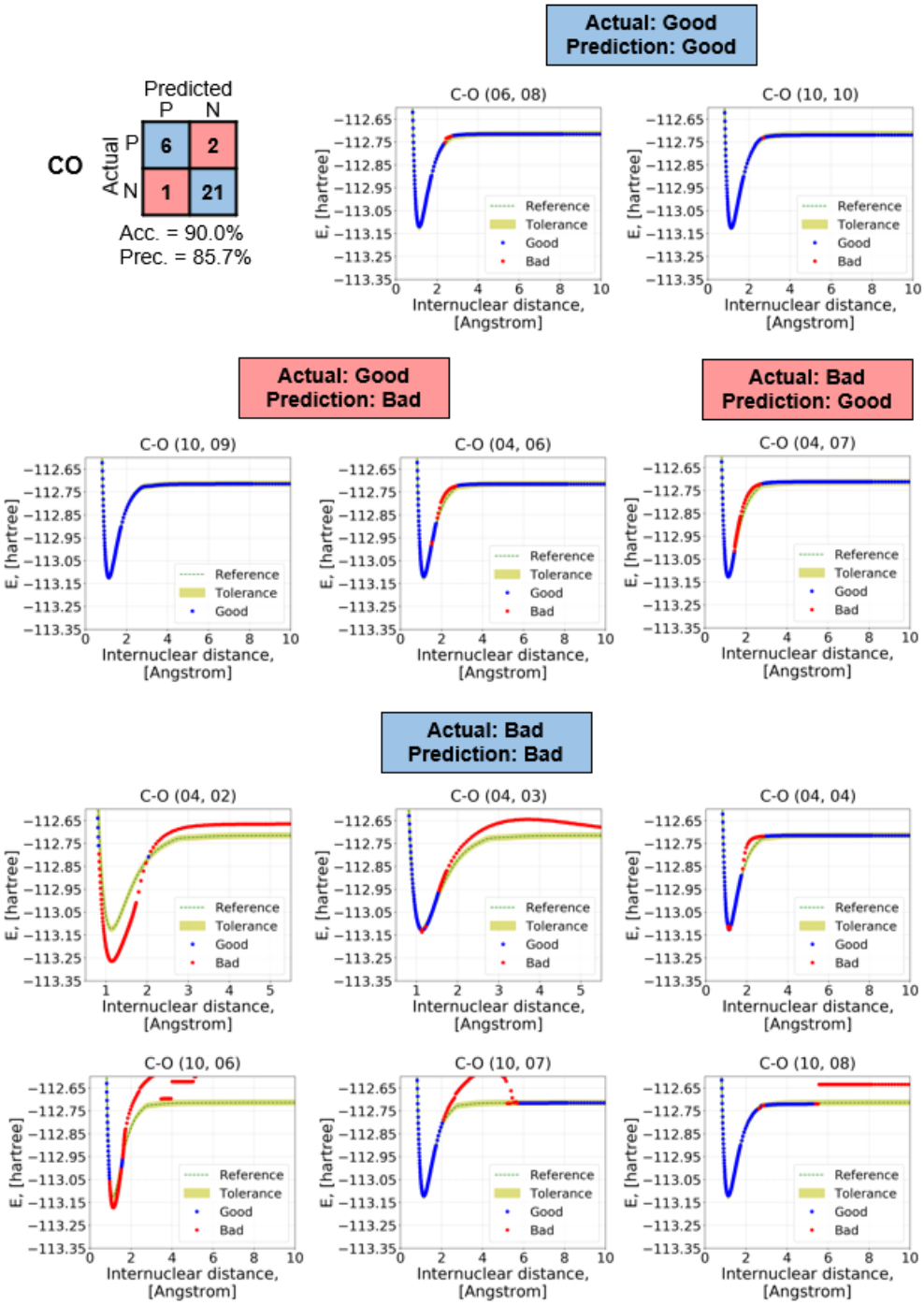

Figure S16. Representative potential energy curves for each case of the confusion matrixes for CO.

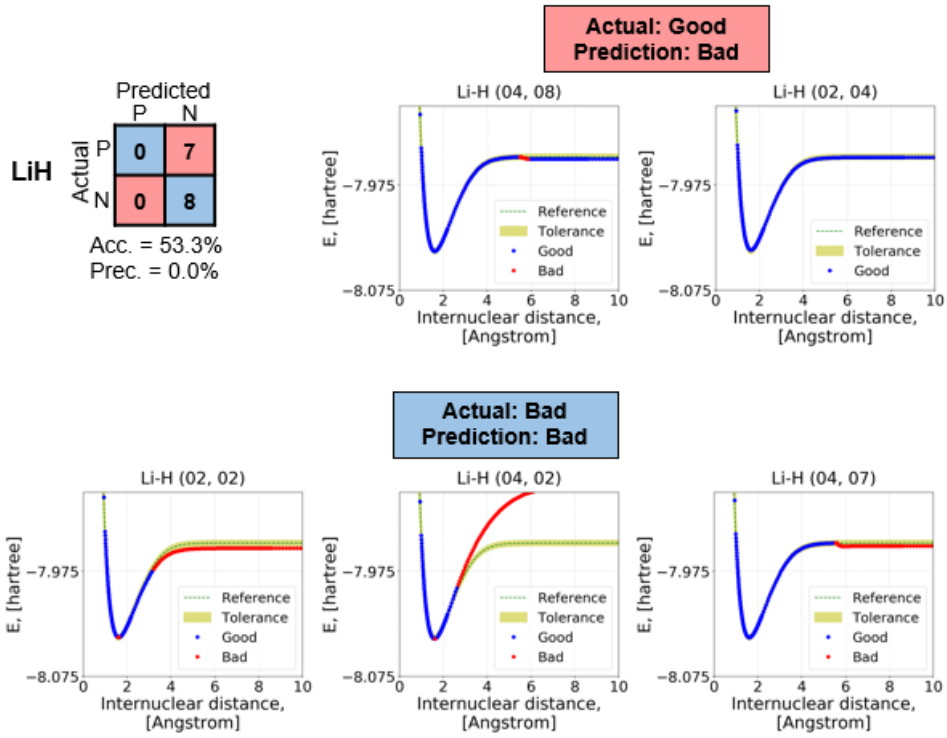

Figure S17. Representative potential energy curves for each case of the confusion matrixes for LiH.

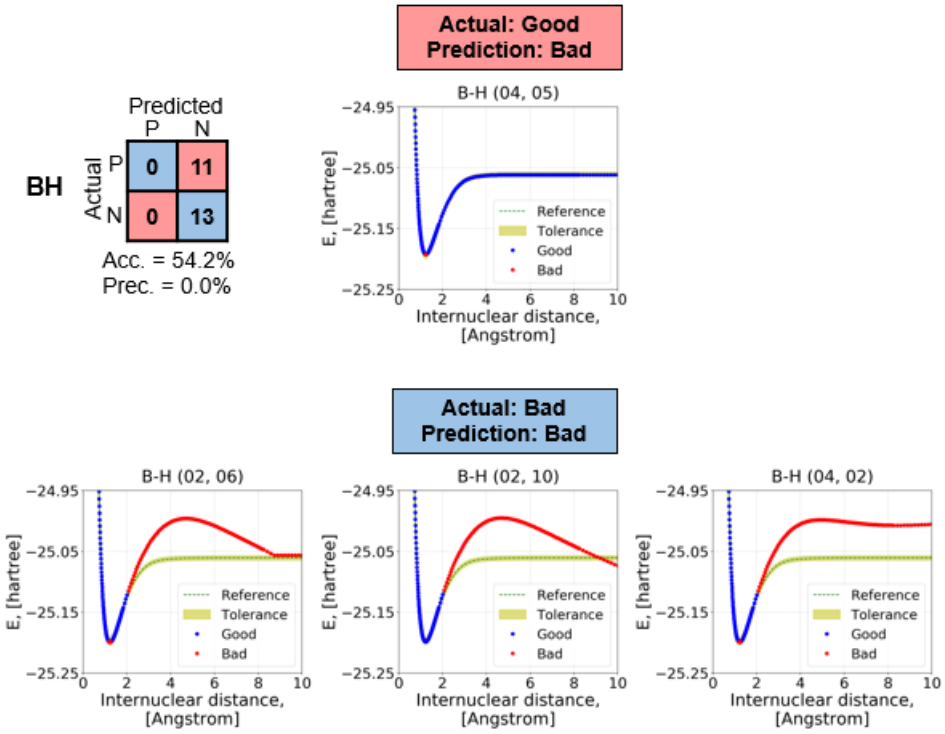

Figure S18. Representative potential energy curves for each case of the confusion matrixes for BH.

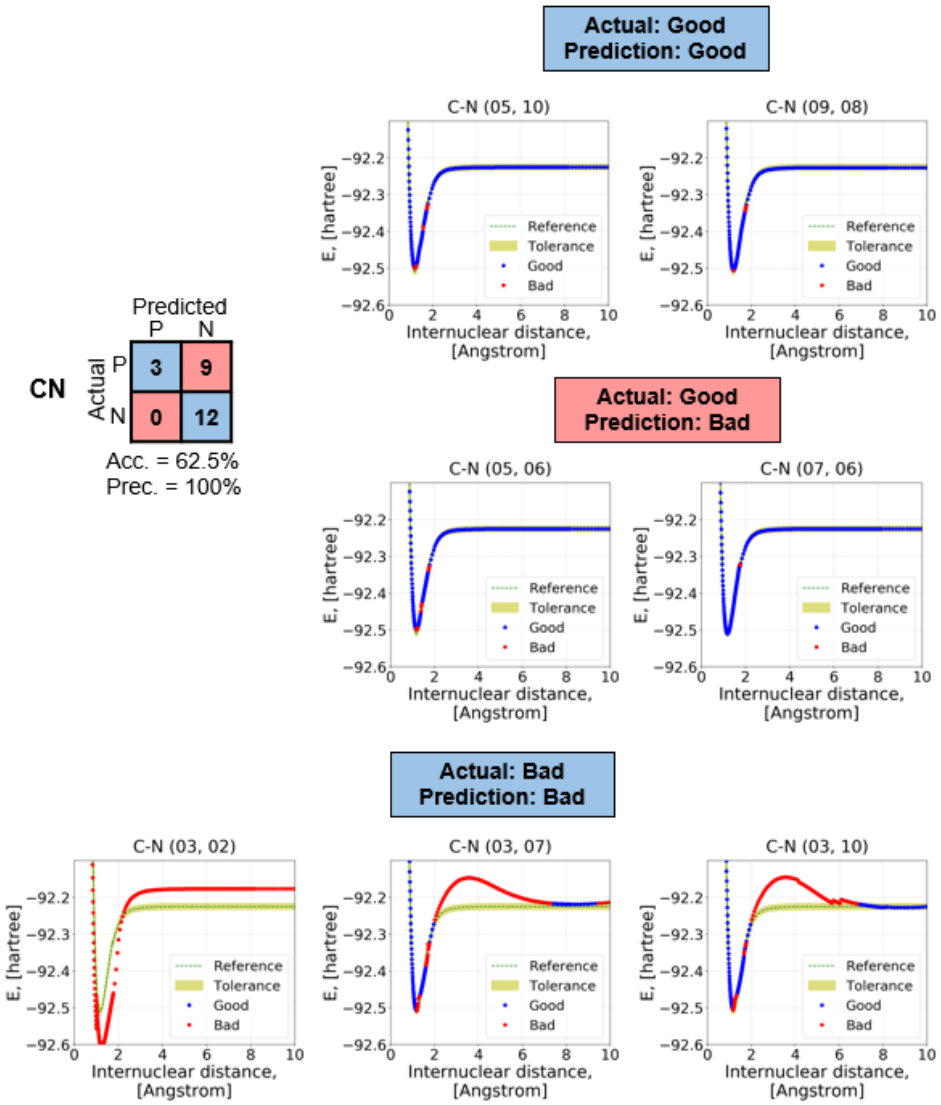

Figure S19. Representative potential energy curves for each case of the confusion matrixes for CN.

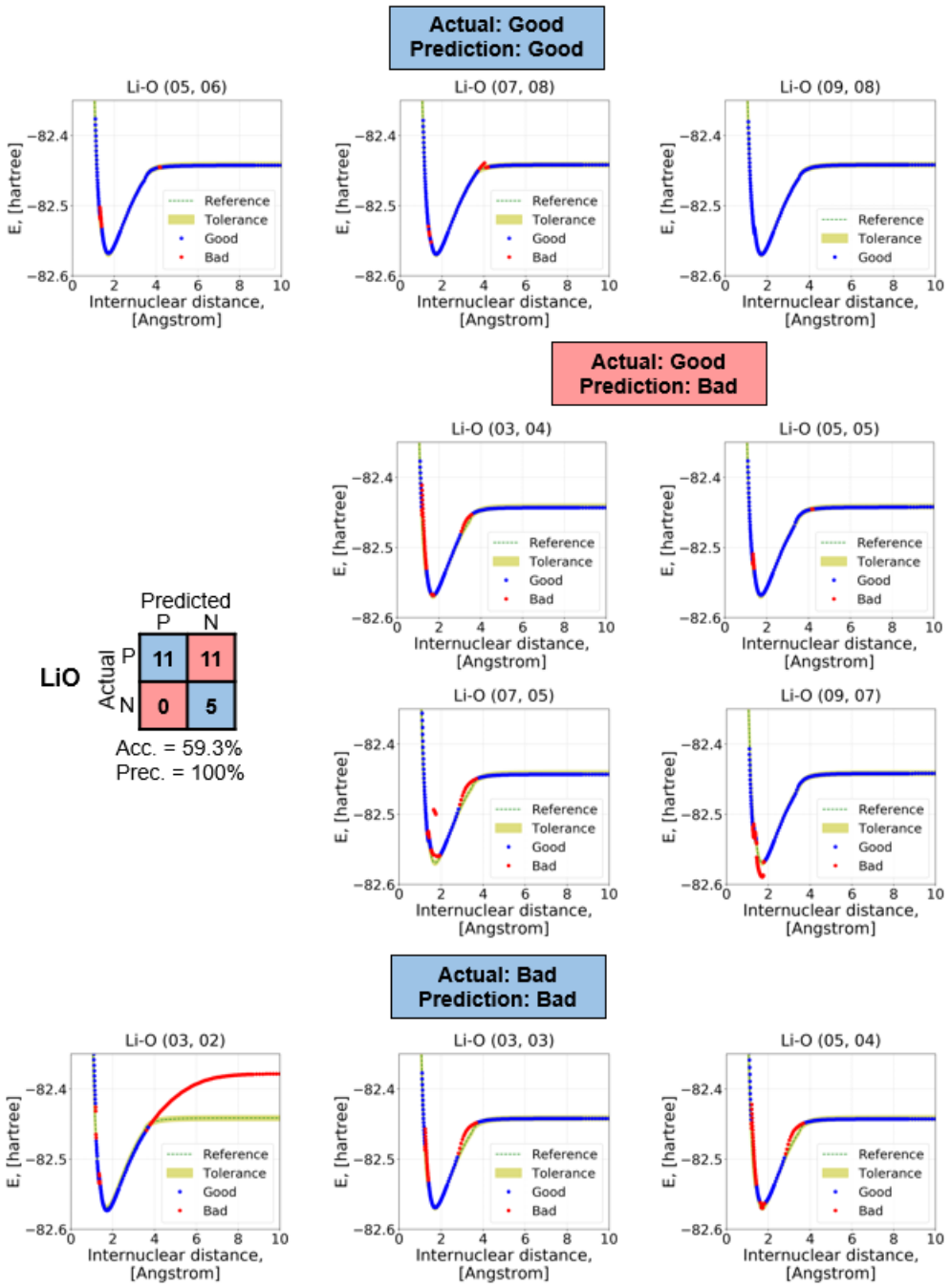

Figure S20. Representative potential energy curves for each case of the confusion matrixes for LiO.

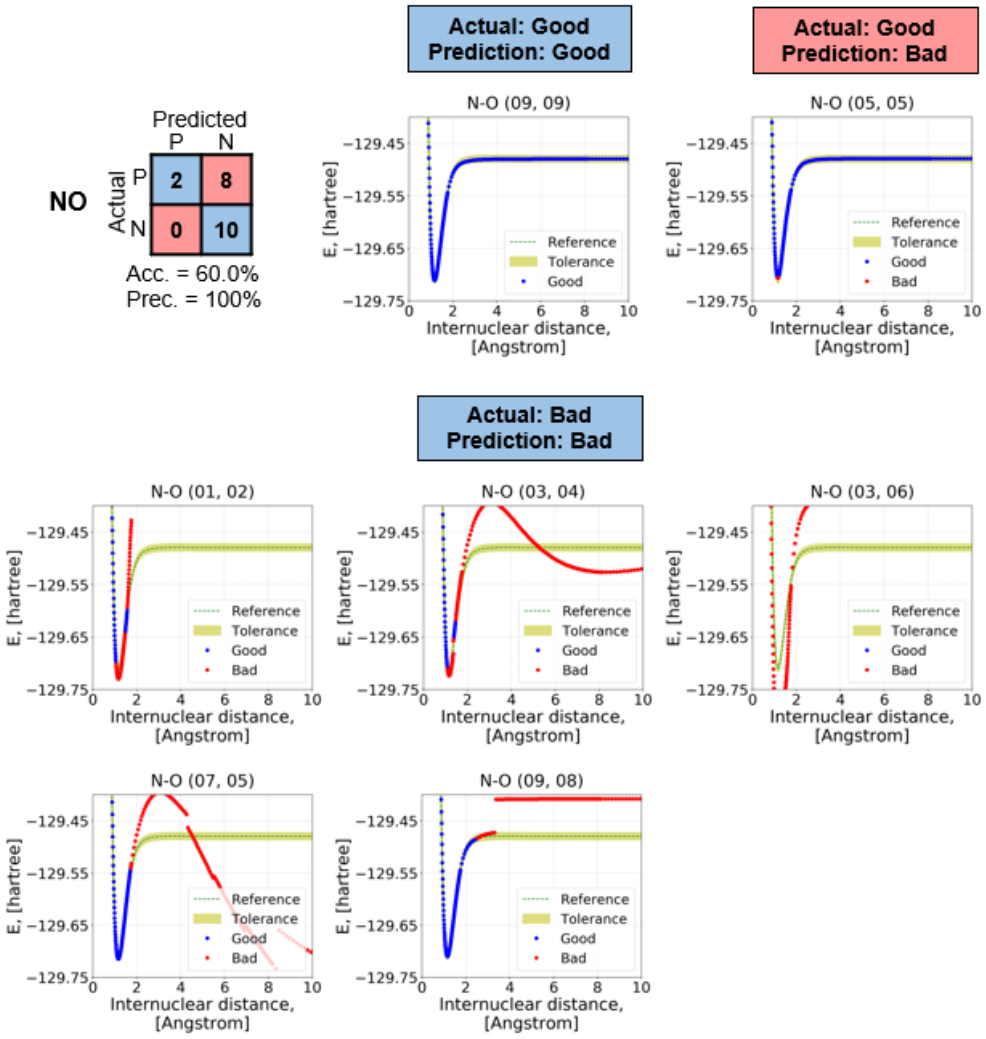

Figure S21. Representative potential energy curves for each case of the confusion matrixes for NO.

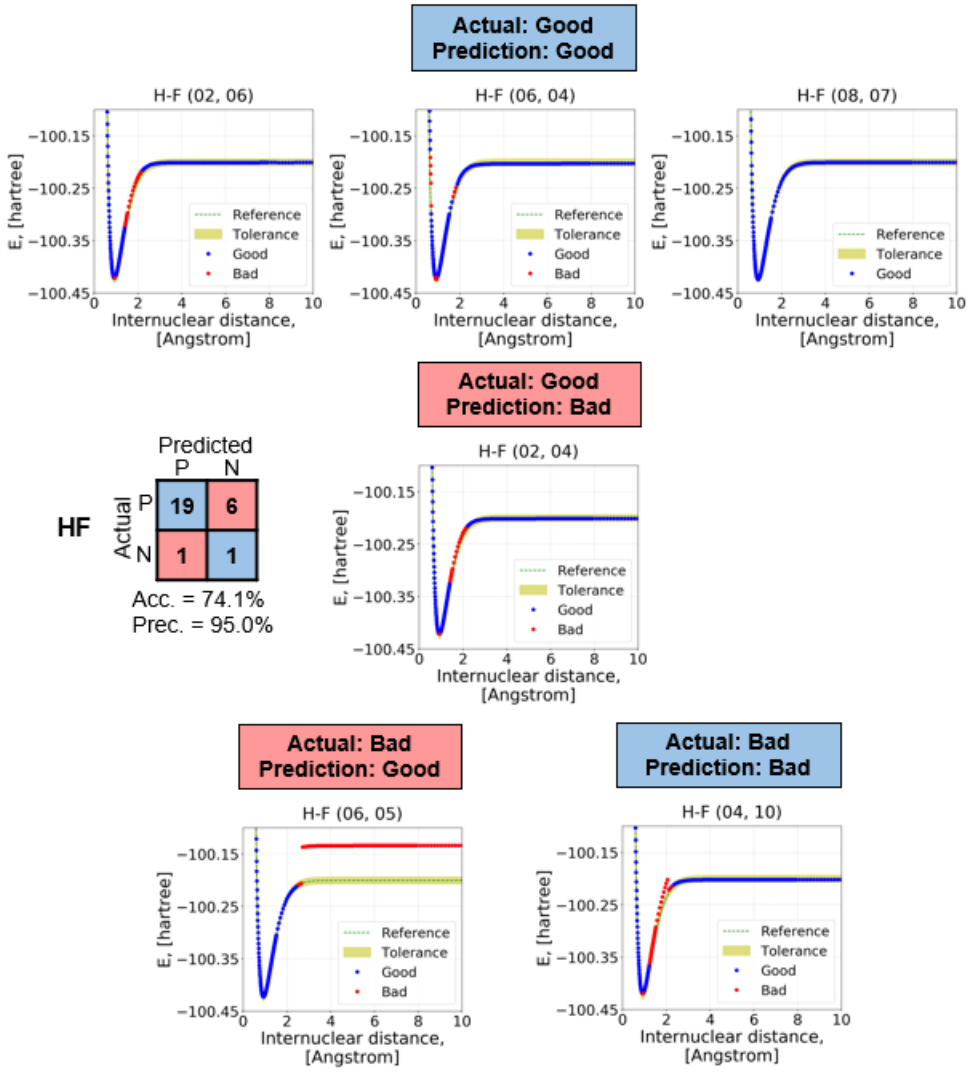

Figure S22. Representative potential energy curves for each case of the confusion matrixes for HF.

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), (7, 6)	70, 210, 210
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* (17), 3973–3976.

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

(3) Editor: Russell D. Johnson III. Experimental data for O2 (Oxygen diatomic) https://cccbdb.nist.gov/exp2x.asp?casno=7782447&charge=0.

(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* (14), 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* (2), 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* (5), 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* (5), 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* (17), 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* (23), 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* (9), 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* (1–3), 142–149.

(13) Forsberg, N.; Malmqvist, P. Å. Multiconfiguration Perturbation Theory with Imaginary Level Shift. *Chem. Phys. Lett.* **1997**, *274* (1–3), 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* (4), 626–635.

(15) Hulburt, H. M.; Hirschfelder, J. O. Potential Energy Functions for Diatomic Molecules. *J. Chem. Phys.* **1941**, *9* (1), 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 O2, N2, and SO in Their Ground Electronic States. *J. Mol. Model.* **2019**.

(17) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0--Fundamental Algorithms for Scientific Computing in Python. **2019**.

(18) *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* (15), 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* (1), 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? *Tree Boost. With XGBoost - Why Does XGBoost Win "Every" Mach. Learn. Compet.* **2016**.

(23) No Title https://www.kaggle.com/competitions.

(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* (1), 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* (2), 145–151.