# Supplementary Information: Δ-Machine Learning toward CCSD Accuracy for Homohalogenated Borane–Phosphine Adducts: Screening Low-Energy Structures from DFT and MP2 Libraries

Okan Köksal[1]

[1] *Center for Molecular Modeling, Ghent University, Technologiepark 903, BE-9052 Zwijnaarde, Belgium*

(Dated: May 5, 2026)

## I. SUPPLEMENTARY COMPUTATIONAL DETAILS AND MACHINE LEARNING VALIDATION

This Supplementary Information contains additional computational and machine learning details related to the study of the homohalogenated borane–phosphine adducts $F_3B{-}PF_3$, $Cl_3B{-}PCl_3$, and $Br_3B{-}PBr_3$. It covers complementary validation of the Δ-machine learning models, comparisons between descriptor-based and SOAP-based representations, principal-component analyses of the feature spaces, optimized fragment reference geometries at different electronic-structure levels, and CCSD(T) refinements for ML-selected low-energy candidates.

### A. Optimized fragment reference geometries

TABLE S1. Optimized fragment geometries used for fragment reference energies at each level. Reported are averages over the three symmetry-related bonds and angles: $\langle r \rangle$ denotes the mean B–$X$ (or P–$X$) bond length, and $\langle \theta \rangle$ the mean $X$–B–$X$ (or $X$–P–$X$) bond angle (deg).

| Fragment | Level | $\langle r \rangle$ (Å) | $\langle \theta \rangle$ (deg) |
|---|---|---|---|
| $BF_3$ | MP2 | 1.31706 | 120.000 |
| $BF_3$ | DFT-D3 | 1.31657 | 120.000 |
| $BF_3$ | CCSD | 1.31165 | 120.000 |
| $BF_3$ | CCSD(T) | 1.31533 | 120.000 |
| $BCl_3$ | MP2 | 1.73817 | 120.000 |
| $BCl_3$ | DFT-D3 | 1.74876 | 120.000 |
| $BCl_3$ | CCSD | 1.74212 | 120.000 |
| $BCl_3$ | CCSD(T) | 1.74454 | 120.000 |
| $BBr_3$ | MP2 | 1.89949 | 120.000 |
| $BBr_3$ | DFT-D3 | 1.91435 | 120.000 |
| $BBr_3$ | CCSD | 1.90538 | 120.000 |
| $BBr_3$ | CCSD(T) | 1.90803 | 120.000 |
| $PF_3$ | MP2 | 1.57406 | 97.606 |
| $PF_3$ | DFT-D3 | 1.58211 | 97.639 |
| $PF_3$ | CCSD | 1.56566 | 97.419 |
| $PF_3$ | CCSD(T) | 1.57226 | 97.433 |
| $PCl_3$ | MP2 | 2.04612 | 100.115 |
| $PCl_3$ | DFT-D3 | 2.07538 | 101.027 |
| $PCl_3$ | CCSD | 2.05131 | 100.186 |
| $PCl_3$ | CCSD(T) | 2.05779 | 100.167 |
| $PBr_3$ | MP2 | 2.22367 | 100.913 |
| $PBr_3$ | DFT-D3 | 2.25917 | 102.199 |
| $PBr_3$ | CCSD | 2.23122 | 101.030 |
| $PBr_3$ | CCSD(T) | 2.23851 | 101.032 |

Table S1 compiles the optimized geometries of the isolated fragments used to compute fragment reference energies at each electronic-structure level. For each fragment, all structures were optimized independently at the corresponding level of theory using the aug-cc-pVTZ basis set. Reported bond lengths and angles represent averages over the three symmetry-equivalent $X$ substituents, i.e., $\langle r \rangle$ corresponds to the mean B–$X$ or P–$X$ bond length and $\langle \theta \rangle$ to

the mean $X$–B–$X$ or $X$–P–$X$ bond angle. For the isolated fragments, geometry optimizations at the CCSD/aug-cc-pVTZ and CCSD(T)/aug-cc-pVTZ levels were carried out in GAMESS using numerical gradients. In contrast, full supermolecule optimizations at the CCSD(T)/aug-cc-pVTZ level were not pursued because of their substantially higher computational cost, i.e., for the donor–acceptor adducts, CCSD(T) was instead used for targeted single-point refinements on selected low-energy structures.

### B. Descriptor definitions and SOAP settings

For the descriptor-based Δ-ML models, each structure was represented by a compact handcrafted feature vector containing selected Cartesian coordinate components, key intermolecular and intramolecular distances, bond angles, and the corresponding low-level energy $E_{\mathrm{low}}$. The descriptor set was chosen to capture both the intermolecular donor–acceptor arrangement and the most relevant intrafragment distortions induced by the stochastic Cartesian perturbations. The same handcrafted descriptor definition was used for the SVR, GBR, and GPR models discussed in the main text. Table S2 collects representative values for the chemically most transparent subset of these descriptors, whereas the Cartesian components are part of the full input vector used in the model but are omitted from the table for readability.

TABLE S2. Representative values of selected handcrafted descriptors for three example configurations of $Br_3B{-}PBr_3$ from the matched CCSD-labeled subset. The full descriptor vector used in the descriptor-based models additionally contains selected Cartesian coordinate components of the distorted geometry. Here we report the chemically most transparent subset of descriptors to illustrate their scale and variation. Energies are given in a.u., distances in Å, and angles in degrees.

| Structure | $E_{\mathrm{MP2}}$ | $E_{\mathrm{CCSD}}$ | $R_{\mathrm{B\cdots P}}$ | $r_1^{\mathrm{BBr}}$ | $r_2^{\mathrm{BBr}}$ | $r_3^{\mathrm{BBr}}$ | $r_1^{\mathrm{PBr}}$ | $r_2^{\mathrm{PBr}}$ | $r_3^{\mathrm{PBr}}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | −15801.653007 | −15801.715614 | 1.984 | 2.165 | 2.188 | 2.153 | 1.919 | 2.010 | 2.008 | 104.49 | 104.83 | 105.48 | 115.30 | 115.83 | 113.06 |
| 2 | −15801.648238 | −15801.711690 | 1.971 | 2.189 | 2.198 | 2.186 | 2.066 | 2.058 | 1.892 | 105.79 | 104.01 | 108.27 | 111.23 | 113.50 | 116.99 |
| 3 | −15801.651584 | −15801.713902 | 2.015 | 2.182 | 2.139 | 2.161 | 1.902 | 2.027 | 1.967 | 106.79 | 107.31 | 103.31 | 116.45 | 115.62 | 113.19 |

As an alternative structural representation, SOAP descriptors were generated with the DScribe package [1] using atomic species present in each system. For the brominated system, e.g., the SOAP representation employed the species set {B, P, Br}, a cutoff radius of 5.5 Å, $n_{\mathrm{max}} = 8$, $l_{\mathrm{max}} = 6$, and a Gaussian smearing width of $\sigma = 0.4$, with inner averaging to obtain a fixed-length descriptor per structure. The corresponding low-level MP2 energy $E_{\mathrm{low}}$ was appended as an additional scalar feature prior to standardization and PCA. The resulting SOAP-based feature vectors were standardized using parameters fitted on the training set, compressed by principal-component analysis (PCA) to 10 components (cf. Subsection I D), and subsequently used as input to SVR.

### C. Validation of descriptor-based and SOAP-based Δ-ML models

Figure S1 compares the energetic structure of the matched CCSD-labeled subsets in terms of both relative-energy distributions and distributions of relative-energy differences. The relative-energy histograms in Figure S1a,c,e show that, for all three complexes, the matched subset is dominated by structures close to the minimum, although the breadth of the low-energy basin depends on the electronic-structure method. For $F_3B{-}PF_3$, CCSD, MP2, and DFT all yield strongly peaked near-minimum distributions, pointing to a comparatively flat and locally consistent low-energy landscape. On the contrary, for $Cl_3B{-}PCl_3$ and especially $Br_3B{-}PBr_3$, the DFT distributions are broader and shifted toward higher relative energies than the corresponding CCSD and MP2 distributions, indicating a more pronounced reshaping of the low-energy ordering. This trend is also evident from the relative-energy-difference histograms in Fig. S1b,d,f. After removal of trivial absolute offsets by referencing each method to its own minimum, the relative-energy differences with respect to CCSD are generally narrower and more strongly centered around zero for MP2 than for DFT. Thus, within the matched subset, MP2 preserves the shape of the CCSD relative-energy landscape more faithfully than DFT, with the contrast becoming particularly clear for the chlorinated and brominated complexes.

Figure S2 illustrates the strong deterioration in predictive performance for $Br_3B{-}PBr_3$ when $E_{\mathrm{MP2}}$ is omitted from the descriptor vector. In contrast to the near-ideal parity obtained when $E_{\mathrm{MP2}}$ is included, the present model exhibits substantial scatter around the regression line, reflected in the reduced $R^2$ value of 0.72 and the increased MAE of $3.09 \times 10^{-4}$ Eh on the held-out non-training subset. The deviations are especially pronounced for structures outside the densely populated low-energy region, indicating that the geometric descriptors alone are insufficient to recover the

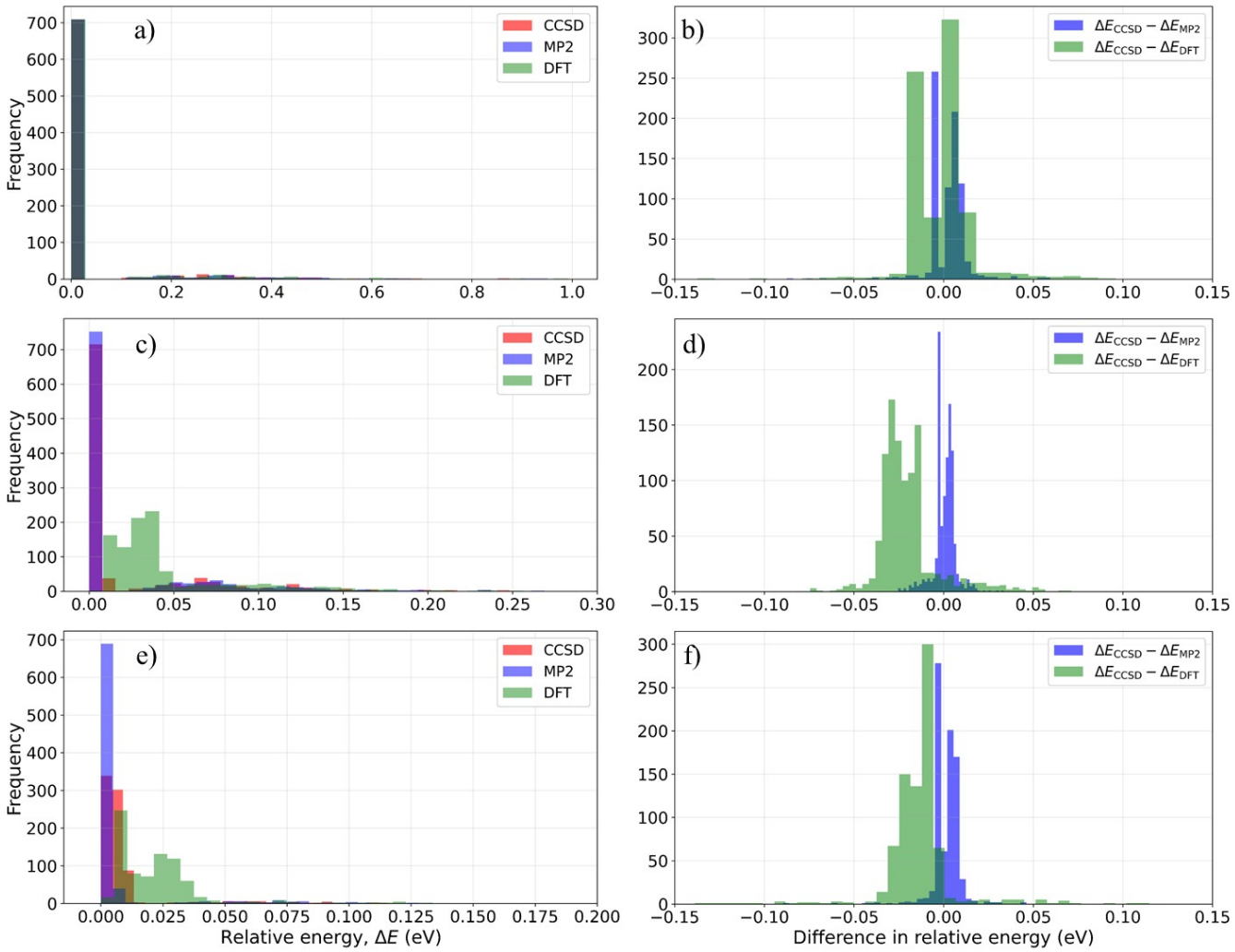


FIG. S1. Relative-energy and relative-energy-difference distributions for the matched CCSD-labeled subsets of a,b) $F_3B{-}PF_3$, c,d) $Cl_3B{-}PCl_3$, and e,f) $Br_3B{-}PBr_3$. Panels a,c,e show histograms of relative energies, $\Delta E = E_i - E_{\min}$, expressed in units of eV for CCSD, MP2, and DFT after referencing each method to its own minimum within the matched subset. Panels b,d,f illustrate the corresponding histograms of relative-energy differences, $\left(E_i^{\mathrm{CCSD}} - E_{\min}^{\mathrm{CCSD}}\right) - \left(E_i^{\mathrm{MP2}} - E_{\min}^{\mathrm{MP2}}\right)$ and $\left(E_i^{\mathrm{CCSD}} - E_{\min}^{\mathrm{CCSD}}\right) - \left(E_i^{\mathrm{DFT}} - E_{\min}^{\mathrm{DFT}}\right)$, which isolate differences in the shape of the low-energy landscape after removal of trivial absolute energy offsets.

CCSD energy landscape with comparable fidelity for the brominated system. These results align with the discussion in the main text and confirm that explicit inclusion of $E_{\mathrm{MP2}}$ is critical for achieving robust predictive accuracy in $Br_3B{-}PBr_3$, whereas its removal leads to a marked loss of model quality.

As shown in Figure S3, the SOAP+PCA+SVR models reproduce the CCSD energies of the held-out non-training subsets with high accuracy for $F_3B{-}PF_3$ and $Cl_3B{-}PCl_3$. For $F_3B{-}PF_3$, the parity plot is essentially ideal, with an $R^2$ value of 0.9986 and a low MAE of $5.33 \times 10^{-5}$ a.u., indicating that the SOAP-based representation captures the relevant structural variation extremely well. Similarly, $Cl_3B{-}PCl_3$ shows excellent predictive performance, with $R^2 = 0.9955$ and an MAE of $9.65 \times 10^{-5}$ a.u., again demonstrating near-quantitative agreement between predicted and explicitly calculated CCSD energies on the corresponding held-out subsets. In contrast, the SOAP-based model for $Br_3B{-}PBr_3$ is noticeably less accurate, with $R^2 = 0.7239$ and an MAE of $2.50 \times 10^{-4}$ a.u., which is apparent from the broader scatter about the ideal parity line. Nevertheless, despite the reduced quantitative accuracy for the brominated system, the SOAP-based workflow leads to the same qualitative conclusions as the descriptor-based models discussed in the main text, underscoring the robustness of the ML-guided screening conclusions with respect to the choice of structural representation.

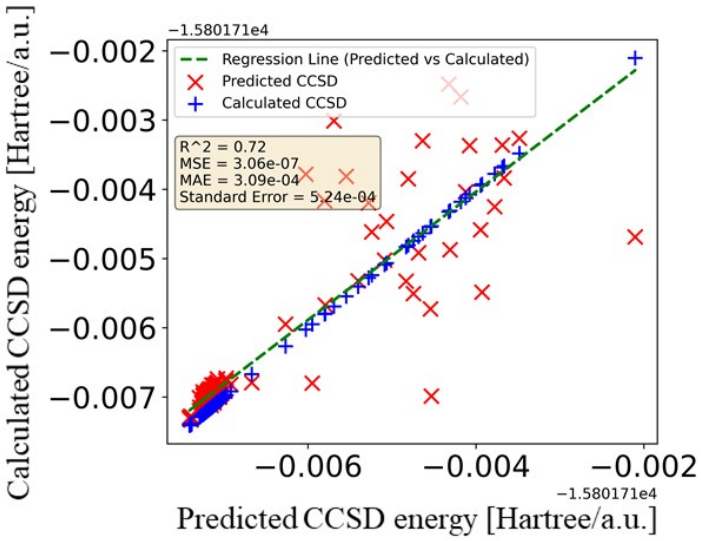


FIG. S2. Parity plot for the held-out non-training CCSD-labeled subset of $Br_3B{-}PBr_3$, showing calculated CCSD energies as a function of the corresponding Δ-ML predicted CCSD energies for the descriptor-based SVR model without inclusion of $E_{MP2}$ as an input feature. Red crosses denote predicted CCSD energies, while blue symbols represent the corresponding CCSD reference values.

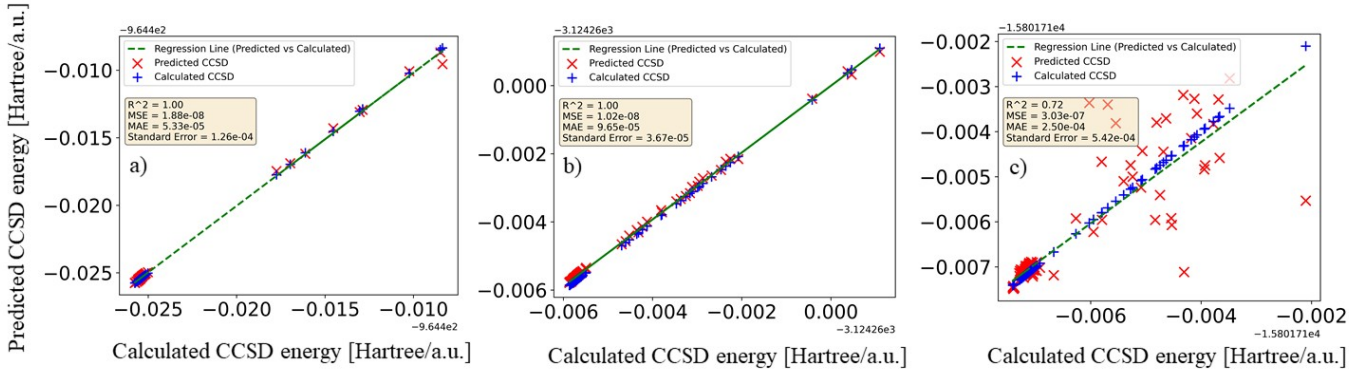


FIG. S3. Parity plots comparing SOAP+PCA+SVR Δ-ML predicted CCSD energies against explicitly calculated CCSD energies for the held-out non-training CCSD-labeled subsets of a) $F_3B{-}PF_3$, b) $Cl_3B{-}PCl_3$, and c) $Br_3B{-}PBr_3$. Red crosses denote SOAP-based predicted CCSD energies, while blue symbols represent the corresponding CCSD reference values. Reported $R^2$, MSE, MAE, and standard error summarize the predictive performance on the corresponding held-out subsets.

### D. Principal-component analysis of SOAP-based feature representations

The PCA explained-variance spectra in Fig. S4 reveal that all three donor–acceptor systems permit substantial dimensionality reduction, although the degree of compression varies across the series. In the SOAP+PCA+SVR workflow, 10 principal components were used for the final reduced representation. For the validation workflow, model selection was based on train-only preprocessing and cross-validation, with emphasis on low-energy accuracy through a tail-weighted CCSD-based loss function. The fluorinated complex, $F_3B{-}PF_3$, exhibits the strongest concentration of variance in the leading principal components, with PC1 alone accounting for 42.1% of the total variance and the first six components capturing approximately 97.5%. The chlorinated system, $Cl_3B{-}PCl_3$, shows intermediate behavior, with a somewhat broader distribution of variance and 94.6% captured by the first six components. The brominated complex, $Br_3B{-}PBr_3$, displays the broadest variance distribution, indicating the highest effective dimensionality among the three systems, although the first six components still account for 94.4% of the variance. Taken together, these results show that the SOAP feature space can be compressed efficiently with only modest information loss, with

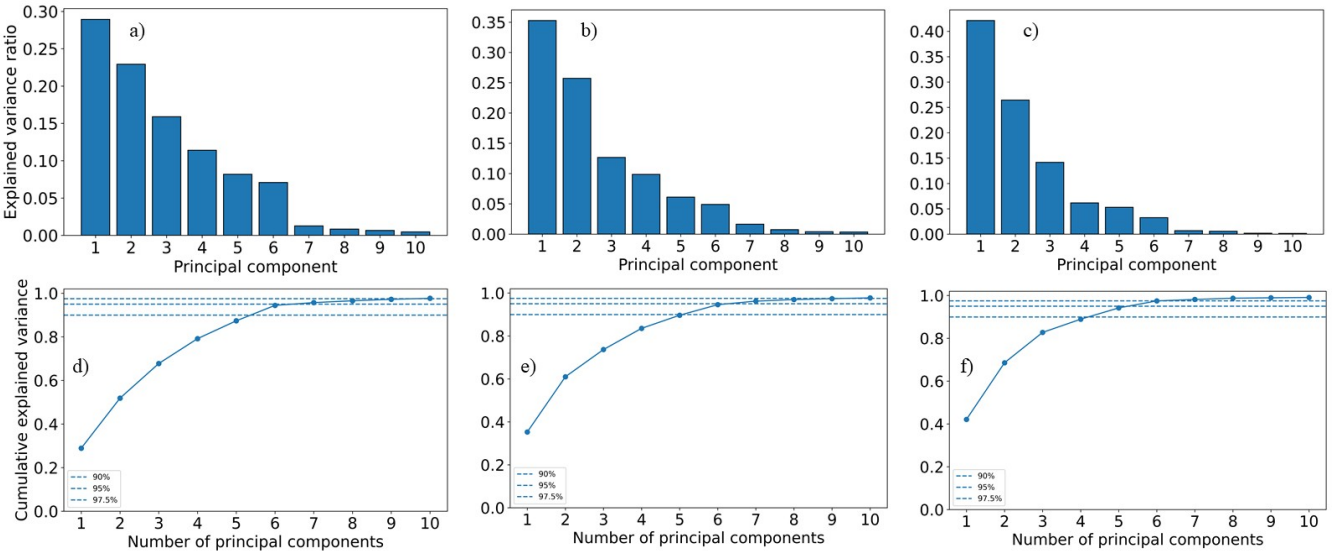


FIG. S4. Explained variance profiles for PCA applied to the SOAP-based feature representations used in the auxiliary Δ-ML validation workflow. SOAP descriptors were standardized using parameters fitted on the training set and subsequently compressed by PCA prior to SVR regression. Top row a–c): per-component explained variance ratios for the first 10 principal components of a) $F_3B{-}PF_3$, b) $Cl_3B{-}PCl_3$, and c) $Br_3B{-}PBr_3$. Bottom row d–f): corresponding cumulative explained variance as a function of the number of retained principal components. Horizontal dashed lines indicate cumulative variance thresholds of 90%, 95%, and 97.5%.

10 retained principal components preserving approximately 99.1%, 98.1%, and 97.7% of the variance for $F_3B{-}PF_3$, $Cl_3B{-}PCl_3$, and $Br_3B{-}PBr_3$, respectively.

### E. CCSD(T) refinement and distance-resolved interaction analysis

Figure S5 compares CCSD and CCSD(T) relative energies for the subset of promising geometries selected with ML guidance and then recalculated at the CCSD(T) level. In each panel, energies are expressed as $\Delta E$ relative to the lowest CCSD(T) structure identified within the candidate subset, and the structures are ordered by increasing CCSD(T) energy to assess ranking stability. The CCSD(T) dataset for each system explicitly includes the equilibrium MP2 geometry and the lowest-CCSD training-set structure, together with additional ML-selected candidates (113 structures for $F_3B{-}PF_3$, 158 for $Cl_3B{-}PCl_3$, and 124 for $Br_3B{-}PBr_3$), thereby enabling a targeted validation of whether the CCSD energetic ordering is preserved upon inclusion of perturbative triples. In summary, Figure S5 shows that the CCSD-based screening successfully captures the relevant low-energy region for all three systems, while CCSD(T) introduces a system-dependent reshaping of the relative energy spread that becomes more pronounced for the chlorinated and brominated complexes.

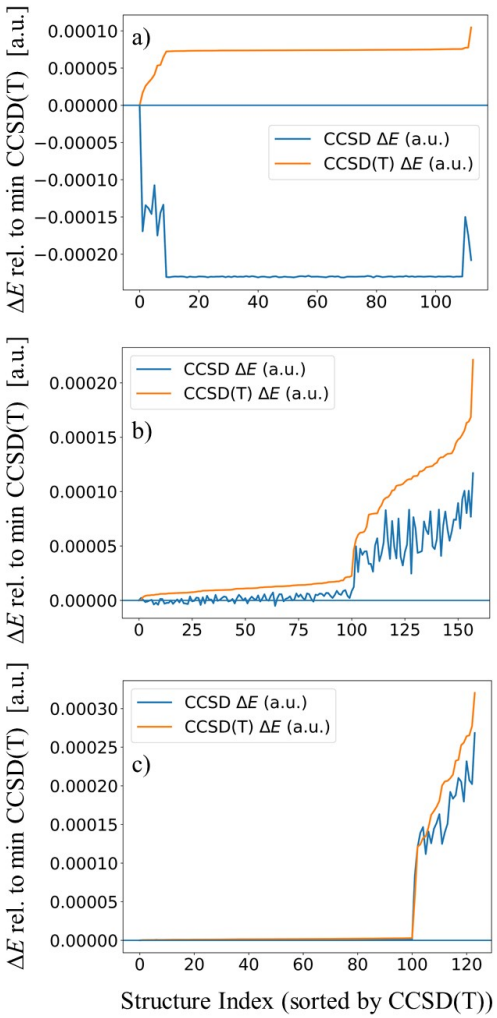


FIG. S5. Relative energy ordering of the ML-selected candidate structures recomputed at CCSD and CCSD(T) level. For each system, $\Delta E$ values are reported in atomic units relative to the lowest-energy CCSD(T) structure within the CCSD(T) subset and structures are indexed after sorting by CCSD(T) energy. Blue curves show CCSD $\Delta E$, orange curves show CCSD(T) $\Delta E$.

[1] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Comput. Phys. Commun. **247**, 106949 (2020).