

Early Energy Prediction in DFT Geometry Optimization: Baseline Sensitivity and the Limits of Convergence Inference

Ahmed Yousif Ghazal

Medical Instrumentation Technology, Technical Engineering College, Northern Technical University, Mosul, Iraq

ORCID: Ahmed Yousif Ghazal 0000-0001-9771-1578

¹ Date of Receiving: 12 September 2026; Date of Acceptance: 26 September 2026; Date of Publication: 30 October 2026

Abstract

Completed geometry optimization histories may support early prediction of final electronic energies, but improvements over current energy can overstate the value of structural machine learning. Seven previously generated adsorption trajectories were reconstructed into 423 unique states. Endpoint priors, energy-based ridge regression, compact geometry features, and interface-centered smooth overlap of atomic positions were compared using nested validation that excluded complete sites. Over records 3–20, the structural model reduced the mean absolute error from 12.340 to 4.262 kilocalories per mole, but the training-endpoint median achieved 3.053 kilocalories per mole. The structural model outperformed this median in only one of seven sites. A common-model evaluation of unseen site pairs yielded 46.3% ordering accuracy, compared with 47.4% for current energy. Prediction-stability triggers occurred in all seven structural-model trajectories, but none of the triggered geometries met all recorded convergence criteria. Entry into a one-kilocalorie-per-mole endpoint band preceded full geometric convergence by 13–85 records. The results of these retrospective measurements show that endpoint-prior controls, common-model ranking tests and separate geometric-convergence tests are important. They are not shown to offer an additional advantage of the tested structural representation to endpoint priors, reliable determination of the most suitable structural representation, or validated improvement in density functional theory computation time.

Keywords: *Density functional theory; Geometry optimization; Endpoint forecasting; Smooth overlap of atomic positions; Baseline comparison; Grouped validation*

Introduction

In electronic-structure geometry optimization, a series of energies, forces, and atomic coordinates is calculated until a stationary structure is reached. The electronic-structure framework is provided by density functional theory (DFT) [1,2], while, In gradient-based optimization, the sequence of the structures that are evaluated is determined by the gradient. These intermediate records contain information beyond the final geometry, although an optimization sequence is an algorithmic search rather than a physical time trajectory.

In the present study, the previously published adsorption calculations on $\text{Al}_{12}\text{P}_{12}$ by BMSF-BENZ [3] are reused. The role of it is to calculate the total electronic energy of an adsorption configuration that was not included in the model fit. All seven configurations have the same molecular composition, nanocage, electronic-structure method and optimizer family.. Consequently, learning the narrow distribution of training endpoints may already yield an accurate prediction, even without an informative structural representation. Related completed-structure calculations have examined hydroxychloroquine adsorption on silicon carbide and aluminum- or silicon-doped carbon nanotubes [4], BMSF-BENZ adsorption on silicon-doped carbon nanotubes [5], and ribavirin adsorption on pristine and doped nanotubes [6]. These four studies provide direct chemical context; the present forecasting evaluation uses only the seven $\text{Al}_{12}\text{P}_{12}$ trajectories from the source study [3].

Geometric and electronic information has been used previously to predict the results of electronic-structure calculations while those calculations are in progress [7]. Smooth overlap of atomic positions (SOAP) describes local

¹ How to cite the article: Ghazal A.Y.; (October 2026) Early Energy Prediction in DFT Geometry Optimization: Baseline Sensitivity and the Limits of Convergence Inference; *International Journal of Universal Science and Engineering*; Vol 13, 91-105

atomic environments while preserving the relevant symmetries [8], and tools like DDescribe and the Atomic Simulation Environment (ASE) exist to implement and utilize it [9,10]. The tools offer a practical foundation for building an interface-centered descriptor. However, their use alone does not establish whether local geometry provides predictive information beyond that provided by energy-based predictors or a prior over final energies.

Continuous-filter neural networks like SchNet [11], transferable neural-network potentials like ANI-1 [12], and more general machine-learned interatomic-potential frameworks [13,14] are examples of recent atomistic machine learning. While highlighting reliance on reference-data quality, chemical coverage, and validation design, these advancements also show significant computational potential.

A larger family of symmetry-aware atomistic representations includes SOAP. The atomic cluster expansion [15] and FCHL descriptors [16] are representative alternatives; recent reviews of atomistic representations highlight invariance, locality, completeness, and transferability as key design factors [17]. Broader reviews of machine learning in molecular and materials science situate these developments within ongoing efforts to accelerate chemical discovery while maintaining physically meaningful validation [18,19].

This study therefore evaluates structural features against deliberately simple endpoint and energy controls. A second objective is to separate endpoint error from candidate ranking: out-of-fold predictions generated by different models do not necessarily reflect the ranking behavior of a single model applied to a shared candidate pool. A third objective is to examine whether stable predictions are accompanied by accurate endpoints and geometrically converged states. All three questions are evaluated retrospectively within one chemical system. The analyses are exploratory and were not prospectively preregistered.

Computational Details

Source calculations and reconstruction

Seven complete optimization logs were selected from the Al-P source folder for sites Br, N8, N9, N58, O35, O42, and S. Parallel folders and property-analysis files were not counted as additional independent trajectories. Source identifiers and SHA-256 hashes are provided with the reproducibility materials. While the route sections specify the B3LYP/6-31G(d) level of theory, geometry optimization and frequency calculations, and a maximum of 1000 self-consistent field (SCF) cycles, the Gaussian output headers identify Revision B.01 of Gaussian 09. The basis designation follows the source logs rather than the 6-31G(d,p) designation in a secondary directive. Each full complex contains 88 atoms: $C_{27}H_{30}Al_{12}BrN_3O_2P_{12}S$ and was considered as a neutral singlet.

Site labels correspond to the starting contact geometry and atomic indices given in the attached coordinate files and are not to be interpreted as representing an invariant nearest contact throughout optimization. Initial and final XYZ structures are included. All chosen logs finish normally, all final geometries satisfy the printed four force/displacement criteria, none of the reported vibrational frequencies are negative.

Each SCF Done energy was matched to the Standard orientation block immediately preceding it. There are 430 such states in the logs. Only one duplicate was found per trajectory when we considered exact equality of printed Cartesian coordinates and electronic energy, resulting in 423 states. Line-number provenance is preserved for every state. The convergence of SCF and convergence of geometry were considered two different properties. The last electronic energy is the supervised endpoint, excluding zero-point, thermal, and free-energy corrections from the target. The task is hence not to predict an adsorption free energy.

Prediction target and causal features

For site s and record k , the target is the remaining electronic-energy correction:

$$y_{s,k} = E_s^* - E_{s,k} \quad (1)$$

The endpoint prediction is reconstructed from the current energy and predicted correction:

$$\hat{E}_{s,k}^* = E_{s,k} + \hat{y}_{s,k} \quad (2)$$

As defined in Eqs. (1) and (2), Hartree differences are converted using 627.5094740631 kcal mol⁻¹ per Hartree. Energy inputs use a fixed numerical offset of -11337 Hartree solely to improve numerical conditioning; no held-out

endpoint is used to set this origin. Energy history contains eight variables: current energy, record index, change from the first record, one-step difference, least-squares slopes over the last three and five records, and two slope-availability indicators. Unavailable slopes and the unavailable first one-step difference are set to zero. All historical information is exclusively derived from available records at or before k .

The minimum drug-heavy-atom/cage distance, the distance between unweighted drug and cage centroids, and the mean and standard deviation of the 36 shortest Al–P distances are the four coordinate-derived variables which the compact geometry panel adds to this history. This gives 12 explicitly defined variables, with no feature selection based on the held-out site.

Structural representations

The interface SOAP center is the midpoint of the instantaneous closest pair consisting of a drug heavy atom and a cage atom. SOAP describes element-resolved neighbor densities and constructs a rotationally invariant power spectrum [8,9]. Parameters are $r_{\text{cut}} = 5.0 \text{ \AA}$, $n_{\text{max}} = 4$, $l_{\text{max}} = 3$, $\sigma = 0.5 \text{ \AA}$, Gaussian-type orbital radial basis, nonperiodic boundaries, and species H/C/N/O/Al/P/S/Br. Each 2112-component SOAP vector is divided by its L2 norm to obtain a unit-length vector. A combined rotation and translation test yielded a maximum difference of 5.0×10^{-16} between corresponding normalized components.

The global control uses the same SOAP settings at all 88 atomic centers, averages their power-spectrum vectors, and L2-normalizes the result. Interface and global variants each append principal component analysis (PCA) scores to the current energy. The hybrid control appends interface PCA scores to the 12-variable compact panel. The interface center is defined as the midpoint between the nearest drug heavy atom and cage atom at each step. This definition preserves symmetry, but a change in the nearest pair can cause an abrupt shift in the center. Such changes were therefore recorded.

Baselines and nested validation

Ten predictors were evaluated: current energy with zero correction; the mean and median of training-site final energies; Ridge using current energy; Ridge using energy and record; the eight-variable energy history; the 12-variable compact panel; energy plus interface SOAP; energy plus global SOAP; and compact features plus interface SOAP. The two endpoint priors use only the six training endpoints, assigning equal weight to each site. They do not access the held-out endpoint.

At each k from 3 to 20, one entire site is held out. The outer training data consist of records 1 through k from the other six sites. Final endpoints of those completed training trajectories are available as labels. This is a controlled retrospective feature-budget experiment, not a claim that all seven calculations could be forecast simultaneously before any endpoint is available. At $k = 20$ there are 120 outer-training rows representing six independent endpoints.

Within each outer training set, an inner leave-one-training-site-out loop selects Ridge α from $\{0.01, 0.1, 1, 10, 100\}$ and, where applicable, PCA dimension from $\{2, 4, 8, 12\}$, bounded by available sample dimension. Every inner split refits PCA on its five training sites, concatenates the selected features, and refits feature standardization. Inner loss is the equal-site mean absolute error at the current checkpoint k . Ties favor fewer PCA dimensions and then stronger regularization. After selection, all preprocessing and Ridge are refitted on the six outer-training sites. No PCA or scaling is fitted globally. At a checkpoint, equal row weights correspond to equal total site weights since each site contributes k rows.

The estimator is an intercept plus L2-regularized coefficients; PCA and nested validation are refitted within each training split. Recent validation studies emphasize the importance of appropriately nested or separated performance assessment since small-sample model assessment might be unstable [20,21]. Prior to this reanalysis, choices of representation were predetermined, although the manuscript's prior results were known. Therefore, rather than providing a rigorous external test, the results of the outer folds provide exploratory comparisons. All out-of-fold forecasts and a few parameters are saved.

Because fold reliance restricts naïve uncertainty interpretation and tuning and performance evaluation on the same resampling scheme can produce optimistic estimates, nested resampling is especially important in the current small-sample setting [20,21].

Summary of errors and uncertainty

The primary metric is the 7-site equal-weight average of the site-level metric, which is the average of the absolute endpoint errors for each site across the 18 checkpoints. For consistency with a discrete trajectory window, this is referred to as integrated mean absolute error (iMAE); no continuous-time integration is implied. The site-level and site-macro summaries are defined by equations (3) and (4).

$$\text{iMAE}_s = \frac{1}{18} \sum_{k=3}^{20} |\hat{E}_{s,k}^* - E_s^*| \quad (3)$$

$$\text{iMAE}_{\text{macro}} = \frac{1}{7} \sum_{s=1}^7 \text{iMAE}_s \quad (4)$$

The errors for all seven sites are reported. Descriptive 2.5th–97.5th percentile bootstrap ranges were obtained by resampling paired site differences 20,000 times using seed 20260912. Because the seven training folds overlap and the sites were not randomly selected from a wide chemical space, these ranges, which represent sensitivity to resampling the observed site summaries, are not calibrated population confidence intervals. There was no row-level significance test used.

As a result, rather than being calibrated predictive uncertainty, the given bootstrap ranges are descriptive. Rather than relying solely on model confidence, uncertainty metrics in prospective chemical screening should be evaluated against error and domain-of-applicability behavior [22].

Ranking and stability analyses

Pooled predictions from leave-one-site-out (LOSO) validation were evaluated using Spearman rank correlation, Top-1 recovery, and selection regret. The resulting rankings combine predictions generated from different training sets. For the endpoint-mean baseline, each held-out prediction is calculated by subtracting the held-out endpoint from the sum of all seven true endpoints and dividing by six. This prediction is a decreasing affine function of the held-out endpoint; consequently, the pooled predictions reverse the true endpoint ranking exactly when all endpoints are distinct. This reversal arises from the evaluation design rather than indicating that the model has learned an inverse chemical relationship.

To compare two unseen sites using the same fitted model, each of the 21 site pairs was held out together. A nested model was trained on the other five sites, with four-site inner training splits, and predicted both held-out endpoints at the same checkpoint. Correct pair ordering and the energy penalty for choosing the wrong member were averaged over 18 checkpoints. Numerical ties were assigned half credit. Current energy, energy Ridge, history, compact geometry, and interface SOAP were tested. The 378 pair/checkpoint comparisons are dependent and must not be treated as 378 independent trials.

A retrospective stability trigger was defined as the first $k \geq 5$ at which the range of the latest three endpoint predictions was at most 1 kcal mol⁻¹. Predictions may come from models refitted at successive checkpoints. Trigger accuracy was evaluated afterwards against the true endpoint at tolerances of 1, 2, and 5 kcal mol⁻¹. Potential record reduction is $(N - k)/N$ and is averaged equally over sites; it is not wall-clock savings. The four Gaussian convergence flags at the trigger were also inspected. Constant endpoint priors are included as a diagnostic: they necessarily satisfy prediction stability immediately, independently of geometry quality.

Energy and geometry agreement and refitted sensitivity analyses

Three additional analyses were specified after inspection of the primary results and are exploratory. First, at absolute electronic-energy residual thresholds of 1, 2, and 5 kcal mol⁻¹, we located each trajectory's first qualifying record and compared it with the first record passing all four printed Gaussian geometry criteria. Endpoint residuals use the eventual energy and are retrospective diagnostics unavailable to an online stopping rule. Force ratios are the maximum of maximum-force/0.00045 and root-mean-square (RMS) force/0.00030; displacement ratios are the maximum of maximum-displacement/0.00180 and RMS-displacement/0.00120, in the corresponding Gaussian convergence-table units. A ratio at most one indicates that both criteria in that family pass. Overflowed source fields remain missing in plots.

SCF convergence and geometry optimization are distinct numerical subproblems. The former concerns convergence of the electronic iterations at a fixed geometry, whereas the latter concerns updates of the nuclear coordinates. This distinction provides a basis for assessing proximity to the endpoint energy, SCF behavior, and geometric convergence separately.

second, each of the seven sites was excluded from the full dataset one at a time. On each remaining six-site cohort we repeated rolling LOSO evaluation at records 3–20 for current energy, training-endpoint median, and interface SOAP. Consequently, inner validation used four sites for training and one for validation, whereas each outer fit needed five sites. Scaling, PCA, and hyperparameter selection were refitted within these subsets using the original grids. This produced 2,268 predictions. We also tabulated exclusion from the original evaluation averages without refitting as a distinct descriptive calculation. Neither procedure supplies independent external validation.

Third, for the original seven-site LOSO predictions we calculated success counts within each of the three absolute-error thresholds at every checkpoint. Pooled counts have 126 site-checkpoint evaluations per method, with 18 correlated observations per site; they are descriptive frequencies, not 126 independent trials or deployment guarantees. These tolerances are illustrative and were not calibrated to an application requirement. No additional DFT calculations were performed.

Results and Discussion

Audited trajectories and endpoints

Table 1. Audited trajectories and final electronic energies.

Site	Unique records	Final energy (Hartree)	Relative energy (kcal mol ⁻¹)	Lowest frequency (cm ⁻¹)
Br	31	-11337.8875355	11.973	1.7064
N8	136	-11337.9066158	0.000	7.2083
N9	20	-11337.8852357	13.416	5.5934
N58	31	-11337.8880429	11.655	4.5116
O35	67	-11337.8911578	9.700	5.0656
O42	92	-11337.8913297	9.592	3.4530
S	46	-11337.8906965	9.990	5.8837

Table 1 summarizes the audited trajectories and endpoints. Fig. 1 shows the complete energy histories. Trajectory lengths range from 20 to 136 records. N8 has the lowest final electronic energy; the next-lowest endpoint, O42, is 9.592 kcal mol⁻¹ higher. The O42–O35 gap is only 0.108 kcal mol⁻¹, whereas the overall endpoint range is 13.416 kcal mol⁻¹. The dataset therefore contains one distinctly low endpoint and several closely clustered higher endpoints. This distribution makes training-endpoint controls especially relevant.

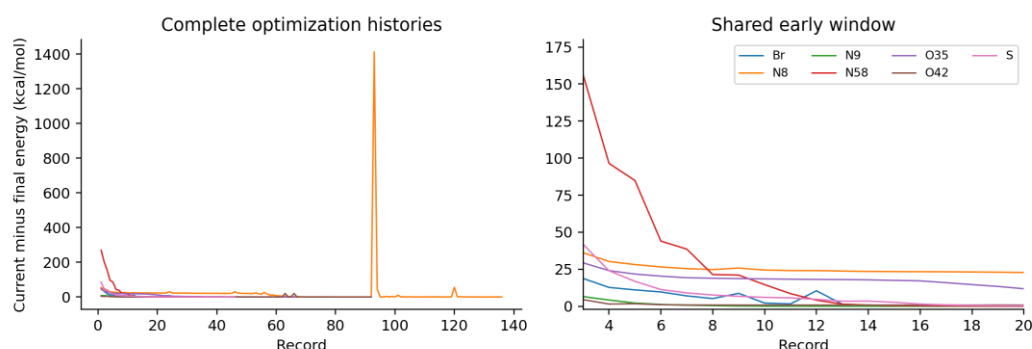


Fig 1. Complete energy histories and an enlargement of the common early window. Energies are shown relative to each trajectory endpoint for retrospective visualization only.

Endpoint priors outperform the structural models

Table 2. Rolling LOSO endpoint errors. All values are kcal mol⁻¹.

Predictor	Macro iMAE	Median site iMAE	Maximum site iMAE
Current energy	12.340	8.155	27.409
Endpoint mean	3.158	2.543	11.054
Endpoint median	3.053	1.810	10.822
Energy Ridge	3.319	2.834	10.824
Energy + record	3.314	2.745	11.129
Energy history	3.776	3.159	11.337
Compact geometry	5.939	4.687	13.315
Interface SOAP	4.262	2.608	11.495
Global SOAP	5.406	5.030	11.372
Compact + SOAP	7.790	8.410	11.841

Table 2 and Fig. 2 compare all ten predictors. Interface SOAP lowers the current-energy error from 12.340 to 4.262 kcal mol⁻¹, a 65.5% reduction. Nevertheless, in the absence of structural inputs, the endpoint median and mean yield 3.053 and 3.158 kcal mol⁻¹, respectively. Energy-only Ridge gives 3.319, while adding record index gives 3.314. Interface SOAP therefore has 39.6% higher error than the endpoint median and 28.4% higher error than energy-only Ridge. SOAP's incremental endpoint-predictive value over these controls is not established by the current comparison.

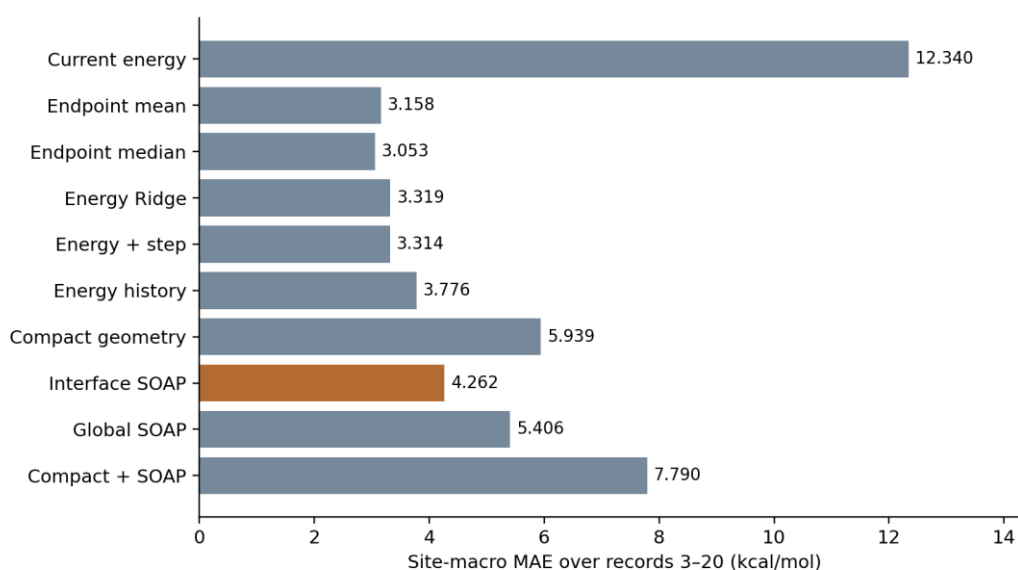


Fig 2. Site-macro endpoint errors for all ten predictors. Interface SOAP improves on current energy but does not outperform the endpoint-prior or energy-only controls.

Energy history gives $3.776 \text{ kcal mol}^{-1}$. The compact geometry, global SOAP, and compact-plus-SOAP controls give 5.939, 5.406, and 7.790, respectively. On average, interface SOAP performs better than global SOAP, but this relative structural advantage is insufficient to outperform the endpoint priors. These findings support a restricted statement about the tested representations; they do not establish that all geometric or hybrid models are inferior.

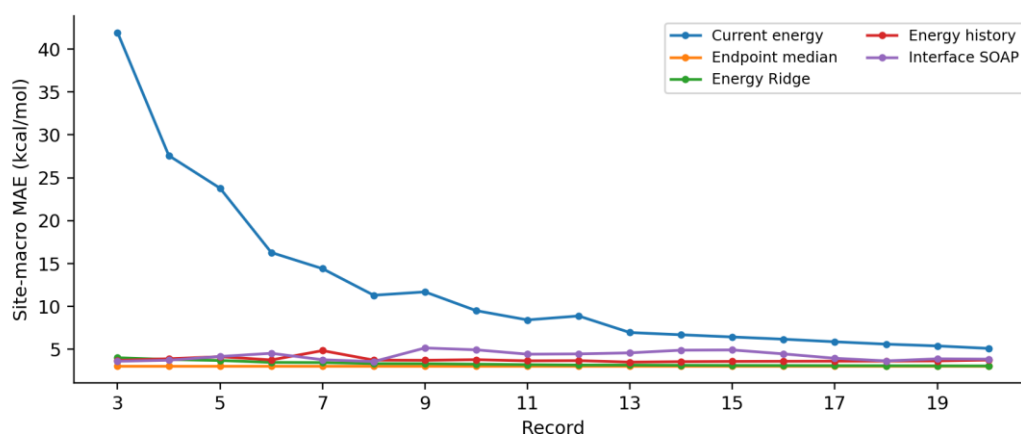


Fig 3. Checkpoint-specific site-macro errors. Constant endpoint priors provide the same prediction at every checkpoint within an outer fold.

Fig. 3 shows the checkpoint-specific errors. Interface SOAP errors show an average of $4.192 \text{ kcal mol}^{-1}$ over records 3–10 and 4.318 over records 11–20. Its average error does not decrease over this range, but the current energy baseline gets closer to the endpoints. The shared window is an equal-record budget, not equal fractional progress: record 20 is the last unique state of N9, but comes before most of the N8 path.

*Site heterogeneity and descriptive sensitivity*Table 3. Site-specific integrated errors in kcal mol⁻¹.

Site	Current energy	Endpoint median	Energy Ridge	Interface SOAP
Br	4.997	2.128	3.118	1.874
N8	25.329	10.822	10.824	11.495
N9	0.863	3.571	4.998	5.229
N58	27.409	1.810	2.834	5.480
O35	18.520	1.122	0.351	2.608
O42	1.111	1.230	0.378	1.459
S	8.155	0.688	0.730	1.690

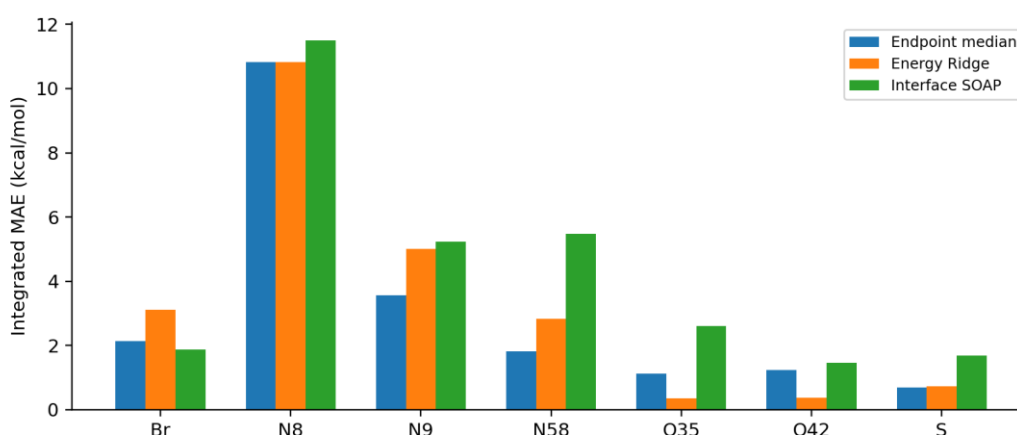


Fig 4. Paired site errors for the endpoint median, energy-only Ridge, and interface SOAP.

Interface SOAP outperforms the endpoint median only for Br (1.874 versus 2.128 kcal mol⁻¹), with larger errors at the other six sites (Table 3 and Fig. 4). Both predictors have their largest site errors at N8: 11.495 and 10.822 kcal mol⁻¹, respectively. Omitting N8 from the descriptive macro summary gives interface SOAP and endpoint-median errors of 3.057 and 1.758 kcal mol⁻¹, respectively; the direction of the difference is therefore not driven solely by N8. This summary is not a retraining experiment.

The paired mean difference, interface SOAP minus endpoint median, is 1.209 kcal mol⁻¹, with a descriptive site-bootstrap percentile range of 0.422–2.153 kcal mol⁻¹. Interface SOAP and energy-only Ridge differ by 0.943 kcal mol⁻¹ on the corresponding measure, with a descriptive range of 0.035–1.788 kcal mol⁻¹. These ranges should be assessed together with the errors for all individual sites, rather than treated as evidence of statistical significance at the population level.

Common-model ranking remains unreliable

The pooled LOSO Spearman correlation of -0.413 for interface SOAP and a perfect -1 for the endpoint mean. All tested models have zero Top-1 recovery in the shared window. The endpoint-mean result shows that negative pooled correlation can be caused by different training endpoints. Interface SOAP selection regret is 11.347 kcal mol⁻¹; this is not solely due to close adjacent gaps, as the penalty is too large (11.347 kcal mol⁻¹) compared to the 13.416 kcal mol⁻¹ endpoint range.

Table 4. Common-model leave-two-site-out pair ranking.

Predictor	Correct ordering (%)	Mean pair regret (kcal/mol)
Current energy	47.4	3.037
Energy Ridge	45.5	2.689
Energy history	45.2	2.928
Compact geometry	44.4	2.993
Interface SOAP	46.3	2.635

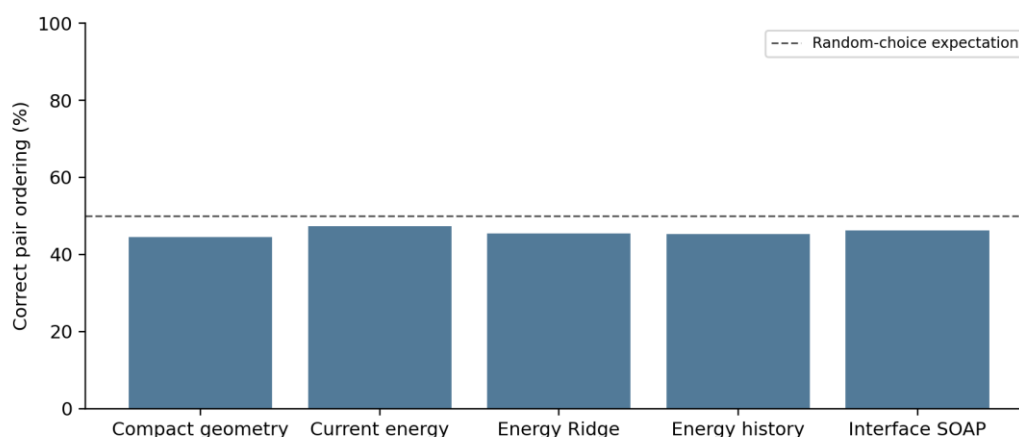


Fig 5. Pair-ordering accuracy when a common model predicts both unseen sites. The dashed line represents the 50% expectation under random choice, not a statistical confidence bound.

Interface SOAP correctly orders 175 of 378 dependent pair/checkpoint comparisons (46.3%), compared with 179/378 (47.4%) for current energy (Table 4 and Fig. 5). Its mean pair regret is lower (2.635 versus 3.037 kcal mol⁻¹), but this exploratory result does not establish reliable candidate selection. None of the tested methods exceeds the 50% random-choice expectation for common-model pairwise accuracy. An external ranking test over seven candidates would require a separate training set.

Prediction stability does not establish geometric convergence

Table 5. Retrospective interface SOAP stability triggers.

Site	Record	Endpoint error (kcal mol ⁻¹)	Potential records avoided (%)	All geometry criteria met
Br	10	0.351	67.7	No
N58	5	5.452	83.9	No
N8	9	11.123	93.4	No
N9	11	6.151	45.0	No
O35	7	2.286	89.6	No
O42	19	0.944	79.3	No
S	16	1.822	65.2	No

Interface SOAP triggers at all seven sites, at a mean record index of 11.0 (Table 5). The mean potential record reduction is 74.9% (Fig. 6). Endpoint errors are within 1, 2, and 5 kcal mol⁻¹ at two, three, and four sites, respectively. The mean trigger error is 4.018 kcal mol⁻¹. None of these seven states satisfies all four recorded geometry-convergence criteria.

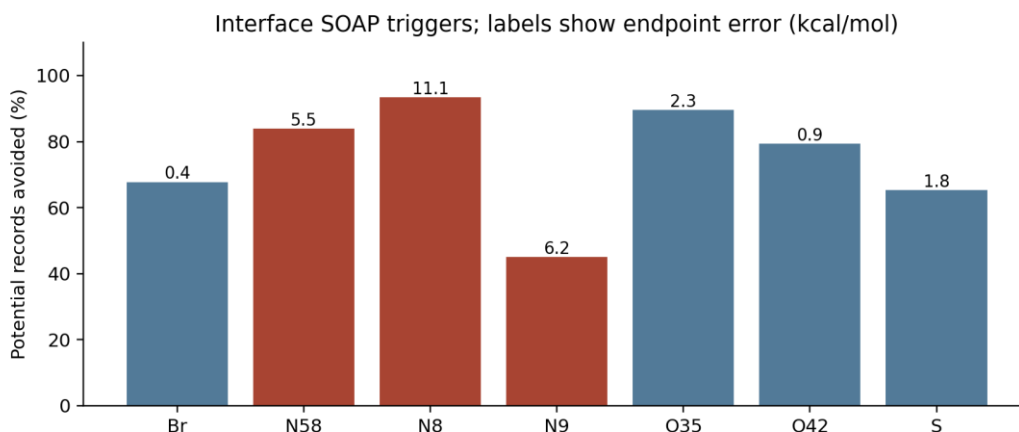


Fig 6. Possible record reduction at interface SOAP triggers. Labels indicate endpoint errors; red bars are errors above 5 kcal mol⁻¹. These are historical record counts, not estimated speed.

For the endpoint-median predictor, the stopping criterion is triggered at record 5 for every site because the prediction is constant, yielding a mean apparent record reduction of 87.9%. This does not indicate a better stopping algorithm: none of these states meets all geometric-convergence criteria. The control indicates that prediction stability alone is not sufficient to serve as an operational convergence certificate. Also, an accurate endpoint prediction does not yield the optimized coordinates, forces and vibrational information that comes from a full DFT calculation.

Source audit of the N8 excursion

At record 93 for N8, the SCF energy after 21 SCF cycles is 1413.213 kcal mol⁻¹ above the final endpoint energy. The reported SCF convergence value is 0.7591×10^{-8} . The same file also gives RMS internal force = 2.287019, and there is a maximum force field greater than the print width. At record 92, both force criteria are met; but according to the printed convergence table, the maximum displacement is 3.29177 in the reported units, and both displacement criteria are not met. The log marks a switch to a linear search around the excursion. Therefore, the spike resides in the source energy, coincident with extreme optimization instability; not induced by the plotting code.

The excursion persists. This is a numerical optimization event and should not be interpreted as a thermally accessible process, a reaction barrier, or a physical dynamical route. It is also beyond the forecasting window and cannot explain the early N8 prediction error by itself. The total examined energy path is 3112.891 kcal mol⁻¹, along with the net relaxation taken to be 54.376. Contact-pair switches are 15 along the whole N8 trajectory, in contrast to 0 to 4 among other sites; their causal role is not established.

Accurate energy can precede geometric convergence

Each trajectory satisfied all four geometry criteria only after first entering the 1 kcal mol⁻¹ endpoint band. The gap varied between 13 records for N9 and 85 for O42 (Fig. 7). N8 made its first appearance at record 68 and satisfied all criteria at record 136, with its later energy excursion showing that first entry need not be permanent. Geometric unconvergence is common in this energy band: 244 of 251 records in this energy band, when viewed across all trajectories, were still geometrically unconverged. These are correlated records, describing paths rather than estimating an independent failure probability. The result is to separate endpoint-energy proximity from convergence of the associated structure; it does not constitute a usable online oracle.

For Br, N8, N9, N58, O35, O42, and S, respectively, the first records within the energy band were 14, 68, 7, 14, 37, 7, and 17; the first fully converged records were 31, 136, 20, 31, 67, 92, and 46. The corresponding gaps were 17, 68, 13, 17, 30, 85, and 29 records.

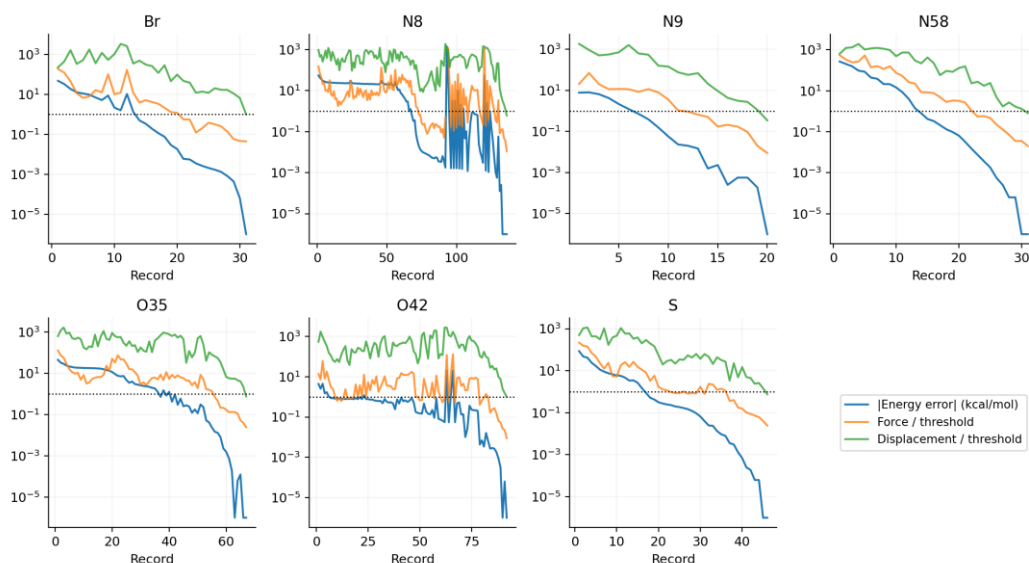


Fig 7. We show retrospective absolute energy residual and normalized force and displacement criteria along each trajectory. The horizontal line denotes 1 kcal mol⁻¹ for the energy trace and 1 for the ratios of the criteria. These quantities have different interpretations but are plotted against the same diagnostic axis; all four criteria must be met for geometric convergence. Missing overflow values are not interpolated.

Sensitivity to removing a complete site

After removing each complete site and repeating nested evaluation, the difference in integrated mean absolute error between interface SOAP and the endpoint median ranged from -1.185 to 2.612 kcal mol⁻¹. Positive values favor the median. SOAP had lower errors in only one of seven reduced cohorts (Fig. 8). Removing N8 yielded errors of 10.176, 2.126, and 0.942 kcal mol⁻¹ for current energy, the endpoint median, and SOAP, respectively. Removing a site changes both the training distribution and the evaluation cohort. These differences therefore cannot isolate a causal effect or justify excluding a challenging site from the primary analysis.

In the order Br, N58, N8, N9, O35, O42, and S for the removed site, the refitted current-energy errors were 13.564, 9.829, 10.176, 14.253, 11.311, 14.212, and 13.038 kcal mol⁻¹. The corresponding endpoint-median errors were 2.773, 2.826, 2.126, 2.532, 3.743, 3.725, and 3.936 kcal mol⁻¹, and the interface SOAP errors were 5.385, 3.487, 0.942, 5.037, 3.829, 4.632, and 5.291 kcal mol⁻¹. SOAP-minus-median differences were 2.612, 0.661, -1.185 , 2.505, 0.086, 0.907, and 1.355 kcal mol⁻¹, respectively.

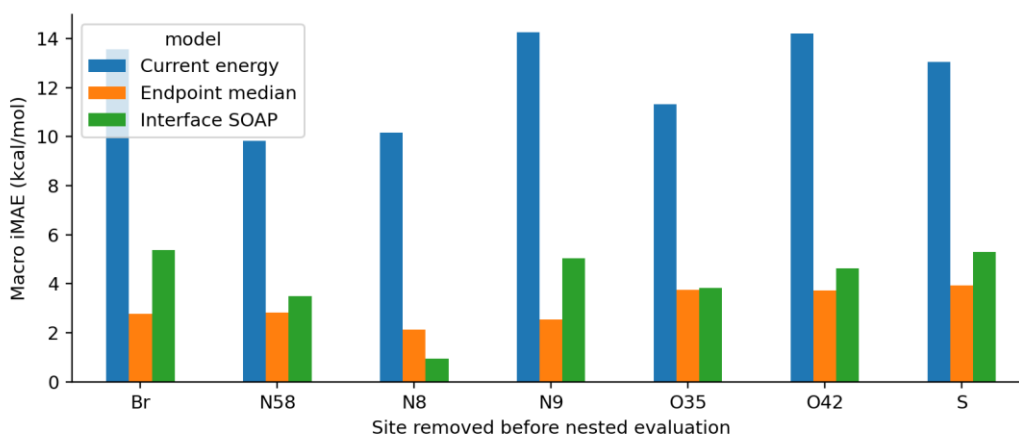


Fig 8. Site-macro errors after removing each site from both training and evaluation and repeating nested leave-one-site-out validation on the six remaining sites. The seven reduced cohorts overlap and are not independent replications.

Practical accuracy depends on the chosen tolerance

At the 1 kcal mol⁻¹ tolerance, the current energy passed in 46/126 evaluations (36.5%), the interface SOAP in 24/126 (19.0%), and the endpoint median in 18/126 (14.3%). At 5 kcal mol⁻¹, the order reversed: the endpoint median passed in 108/126 (85.7%), SOAP in 77/126 (61.1%), and current energy in 63/126 (50.0%). Hence, the median's lower iMAE does not correspond to being the best at every tolerance. At the intermediate tolerance of 2 kcal mol⁻¹, current energy, the endpoint median, and interface SOAP succeeded in 54/126 (42.9%), 72/126 (57.1%), and 43/126 (34.1%) evaluations, respectively. Fig. 9 presents the underlying checkpoint counts. It is important to base conclusions not on pooled correlated evaluations treated as if they were independent evidence.

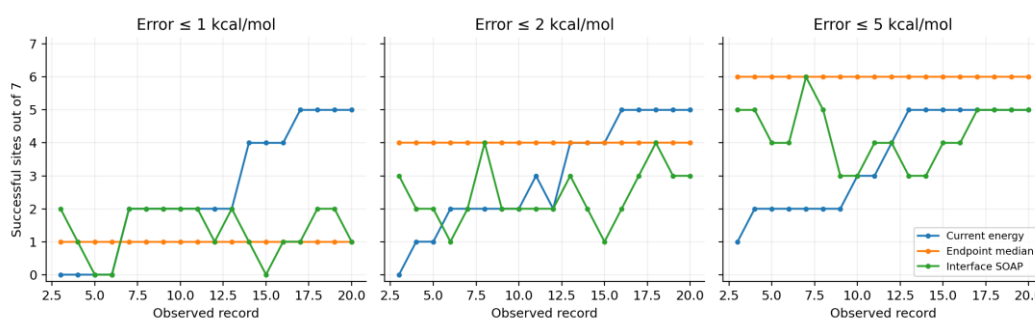


Fig 9. Count of sites with endpoint prediction error at most each tolerance for records 3–20. Each point consists of seven sites; the repeated checkpoints and common training sets do not allow for an independent binomial interpretation.

Interpretation and limitations

The main result is baseline sensitivity. Even if the geometry-aware model substantially outperforms the current energy, its distribution of completed training end points is better when the dataset is fixed-composition. Thus, only current-energy and high dimensional structural baseline reporting would give a partial indication of structural predictive value. The endpoint median is a task-matched control model which does not use the held-out endpoint and is already provided for all fitted models.

The dataset contains seven trajectories, rather than hundreds of independent chemical endpoints. Adding correlated records improves resolution along each optimization but does not add chemical diversity. The low N8 endpoint is inadequately predicted when N8 is omitted from the training, aligning with the challenge of extrapolating beyond the remaining endpoints. There are 7 sites that are not amenable to a general population level interpretation. Here, common-model pair comparisons and the full site-error table provide more information than do row-level statistical tests.

The ranking analysis offers a further methodological insight. Although pooled cross-validation predictions can support a specified marginal error calculation, they may misrepresent how a single model ranks a common pool of candidates. The endpoint-mean reversal illustrates this exact mismatch. Holding out two candidates together eliminates this mismatch, but pair-ordering accuracy remains insufficient to support reliable selection in this dataset. A lower mean pair regret should therefore be reported as an exploratory finding rather than as evidence of screening utility.

Although they distinguish between local and global representations, structural controllers do not exhibit a structural advantage over straightforward energy or endpoint predictors. This may be attributed to the absence of endpoint diversity, miss-specification of the linear model, noisy or redundant directions of the features, and the instability of the instantaneous center of contact. The worse hybrid result does not pinpoint the cause. If one were to search more broadly the descriptors on the same seven sites, there would be more choices available in the selection process; however, the current paper does not claim to have found the solution to the issue by only searching these sites.

Modern E(3)-equivariant interatomic models such as NequIP [23] offer significantly richer geometric inductive biases, however assessing such models on the existing seven-site cohort would still not provide independent chemical validation.

The stopping analysis is purposely negative. Spatial (endpoint) prediction and geometry optimization yield different information. The constant predictor might be perfectly stable even though the current structure is not converged, and N8 demonstrates that small residuals in energy and in the force criteria can exist in tandem with failed displacement criteria even if there is a large excursion. Independent trajectories, quantified computing cost, explicit tests for structural convergence, and calibrated uncertainty would all be necessary for a potential controller. All of this is not provided by a retrospective count of records.

One level of theory, one molecule–cage composition, seven carefully chosen initial configurations, correlated checkpointing, overlapping cross-validation training sets, and the retroactive nature of representation decisions are some of the limitations. This work did not involve any pretrained potentials, novel structure development, external chemical datasets, or new DFT calculations. The given electronic-structure level is not a newly validated representation of experimental adsorption energetics, but rather a source-data condition. There is no proof of wall-clock acceleration, global optimization, adsorption thermodynamics, or generic chemical transfer.

The additional diagnostics further strengthen the interpretation of practical utility. Near-endpoint energy does not guarantee a converged structure, and a predictor chosen based on mean absolute error may have a different ranking under a strict tolerance. Presenting both full convergence diagnostics and tolerance-dependent site counts makes this tradeoff explicit. Refitted site exclusion probes sensitivity within this small dataset (it cannot demonstrate transfer across molecules or cages). All additional analyzes are retroactive and do not represent preregistered validation; the initial seven-site results are still the principal conclusions.

Conclusion

Seven legacy BMSF-BENZ/ $\text{Al}_{12}\text{P}_{12}$ optimization trajectories were reconstructed into 423 unique, source-traceable states. Interface-centered SOAP achieved a rolling LOSO endpoint error of $4.262 \text{ kcal mol}^{-1}$, substantially below the current-energy baseline but above training-endpoint median, mean, and energy-only Ridge controls. Hence, compared to basic endpoint priors, the data do not demonstrate an additional prediction benefit of the tested SOAP model.

Prediction-stability triggers did not coincide with fully converged geometries, and a common-model pair-ranking analysis produced an accuracy of 46.3% for interface SOAP. A reproducible assessment of baseline dependence, stopping-rule limitations, and cross-validation ranking artifacts in a small fixed-composition trajectory dataset is the contribution supported by these results. Affirmative claims about DFT acceleration or structural superiority require additional Independent corroboration.

Prior to full geometric convergence, all seven trajectories approached the 1 kcal mol^{-1} energy band by 13–85 records. Error tolerance also served as a factor in accuracy rankings. When evaluating early forecasting, these new findings encourage reporting energy error, geometric convergence, and tolerance-specific success separately.

Supporting Information (SI)

The source static adsorption calculations were published previously [3]. The present contribution is a trajectory-level reanalysis and controlled predictive evaluation. Source-file identifiers and hashes, extracted state tables, initial and final structures, representation definitions, split logic, selected hyperparameters, predictions, convergence audits, and executable analysis scripts accompany this manuscript. The raw logs remain in the author-controlled source collection. No public repository identifier is claimed here.

Acknowledgements

The author is grateful to Northern Technical University in Iraq for their support in finishing this study.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Zaid H. Al-Sawaff: conceptualization, methodology, computational analysis, data analysis, and writing of the original manuscript. Ahmed Yousif Ghazal: data preparation, analysis, and review and editing of the manuscript. Entisar Younis Abd Al-Jabbar: review and editing of the manuscript. All authors reviewed and approved the final version of the manuscript.

References

- 1- Hohenberg, P., and Kohn, W., 1964, Inhomogeneous electron gas, *Phys. Rev.*, 136, B864–B871, <https://doi.org/10.1103/PhysRev.136.B864>.
- 2- Kohn, W., and Sham, L.J., 1965, Self-consistent equations including exchange and correlation effects, *Phys. Rev.*, 140, A1133–A1138, <https://doi.org/10.1103/PhysRev.140.A1133>.
- 3- Al-Sawaff, Z.H., Dalgic, S.S., Najim, Z.A., Othman, S.S., and Kandemirli, F., 2022, A comparative density functional theory study of BMSF-BENZ chemisorption on $\text{Zn}_{12}\text{O}_{12}$, $\text{Al}_{12}\text{P}_{12}$ nanocages, *Phys. Chem. Solid State*, 23 (1), 120–133, <https://doi.org/10.15330/pcss.23.1.120-133>.
- 4- Al-Sawaff, Z.H., Dalgic, S.S., Kandemirli, F., Monajjemi, M., and Mollaamin, F., 2022, DFT study adsorption of Hydroxychloroquine for treatment COVID-19 by SiC nanotube and Al, Si doping on carbon nanotube surface: A drug delivery simulation, *Russ. J. Phys. Chem. A*, 96 (13), 2953–2966, <https://doi.org/10.1134/S003602442213026X>.
- 5- Al-Sawaff, Z.H., Dalgic, S.S., and Kandemirli, F., 2022, A density functional theory (DFT) study on silicon doped carbon nanotube Si-CNT as a carrier for BMSF-BENZ drug used for osteoporosis disease, *Momento*, 65, 1–24, <https://doi.org/10.15446/mo.n65.99010>.
- 6- Dalgic, S.S., Al-Sawaff, Z.H., Dalgic, S., and Kandemirli, F., 2023, A comparative DFT study on Al- and Si-doped single-wall carbon nanotubes (SWCNTs) for Ribavirin drug sensing and detection, *Mater. Sci. Semicond. Process.*, 158, 107360, <https://doi.org/10.1016/j.mssp.2023.107360>.
- 7- Duan, C., Janet, J.P., Liu, F., Nandy, A., and Kulik, H.J., 2019, Learning from failure: Predicting electronic structure calculation outcomes with machine learning models, *J. Chem. Theory Comput.*, 15, 2331–2345, <https://doi.org/10.1021/acs.jctc.9b00057>.
- 8- Bartók, A.P., Kondor, R., and Csányi, G., 2013, On representing chemical environments, *Phys. Rev. B*, 87, 184115, <https://doi.org/10.1103/PhysRevB.87.184115>.
- 9- Himanen, L., Jäger, M.O.J., Morooka, E.V., et al., 2020, DScRibe: Library of descriptors for machine learning in materials science, *Comput. Phys. Commun.*, 247, 106949, <https://doi.org/10.1016/j.cpc.2019.106949>.
- 10- Larsen, A.H., Mortensen, J.J., Blomqvist, J., et al., 2017, The Atomic Simulation Environment—A Python library for working with atoms, *J. Phys.: Condens. Matter*, 29, 273002, <https://doi.org/10.1088/1361-648X/aa680e>.

- 11- Schütt, K.T., Sauceda, H.E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R., 2018, SchNet—A deep learning architecture for molecules and materials, *J. Chem. Phys.*, 148 (24), 241722, <https://doi.org/10.1063/1.5019779>.
- 12- Smith, J.S., Isayev, O., and Roitberg, A.E., 2017, ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, *Chem. Sci.*, 8, 3192–3203, <https://doi.org/10.1039/C6SC05720A>.
- 13- Deringer, V.L., Caro, M.A., and Csányi, G., 2019, Machine learning interatomic potentials as emerging tools for materials science, *Adv. Mater.*, 31, e1902765, <https://doi.org/10.1002/adma.201902765>.
- 14- Unke, O.T., Chmiela, S., Sauceda, H.E., et al., 2021, Machine learning force fields, *Chem. Rev.*, 121, 10142–10186, <https://doi.org/10.1021/acs.chemrev.0c01111>.
- 15- Drautz, R., 2019, Atomic cluster expansion for accurate and transferable interatomic potentials, *Phys. Rev. B*, 99, 014104, <https://doi.org/10.1103/PhysRevB.99.014104>.
- 16- Christensen, A.S., Bratholm, L.A., Faber, F.A., and von Lilienfeld, O.A., 2020, FCHL revisited: Faster and more accurate quantum machine learning, *J. Chem. Phys.*, 152, 044107, <https://doi.org/10.1063/1.5126701>.
- 17- Musil, F., Grisafi, A., Bartók, A.P., Ortner, C., Csányi, G., and Ceriotti, M., 2021, Physics-inspired structural representations for molecules and materials, *Chem. Rev.*, 121, 9759–9815, <https://doi.org/10.1021/acs.chemrev.1c00021>.
- 18- Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., and Walsh, A., 2018, Machine learning for molecular and materials science, *Nature*, 559, 547–555, <https://doi.org/10.1038/s41586-018-0337-2>.
- 19- von Lilienfeld, O.A., and Burke, K., 2020, Retrospective on a decade of machine learning for chemical discovery, *Nat. Commun.*, 11, 4895, <https://doi.org/10.1038/s41467-020-18556-9>.
- 20- Varoquaux, G., 2018, Cross-validation failure: Small sample sizes lead to large error bars, *NeuroImage*, 180, 68–77, <https://doi.org/10.1016/j.neuroimage.2017.06.061>.
- 21- Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A.J., 2019, Machine learning algorithm validation with a limited sample size, *PLoS ONE*, 14 (11), e0224365, <https://doi.org/10.1371/journal.pone.0224365>.
- 22- Janet, J.P., Duan, C., Yang, T., Nandy, A., and Kulik, H.J., 2019, A quantitative uncertainty metric controls error in neural network-driven chemical discovery, *Chem. Sci.*, 10, 7913–7922, <https://doi.org/10.1039/C9SC02298H>.
- 23- Batzner, S., Musaelian, A., Sun, L., et al., 2022, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nat. Commun.*, 13, 2453, <https://doi.org/10.1038/s41467-022-29939-5>.