ORIGINAL RESEARCH **Open access**

Active Learning-Guided Sampling for Predicting Association and Dissociation Free Energies of Polycyclic Aromatic Compounds

Lars Andersen[1*], Freya Nielsen[1], Mette Sørensen[2]

Abstract

The progression of Polycyclic Aromatic Compounds (PACs) from molecular entities to soot particulates is pivotal for deciphering soot formation chemistry. Valuable perspectives into this progression can be derived by mapping the free energy landscape of PAC dimerization. However, the prohibitively vast universe of conceivable PAC dimers precludes exhaustive simulation, forcing investigators to fit machine learning models on a fraction of the data and extrapolate to the remainder. Accordingly, we put forward and test an active learning protocol designed to pinpoint the most informative PACs for fitting a machine learning model tasked with estimating PAC association and dissociation free energies. Contrasting active learning against random selection revealed that the former delivers swifter loss reduction, attaining comparable precision with a substantially smaller training corpus. Once fitted, the model reliably characterized previously unobserved PACs and proved resilient to modifications in the sampling domain utilized during fitting. In a broader context, this investigation highlights the capacity of active learning to streamline experimental or computational design and sharpen insights derived from resource-intensive models within specialized fields.

Keywords SHAP, Free energy, Polycyclic aromatic compounds (PACs), Active learning, Gaussian process regression, Representer theorem

*Correspondence:
Lars Andersen
laa@energy.aau.dk
[1] Department of Artificial Intelligence for Energy Networks, Faculty of Engineering, Aalborg University, Aalborg, Denmark
[2] Department of Intelligent Energy Systems, Faculty of Engineering, University of Copenhagen, Copenhagen, Denmark

Introduction

The combustion of fossil fuels releases hazardous emissions, including products of incomplete burning, most notably soot [1], a leading atmospheric contaminant with well-documented detrimental effects on cardiopulmonary health [2, 3]. Even so, several physicochemical phenomena underpinning soot genesis remain inadequately characterized and are the subject of ongoing inquiry. A case in point is the transformation of relatively small polycyclic aromatic compounds (PACs) into the larger nanoscale particles that act as embryonic soot.

At present, the pathway from gas-phase PACs to their condensed-phase counterparts remains contested [4], an impasse attributable to the twin obstacles of limited observational resolution and sheer systemic intricacy. A broad consensus holds that chemical interactions among PACs are indispensable during the earliest phases. Yet, the sheer multitude of candidate species—numbering in the millions [5]—frustrates efforts to formulate a resilient framework. Certain kinetic formulations implicate PACs harboring up to 4–5 fused rings as the principal agents of nascent soot, citing their inherent stability [6, 7]. Nonetheless, definitive proof that such species exclusively, or even predominantly, orchestrate soot

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

inception is lacking, especially given their dearth of promptly reactive moieties, such as free radicals, and their tendency to yield stable dimers and clusters [6, 8].

In pursuit of an unbiased scheme capable of singling out the species most instrumental to soot inception across diverse flame environments, we computed free energy (FE) trajectories of PAC association to extract physically plausible collision rates. These realistic rates can subsequently inform rate constants for radical-mediated reactions [9, 10] or help gauge aggregate persistence, given that aggregates might engage with other entities before dissociation. Conventionally, kinetic rate coefficients are derived via collision or transition state formalisms [11], which invoke the FE barrier of the underlying process. That barrier is accessible through biased Molecular Dynamics (MD) computations [12, 13]. This tactic, however, remains hamstrung by the enormous diversity of plausible PACs. An earlier mitigation strategy sought to distill recognizable patterns by marrying heuristic physicochemical intuition with a sparse set of observations [14]. While partly fruitful, such simplistic correlations typically break down when confronted with a more expansive chemical palette, as the mapping between thermodynamic or kinetic observables and molecular descriptors proves stubbornly nonlinear [13, 15].

A more potent alternative has emerged in the form of machine learning (ML) surrogates trained on a curated subset of PAC FEs and subsequently deployed to infer the remainder of the combinatorial landscape [16, 17]. Although these surrogates have yielded informative and modestly successful outcomes, they raise pressing questions concerning the generalizability and representativeness of the underlying training collection. In an ideal scenario, this collection would be minimal to curb computational expenditure, yet comprehensive enough to encapsulate the molecular attributes of primary interest. To date, however, no a priori criterion exists to delineate such an optimal subset. Blind stochastic selection from the pool of potential dimer pairs invites inefficiency and runaway simulation costs. A promising alternative for identifying the PACs that optimally constrain an ML surrogate is active learning (AL), a systematic framework for sequential training set curation.

Contemporary methods, including Bayesian optimization and AL, iterate between fitting an ML surrogate and identifying the most advantageous samples to annex to the training pool. Here we concentrate on AL, whose objective is to minimize the predictive uncertainty of a probabilistic ML model through judicious data acquisition. Bayesian optimization, by contrast, is geared toward extremizing a designated property (e.g., FE), an objective orthogonal to our present purpose [18]. Since AL acquires training exemplars incrementally, it shines in settings replete with unannotated data where determining the target quantity (e.g., FE) is onerous or costly [19, 20]. AL can achieve high predictive fidelity while interrogating merely a slice of the accessible repository, thereby affording considerable time and resource savings [21].

A body of literature has advanced diverse AL selection heuristics for machine learning workflows: visual tracking powered by convolutional neural networks [22], strategies to curtail the number of candidates evaluated within an AL loop [23], and approaches to heighten sample efficiency in Multi-Objective Reinforcement Learning [24], among others. Further recent contributions have formulated an AL methodology for deep Gaussian processes via Markov Chain Monte Carlo [25], a domain-adaptation enhancement for AL grounded in energy-based models [26], and a hybrid paradigm merging AL with federated learning to alleviate the annotation burden while boosting model accuracy [27]. Moreover, AL has been fruitfully deployed across a spectrum of applied chemistry challenges, encompassing polymer design [28], the discovery of stable iridium oxide polymorphs [29], oligopeptide sequences predisposed to form nanoaggregates [30], and optimal stoichiometries for materials exploration [31]. Select methodologies fuse AL with generative ML frameworks to expedite the identification of high-entropy Invar alloys [32, 33], as well as candidate photocatalysts and electrocatalysts [34]. The Gaussian process enjoys particular favor as an AL backbone due to its inherently Bayesian (probabilistic) character, offering both a mean estimate and an associated uncertainty that serves as a natural acquisition metric [35]. Concretely, instances where the model harbors the greatest predictive uncertainty are deemed the most informative candidates for annotation and inclusion in the training regimen [36]. AL grounded in Gaussian process regression has been directed toward refining force fields for MD exploration [37], extracting governing differential equations from noise-corrupted empirical data [38], reducing annotation overhead in hyperspectral image classification [39], guiding experimental campaigns for the geometric control of composite fuselage components [40], and constructing an autonomous framework for reconstructing radiation dose-rate distributions within nuclear environments [41].

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The present contribution documents that active learning (AL) can yield substantial computational dividends when interrogating the FE of PACs. To this end, we alternate between nominating a PAC pair—initially at random and subsequently in accordance with the AL query criterion—and refining a probabilistic ML surrogate. We illustrate how this cyclical strategy begets an accurate regression engine while drastically curtailing the requisite number of MD trajectories. Additionally, we benchmark the surrogate's fidelity on PACs withheld from the training partition under various data-splitting regimes. Finally, we harness explainable artificial intelligence (XAI) tooling [42] to deconvolve how the surrogate links each input feature to the resultant FE estimate. In aggregate, these contributions sketch a template for leveraging AL to impute key properties across combustion science and, more broadly, in any domain governed by sprawling reaction networks whose collective behavior dictates system outcomes.

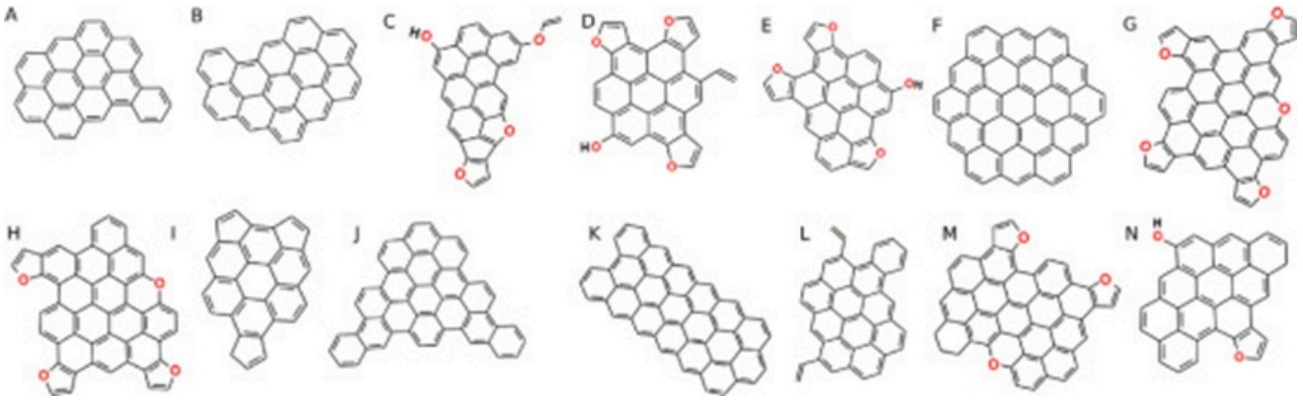
Materials and Methods

Dataset

The free energy profiles and molecular descriptors employed in this study originate from earlier computational work [16, 17]. A complete account of the derivation procedures is available in those publications. We provide here a concise summary of the data's salient characteristics and the nature of the input representations. The descriptor pool consists of 311 molecular features, ranging from elementary properties—mass, carbon-to-hydrogen fraction, number of rings—to more sophisticated attributes that encode charge distribution, polarizability, and topological information. The latter category includes Weighted Holistic Invariant Molecular (WHIM) descriptors [43], Charged Partial Surface Area (CPSA) descriptors [44], Van der Waals Surface Area (VSA) descriptors [45], and tessellation-based descriptors [46].

The investigation encompasses 315 physical scenarios, formed by taking every unordered pair of the 14 molecules displayed in **Figure 1** and evaluating each at three temperature levels: 750 K, 1000 K, and 1250 K. These PACs were deliberately selected during our prior work [15-17] to capture a broad spectrum of properties influencing the dimerization process. A single PAC, p, is encoded through a vector $v_p \in \mathbb{R}^{311}$ reflecting its dimensions, geometry, and elemental composition. Computing any of these features carries negligible cost, as they follow directly from the molecular formula or a canonicalized structural representation. To construct the feature vector for a PAC pair (p, q), designated $x_{pq}^{T} \in \mathbb{R}^{312}$, the harmonic mean h of the two individual vectors v_p, v_q is taken, after which the temperature is appended: $x_{pq}^{T} \doteq \left[h(v_p,\ v_q)^{\top} T\right] \in \mathbb{R}^{312}$ [16, 17]. The use of a symmetric averaging operation ensures that swapping the two molecules leaves the pair descriptor unchanged.

Two separate quantities serve as prediction targets for each system. The first is the association FE (FEA), written a_{pq}^{T}, which measures the free energy gap between the transition state and the fully separated state. The second is the dissociation FE (FED), written d_{pq}^{T} , which measures the gap between the transition state and the bound dimer state. Put compactly, the data consist of triplets (x_{pq}^{T}, a_{pq}^{T},d_{pq}^{T}) for $p, q \in \{A, \ldots, M\}$, $T \in \{750, 1000, 1250\}$, and the central challenge is to infer the computationally intensive scalars a_{pq}^{T}, d_{pq}^{T} from the lightweight vector x_{pq}^{T}.



© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Figure 1. Structure of the PACs considered in this study.

Active learning

Standard supervised learning operates under the assumption that all training input–output pairs are available at the outset and focuses on faithfully reproducing their relationship. Active learning shifts the emphasis toward attaining maximal model performance while expending the minimal number of labeled instances [47], a priority when annotation is resource-intensive. This is accomplished through a query strategy that iteratively identifies which unlabeled examples would be most beneficial to annotate and add to the training set. A frequently adopted criterion is to train a model that reports its own predictive uncertainty and to select the example for which that uncertainty is largest [31].

Gaussian Process Regression (GPR) is a natural vehicle for uncertainty-aware regression [48]. A Gaussian Process (GP) is formally characterized by a mean function together with a covariance, or kernel, function [49]. The kernel is the centerpiece of the GP: it dictates how input–output associations are learned during fitting and how predictions for novel inputs are interpolated [48]. Conceptually, the kernel gauges how similar any two points in the input space are. For an input vector x, the GP model f linking the FE to the feature space is defined such that [49]:

$$f(\boldsymbol{x}) \sim GP(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{x}')) \quad (1)$$

With the zero vector 0 acting as the GP mean and Σ (x, x′) encoding the covariance between x and x′. When the GP is kernel-based, Σ (x, x′) takes the form [49]

$$\boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 k(\boldsymbol{x}, \boldsymbol{x}') \quad (2)$$

Where σ^2 is a scalar variance parameter and k (x, x′) is the kernel function. A kernel k can be understood as an extended inner product or distance metric that maps two inputs (x, x′) into a real number. Its role in ML resembles that of a correlation coefficient: it scores the likeness of x and x′. Broadly speaking, k (x, x′) should increase as the inputs become more strongly correlated [48]. Many kernel choices exist; guided by the strong prior performance of LASSO on FE prediction tasks [16, 17], we use the linear kernel [49]. With $\sigma_0^2 \in (0, \infty)$, this kernel reads [49]:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_0^2 + \boldsymbol{x}^T \boldsymbol{x}', \quad (3)$$

Which corresponds to Bayesian linear regression under a $\mathcal{N}(0, \mathbb{I}_d)$ weight prior for the d-dimensional identity matrix $\mathbb{I}_d$ and a $\mathcal{N}(0, \sigma_0^2)$ bias prior [49]. When the GP is queried at a fresh point $\hat{\mu}(\boldsymbol{x}^*)$, its predictive mean $(\boldsymbol{x}^*)$ is [49]:

$$\hat{\mu}(\boldsymbol{x}^*) = \boldsymbol{r}^T(\boldsymbol{x}^*)(\boldsymbol{R} + \alpha \mathbb{I}_N)^{-1} \boldsymbol{y} \quad (4)$$

Where the vector $\boldsymbol{r}(\boldsymbol{x}^*) = \left[\mathrm{k}\left(\boldsymbol{x}^*, \boldsymbol{x}_1\right)\ \mathrm{k}\left(\boldsymbol{x}^*, \boldsymbol{x}_2\right)\ \ldots\ \mathrm{k}\left(\mathrm{x}^*, \boldsymbol{x}_N\right)\right]^T$ collects the cross-correlations between $(\boldsymbol{x}^*)$ and the N training points, R is the N×N training correlation matrix with entries $(\boldsymbol{x_t}, \boldsymbol{x_s})$,), $R_{t,s} = k(\boldsymbol{x_t}, \boldsymbol{x_s})$, $t, s = 1, 2, \ldots, N$, $\boldsymbol{y} \in \mathbb{R}^N$ holds the training outputs (here, FE values), and is a small positive nugget term that guarantees the matrix inverse exists. The associated predictive variance $\boldsymbol{x}^*, \hat{\sigma}^2(\boldsymbol{x}^*)$ is [49]

$$\hat{\sigma}^2(\boldsymbol{x}^*) = k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{r}^T(\boldsymbol{x}^*)(\boldsymbol{R} + \alpha \mathbb{I}_N)^{-1} \boldsymbol{r}(\boldsymbol{x}^*)\,. \quad (5)$$

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The standard deviation produced by the GPR supplies a natural criterion for sample acquisition [28]. One can evaluate σ̂ for every candidate and launch an MD simulation for the candidate whose uncertainty is highest. The model is not trusted to give an accurate prediction for that point; instead, the newly simulated FE value is fed back into the training set to refit the GPR, as outlined in **Figure 2.** As the number of queried examples grows, the GPR's fidelity is expected to rise, surpassing what would be obtained by randomly selecting the same number of PAC pairs.

An additional sampling baseline is stratified sampling, which partitions the original sample pool into disjoint strata and then draws equally from each [50, 51]. Applied here, the dataset is divided into three strata based on the simulation temperatures (750, 1000, and 1250 K). Thirty-three samples are randomly drawn from each stratum, and one extra sample is randomly chosen from the 1000 K stratum to yield a total of 100 samples. One nuance is that, while stratifying by temperature seems straightforward, the higher-temperature strata inherently carry less signal about the molecular determinants of the FE landscape, since PAC behavior drifts toward the ideal gas limit as temperature increases.

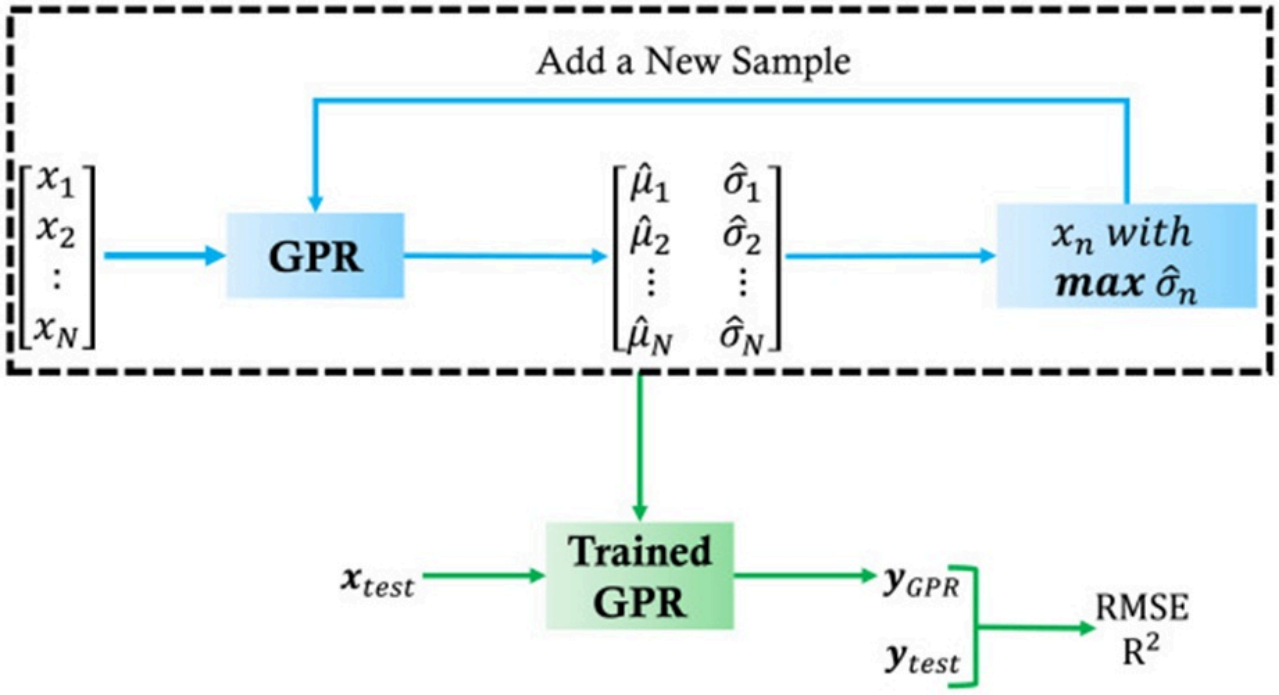


Figure 2. Active learning approach using the standard deviation provided by GPR. xi is the input sample where $i \in [1, 2, \ldots, N]$, μ̂ is the mean of GPR prediction, and σ̂ is the GPR prediction standard deviation.

The GPR was trained on 214 samples and tested on the remaining 101, with performance judged by the Root Mean Squared Error (RMSE) and the coefficient of determination (R2). Implementation relied on the Python libraries scikit-learn-1.2.2 (sklearn) [52] and the active learning package modAL-0.4.2.1 [53]. Default settings were kept for the linear kernel and the GPR, except for the hyperparameter α, which was set to unity after evaluating the alternatives (0.001, 0.01, 1, 0.1, 2, 3). The scalar α is added to the diagonal of the kernel matrix during fitting to maintain positive definiteness [54]. Input and output features were both scaled with MinMaxScaler, mapping every feature into the [0,1] interval [55]. Alongside the linear kernel, several alternatives—Radial Basis Functions, Matern, Rational quadratic, and Exponential Sine Squared—were examined. All proved substantially less effective.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

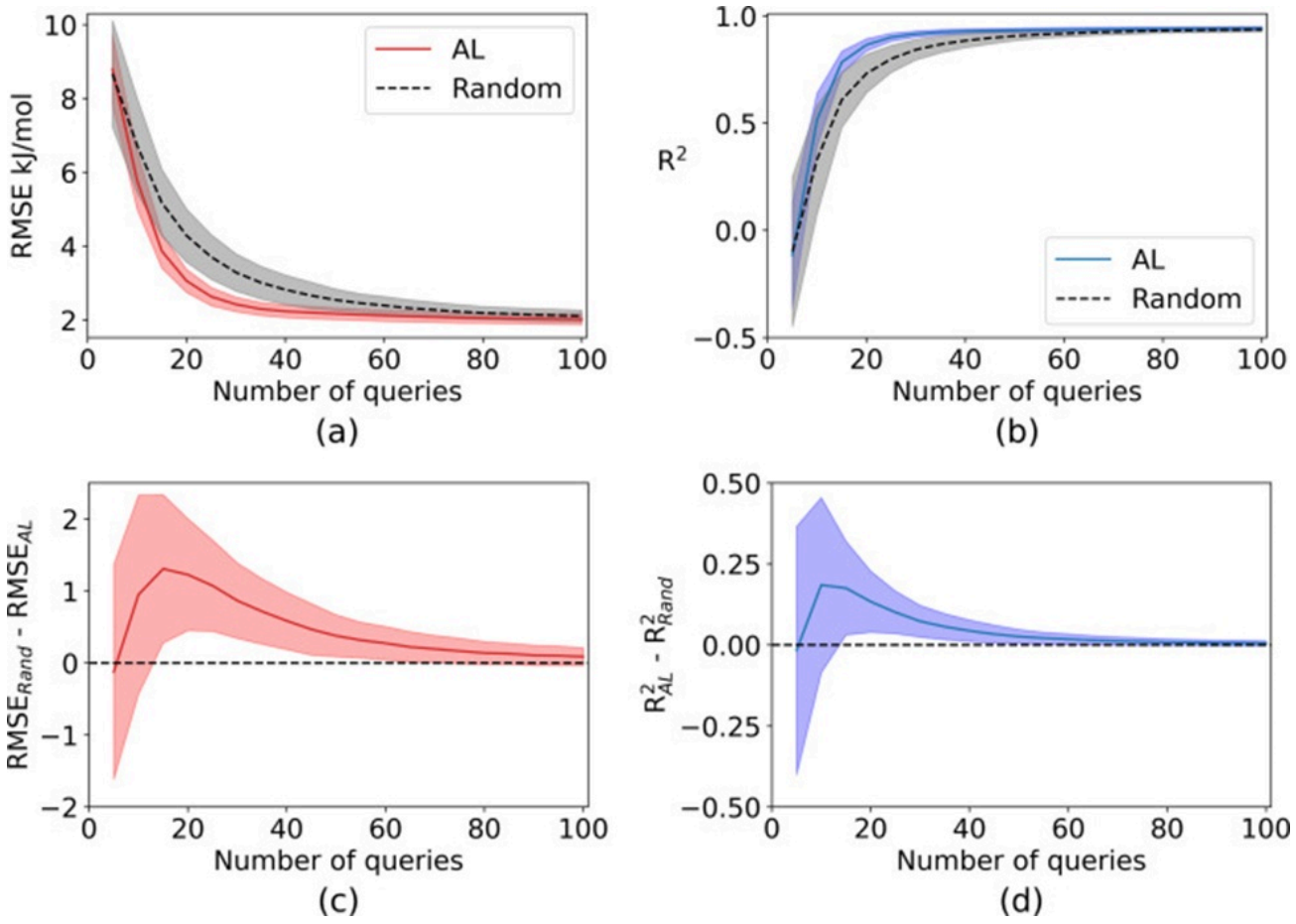


Figure 3. The mean (a) RMSE and (b) R2 computed over 100 runs with varying initial sampling/testing splits for both active learning (AL) and random samples (Random) for association FE. The mean difference of (c) RMSE (kJ/mol) and (d) R2 between random (Rand) and active learning (AL) over the same 100 runs for association FE. Shaded areas indicate ±1 standard deviation.

Results and Discussion

Active learning vs. Random samples

We contrasted the AL-driven GPR (AL-GPR) with a randomly queried GPR (Random-GPR) across 100 independent random partitions of the data into training and testing subsets, as summarized in **Figure 3**. Each partition (case) employs a different assignment of samples to the training and test roles, analogous to a repeated cross-validation procedure for probing model reliability. The outcomes clearly indicate faster learning under the AL protocol. At 40 acquired samples, AL-GPR for FEA attains an RMSE of 2.23 ± 0.17 kJ/mol and an R2 of 0.93 ± 0.01; the corresponding figures for Random-GPR are 2.82 ± 0.39 kJ/mol and 0.88 ± 0.03. To reach an R2 of approximately 0.93 with random sampling requires roughly 35 extra samples, which equates to around 3500 additional core hours of MD simulation. AL-GPR is thus both more accurate and more consistent than its random counterpart, and its advantage holds irrespective of which particular samples seed the training process.

A similar but somewhat weaker pattern emerges for the dissociation free energy (FED). After 70 queries, AL-GPR delivers an RMSE of 6.53 ± 0.48 kJ/mol and an R2 of 0.86 ± 0.02, whereas Random-GPR yields 6.99 ± 0.66 kJ/mol and 0.84 ± 0.03. The AL variant again exhibits better accuracy and tighter confidence bands. The fact that GPR learns more quickly for FED as well, yet achieves a lower absolute performance than for FEA, likely reflects the greater underlying physical

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

complexity of the dissociation process; matching the FEA performance would demand more numerous and more varied training examples. The relative advantage of AL becomes particularly evident in the mean-difference plots (panels c and d of **Figure 3**). These confirm a statistically significant edge for AL over random sampling, even after accounting for the performance ceiling imposed by the finite dataset and the partial overlap of uncertainty bands when fewer than 20 queries have been made **(Figures 4 and 5).**

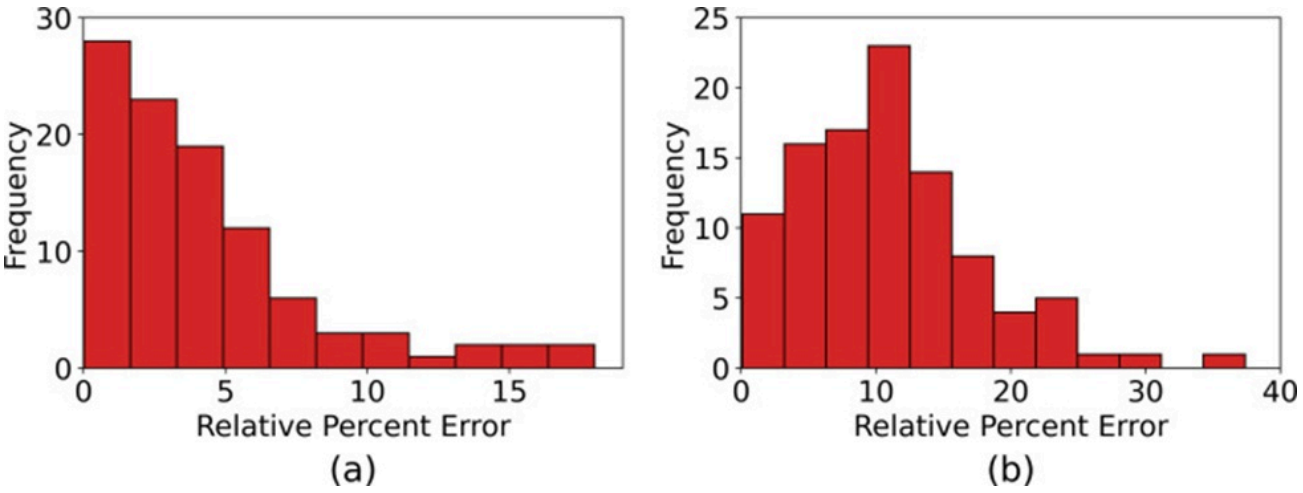


Figure 4. The distribution of relative percent error for predicting (a): Association Free Energy (FEA) and (b): Dissociation Free Energy (FED).

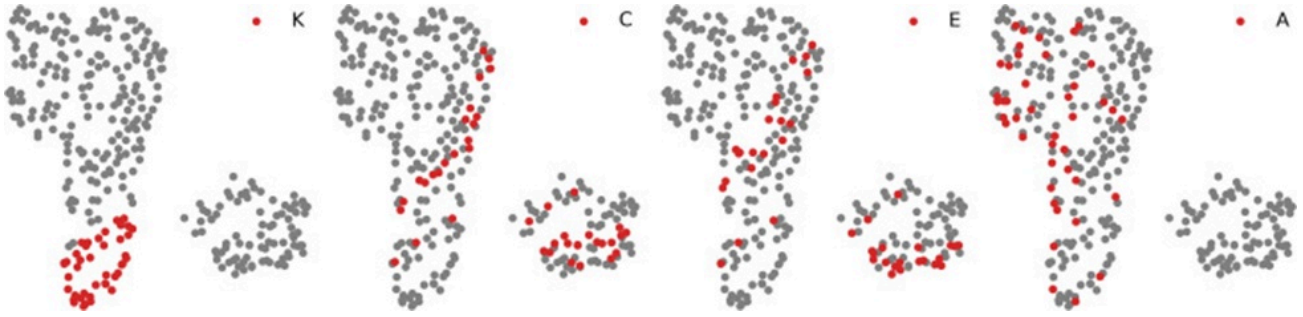


Figure 5. Input features space visualization using UMAP (correlation metric, 25 neighbors). The same representation is repeated to highlight all the datapoints that include a specific molecule, highlighted in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A similar picture took shape when examining stratified sampling: it beat random selection but could not match the efficiency of AL. The only window where stratified sampling held a marginal, statistically indistinguishable edge over AL occurred below 20 queries, simply because stratifying forces the inclusion of all three temperatures from the outset. However, since information content varies considerably across strata, AL quickly closes the gap and then pulls ahead. Beyond that, in a setting where temperature is a continuous degree of freedom rather than a small fixed set, the AL loop can additionally recommend which temperature values to simulate next—an advantage that would likely widen its lead over any stratification scheme.

Multiple routes were pursued to raise the predictive quality for FED. Three are worth highlighting: swapping the single GP for an ensemble (multi-GPR) within the AL loop; employing transfer learning, where a GPR first trained on FEA is repurposed as the starting point for FED training; and shrinking the feature space with Principal Components Analysis (PCA). In the first case, ensembles of 3, 5, 7, and 9 GPRs were stacked against the single-GP baseline. Using the same test setup as the single GP (RMSE = 6.79 kJ/mol, R2 = 0.868), the ensembles delivered nearly identical performance,

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

with the 5-GPR variant showing a whisper of improvement (RMSE = 6.73 kJ/mol, R2 = 0.871). The second idea hinged on the observation that GPR requests near-identical samples for both FEA and FED; transfer learning could, in principle, redirect some of that redundant budget toward more informative FED queries. To probe this, a GPR was run on FEA for 101 queries and then fine-tuned on FED for an additional 100 queries to select new training points. The outcome offered no uplift (RMSE = 6.84 kJ/mol, R2 = 0.866). The third tack compressed the 312-dimensional input space down to 10 principal components, preserving 95% of the original variance. Even with only a 5% information loss, the R2 plummeted to 0.73. Collectively, these null results point to a bottleneck rooted in the MD data themselves, suggesting that a richer and more varied collection of PACs will be essential for markedly better FED predictions.

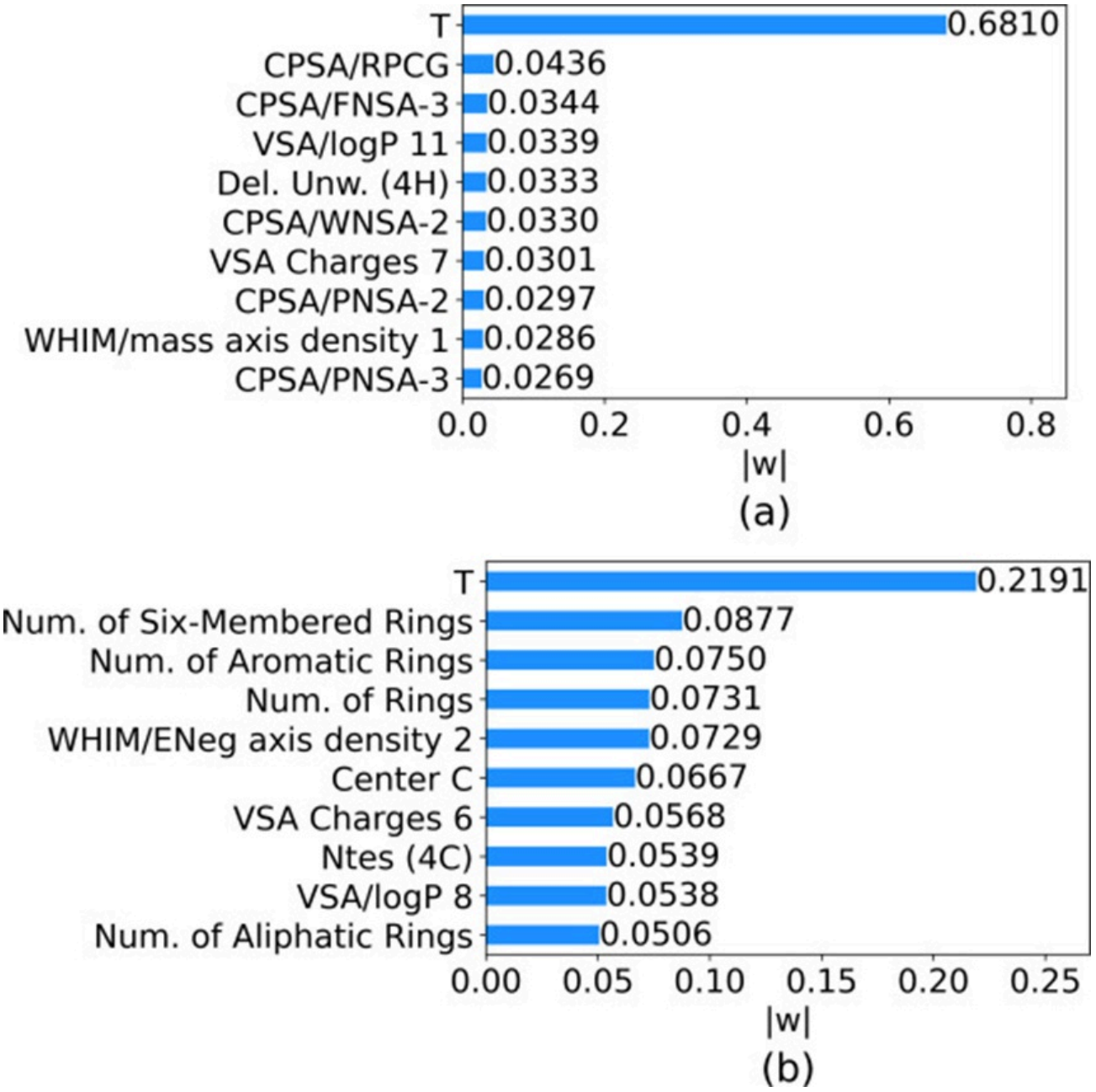


Figure 6. Ten features with the largest absolute values of weights (w) obtained by Representer Theorem to predict (a) FEA and (b) FED. A higher weight value indicates higher importance.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Model robustness

To dissect the prediction errors more finely, **Figure 4** displays how the relative percent error $\frac{|True-Predicted|}{True}\%$) is distributed over the 101 test samples, separately for association free energy (left) and dissociation free energy (right). For FEA, 61 of the 101 samples (60%) land below the dataset-wide mean relative error of 4.34%. The samples that fare worst, sitting in the ≥ 10% error tail, are invariably those simulated at T = 750 K. Turning to FED, the errors cluster more symmetrically around a mean of 10.91%, with 56 samples (55%) falling beneath that average. The largest errors concentrate on molecules C and K, trailed by E, L, and M.

To push the generalizability assessment further, we designed a leave-one-molecule-out protocol that rules out any overlap between training and test pools. Seven separate trials were run; in each, every sample featuring a particular molecule from the set {A, C, E, G, I, K, M} (42 samples apiece) was stripped from the 214-sample selection pool, forcing that molecule to appear exclusively during testing. The outcomes are compiled in **Table 1** and set against the earlier full-dataset R2 figures for FEA (0.96) and FED (0.89).

The table confirms that FEA for held-out molecules can be recovered reliably. In contrast, FED presents a mixed bag, with notably poor results for C and K. This sharp contrast highlights an urgent need for more judicious exploration of the chemical space when targeting FED—exploration that explainable AI could direct to prioritize molecules that yield dependable FED estimates even on unfamiliar PACs.

Table 1. AL-GPR performance for predictions involving molecules not included in the sampling space (214 samples). RMSE is in (kJ/mol); ΔR2 is the relative difference between using the full dataset and a reduced one where all the FE involving the listed PAC are only used in the test set. ΔR2% is the percent difference in R2, i.e., $\Delta R^2\% = \frac{R^2_{ref}-R^2_k}{R^2_{ref}}$. The reference values for R2 are 0.96 for association FE and 0.89 for dissociation FE.

	Dissociation FE ΔR^2%	Dissociation FE R^2	Dissociation FE RMSE	Association FE ΔR^2%	Association FE R^2	Association FE RMSE
A	9.46	0.81	4.96	0.00	0.96	1.58
C	174.09	−0.66	7.83	1.14	0.95	2.01
E	46.53	0.48	7.26	3.40	0.93	2.27
G	61.62	0.34	12.70	3.09	0.93	2.22
I	1.35	0.88	6.17	2.99	0.93	2.19
K	191.52	−0.81	8.18	4.89	0.91	8.18
M	68.21	0.28	8.61	0.97	0.95	1.94

To explain why GPR stumbles on C and K, we projected the input feature space into two dimensions via Uniform Manifold Approximation and Projection (UMAP) [56]. The map, shown in **Figure 5**, hints that K's poor showing might partly stem from its data points lying in a somewhat isolated region of the feature landscape. The same cannot be said for C, the second-worst case: its points appear well intermingled with those of other molecules, E being a prime example. Chemically, the puzzle deepens: C is a close structural cousin of D (both contain side chains, five-membered rings, furanyl rings, and have comparable masses), while K is nearly a twin of I and J (similar tallies of six-membered rings, only minor rearrangements). These similarities make the weak FED predictions for C and K tough to reconcile through chemical reasoning alone.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

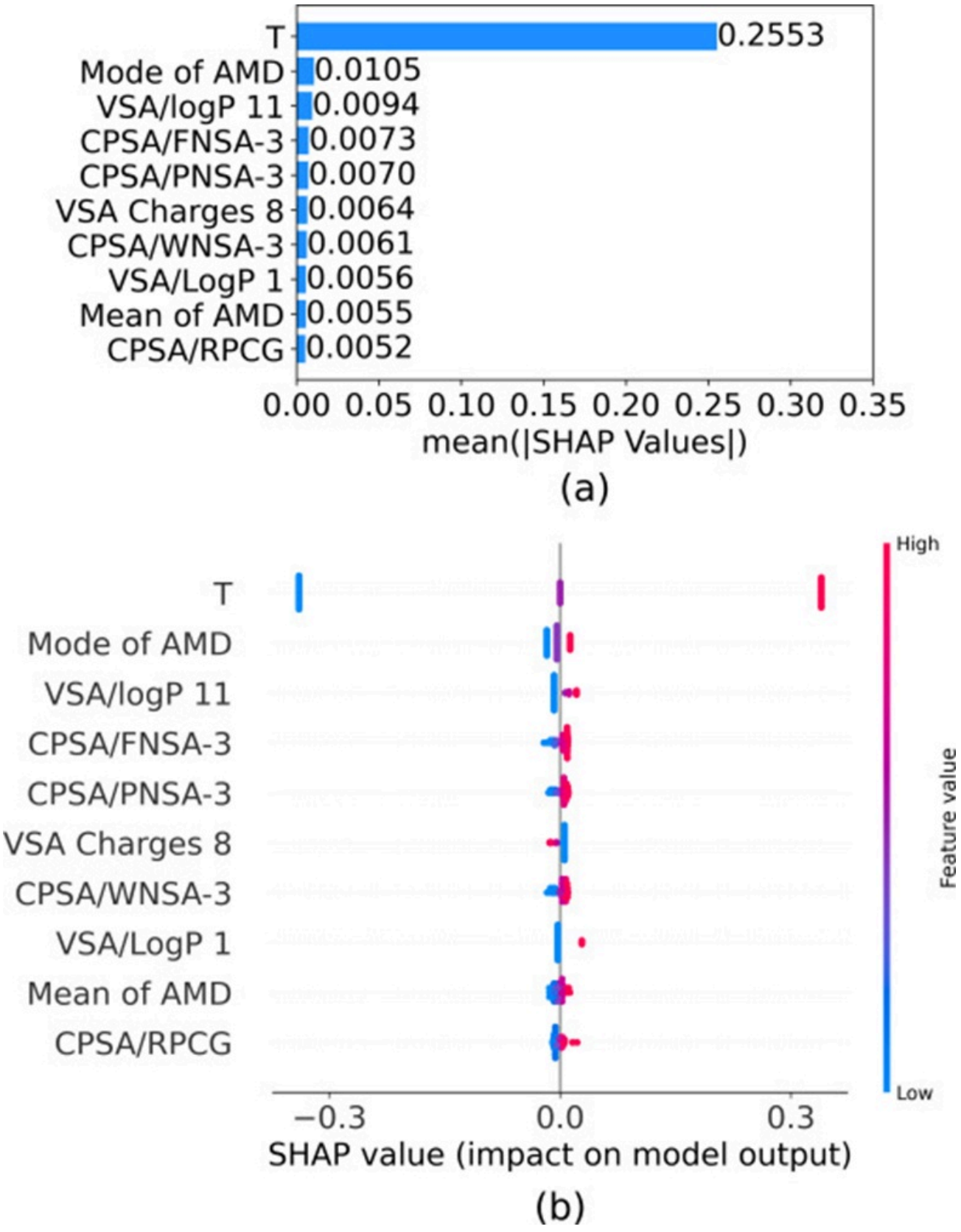


Figure 7. (a) Top important features to predict FEA obtained by SHAP framework, a higher mean absolute of SHAP values (|SHAP values|) indicates higher importance. (b) SHAP bee swarm plot to show the effect of important features'

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

values on FEA, a positive SHAP value indicates that this feature improves FEA prediction and vice versa for negative SHAP values.

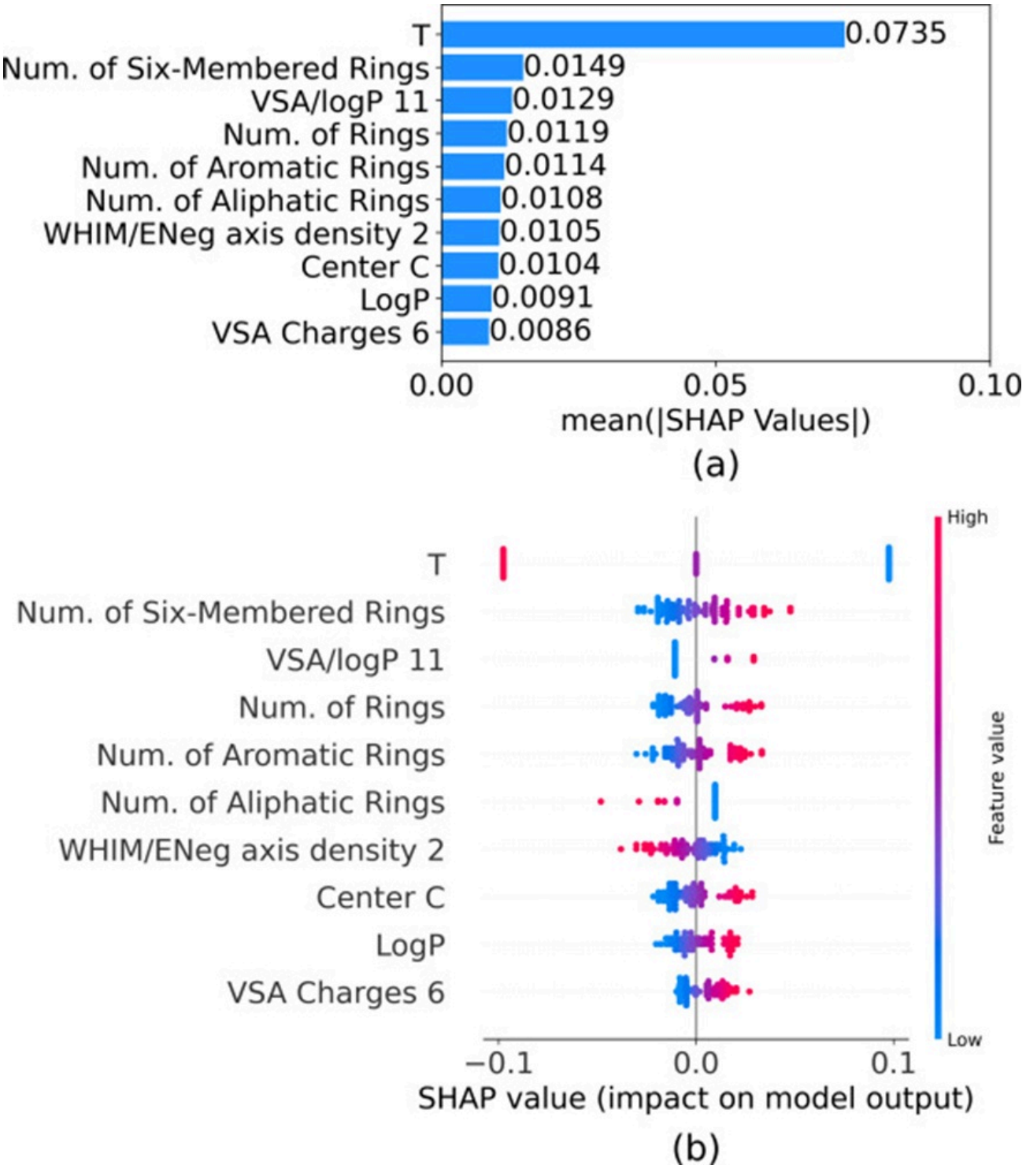


Figure 8. (a) Important features to predict FED obtained by SHAP framework, a higher mean absolute of SHAP values (|SHAP values|) indicates higher importance. (b) SHAP bee swarm plot to show the effect of important feature values on FED, a positive SHAP value indicates that this feature improves FED prediction, and vice versa for negative SHAP values.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Physical interpretation

Making model predictions interpretable carries enormous practical value, especially when computational limitations constrain further data generation and when the ultimate aim is to construct simpler, potentially generalizable analytical relationships. Toward this end, we employ both the representer theorem [57] and the SHapley Additive exPlanations (SHAP) framework [58] to open up the black box of our machine learning models. Each approach yields complementary perspectives on which features matter most, how they interact with one another, and how their magnitudes shape the final predictions.

Since we operate with a linear kernel (a standard inner product), the nonparametric representer theorem grants us direct access to a global feature importance ranking from the GPR fit. With a linear kernel, the theorem establishes a precise equivalence linking the kernel regression predictor to a conventional linear model. In formal terms, given N training samples xs, for s = 1, 2, …, N, the GPR prediction admits the reparameterization:

$$f(\vec{x}) = \sum_{s=1}^{N} \gamma_s k(\vec{x}, \vec{x}_s) = \vec{w}^T \vec{x} + b \quad (6)$$

Where $\boldsymbol{\gamma} \in \mathbb{R}N$ stores the kernel regression coefficients, $\overrightarrow{w} = \sum_{s=1}^{N} \gamma_s \overrightarrow{x}_s$ is the resulting weight vector, and $= \sigma_0^2$ corresponds to the bias.

Figure 6 highlights the ten features possessing the largest absolute weight values (w) for FEA in the upper panel and for FED in the lower one, with a greater absolute weight signaling higher importance. Analogously, the upper panels of **Figures 7 and 8** display the ten most salient features for FEA and FED, respectively, as ranked by SHAP analysis. Features are ordered from most to least influential, where an elevated mean absolute SHAP value (|SHAP values|) denotes greater impact.

Both the representer theorem and SHAP converge on a single message: temperature is the commanding factor for FEA. The CPSA family of descriptors, appearing in **Figures 6 and 7**, quantifies polar contacts between molecules [44]. The mean and mode of the atomic mass distribution (AMD) represent the average and the most probable mass, respectively, across each dimer's mass profile. Delaunay Unweighted tessellations involving four hydrogen atoms (Del. Unw. (4H)) originate from a geometric decomposition methodology [46] and act as a surrogate marker for the arrangement of specific chemical moieties [16]. VSA-based descriptors (e.g., VSA Charges 7, VSA/LogP 11) capture how much surface area falls into a given property window, such as a particular partial charge interval [45]. To illustrate, VSA Charges 7 sums the van der Waals surface area contributed by atoms whose partial charge ranges from −0.05 to 0 [16]. WHIM descriptors encapsulate the dispersion of atoms along the three principal axes of inertia of the molecule [17, 43].

For FED, the representer theorem and SHAP analysis display a striking degree of overlap. Temperature once again sits at the very top and carries substantially more weight than any molecular descriptor. Seven of the nine leading features appear in both rankings; the two points of departure are the number of tessellations with four carbon atoms (Ntes (4C)) and VSA LogP 8, which replace LogP and VSA/LogP 11 highlighted by SHAP. The counts denoted by Num. of Six-Membered Rings, Num. of Aromatic Rings, and Num. of Aliphatic Rings enumerate distinct cyclic architectures within the PACs. At the same time, Num. of Rings represents the overall sum of rings. Center C refers to an aromatic carbon lacking a hydrogen substituent, and logP is a widely used polarity index. The strong concordance between SHAP and the weight vector $\vec{w}$ extracted via the representer theorem is unsurprising, as we relied on kernelized SHAP, which employs local linear regression to infer feature importance [58]. When the underlying model is a kernel machine like a Gaussian Process with a smooth kernel and modest feature interactions, the SHAP importance scores naturally gravitate toward the structure encoded by the representer theorem: features that distinguish the most pivotal training points tend to garner large SHAP values. This close agreement is both gratifying and serves as an independent sanity check on the SHAP analysis conducted with the kernel explainer.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

The lower panels of **Figures 7 and 8** reveal how the actual numerical values of these features influence the predictions of FEA and FED. A positive SHAP value pushes the predicted free energy above the average prediction, whereas a negative value drags it below. Temperature, as expected, utterly dominates the FEA landscape (upper panels of **Figures 6 and 7**), with the other features exerting comparatively minor influence. This finding explains why the model generalizes so well to molecules never seen during training, as reported in **Table 1**. The underlying physics is straightforward: FEA—the free energy separating the transition state from the isolated molecules—is overwhelmingly dictated by the behavior of a gas-phase molecule in isolation. In that dilute limit, the molecule closely approximates an ideal gas particle, so its detailed structural descriptors matter very little until the system approaches the transition state.

Shifting the same analytical machinery to FED **(Figures 6 and 8)** reveals that, while temperature retains its status as the foremost predictor, its relative dominance is notably diluted relative to the FEA case. The other influential attributes are connected to molecular dimensions—counts of various ring types (aromatic, aliphatic, total, six-membered)—charge topology (e.g., VSA, logP), and distinctive structural signatures of the PACs (e.g., center C, Ntes). Although derived through a markedly different mathematical route, these descriptors strongly echo those we previously pinpointed using LASSO [16, 17]. The reproducibility of these findings cements the connections we earlier posited between PAC structural features and the free energy landscape. As before, the bee swarm plot (lower panel of **Figure 8**) captures how departures from the mean value of each feature correlate with the free energy. Nearly every feature bears a positive relationship with FED, with two noteworthy exceptions: the prevalence of aliphatic rings and the electronegativity density distribution along the second principal axis of inertia (WHIM/ENeg axis density 2). These trends align with physical intuition, though the latter descriptor's role demands a closer look. WHIM axial density quantifies atomic clustering around an inertia axis using distributional kurtosis. Given that the PACs in this dataset are virtually planar, the second inertia axis lies approximately in the molecular plane; consequently, the electronegativity distribution along this axis tends to be rather flat and symmetric, since the electronegativity contrast between hydrogen and carbon is tiny (a mere 5.6% difference). Oxygen atoms, by contrast, are considerably more electronegative and sit at the molecular periphery; their presence substantially reduces the kurtosis, which is reflected in a depressed value of WHIM/ENeg axis density 2. Within this context, then, the descriptor operates as a convoluted proxy for oxygenated functional groups, whose inclusion generally compromises the stability of aggregated PACs—exactly the trend observed in the plot.

Limitations and scaling

These results cast light on both the strengths and the constraints of the proposed methodology. While we have established that AL is a workable and encouraging tactic for navigating the expansive chemical space of PACs, projecting the concrete computational savings—and the prerequisites—onto a much larger dataset is not straightforward. On the one hand, the compact dataset used here is perfectly adequate as a proof of principle; on the other, its limited scale inherently curtails the benefits one could extract from a tightly coupled AL/MD loop. Three ingredients dictate the computational expense of such a framework: (a) populating a library of candidate PACs, (b) performing MD simulations for a given collision pair, and (c) retraining the AL model. We already possess reliable, experimentally vetted procedures that select a representative palette of molecules present during soot inception at trivial computational cost [59]. An individual MD simulation can currently be wrapped up in roughly 100 core hours or fewer, and the workflow lends itself well to parallel execution (the free energy reconstruction step is highly parallel). Assembling a larger corpus of training data is therefore not computationally prohibitive. Finally, the Gaussian process hinges on inverting the correlation matrix **(Eq. 4)**, an operation that becomes burdensome only for datasets exceeding about 10,000 samples [49]. Consequently, when mapping out a scaled-up deployment of this approach, only two factors genuinely warrant attention: the anticipated number of data points required, and the interface between the AL engine and the MD simulator. Foretelling precisely how many samples must be acquired to squeeze model uncertainty down to levels comparable with the intrinsic MD noise is challenging. Still, extrapolating from the trajectory of prior investigations—which made do with suboptimal choices of compounds and temperatures—it seems plausible that a carefully curated set of a few hundred to a thousand samples could prove sufficient. The final substantive obstacle, then, is the need to bridge the feature space back to the molecular space, so that the AL strategy can rationally propose which collisions to simulate next when training a larger surrogate. That topic, however, extends well beyond the current scope and will be the focus of forthcoming work.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Conclusion

Understanding the physical assembly of polycyclic aromatic compounds into incipient soot particles is a cornerstone of soot formation chemistry. To tackle this problem, we put to the test an active learning protocol coupled with Gaussian Process Regression, designed to efficiently single out the PACs most worth simulating for training models that forecast association and dissociation free energies. The fundamental premise of AL is to squeeze maximal predictive performance out of a supervised ML model while keeping the tally of molecular dynamics runs as low as possible. Our findings lay bare a substantial edge for active learning over both random selection and stratified sampling on this task. A side-by-side comparison spanning 100 independent random splits demonstrated that AL-GPR reached faster loss convergence and called for considerably fewer training instances to hit the same accuracy mark (e.g., 40 versus 75 samples for FEA), with the promise of far greater dividends in a more exhaustively populated PAC space.

We examined the model's resilience and extrapolative power by challenging it with PACs deliberately omitted from the training pool. The results confirmed that an AL-trained surrogate is robust and reliably handles previously unseen PACs, especially for FEA. Predicting FED, in contrast, turned out to be far less straightforward, delivering lower overall fidelity and conspicuously large errors for a few specific molecules. Attempts to shore up FED accuracy via ensemble GPRs, transfer learning, and dimensionality reduction did not move the needle appreciably. This outcome bolsters the view that the current performance ceiling is baked into the available MD data, and that a chemically richer PAC portfolio, combined with sharper sampling heuristics, will be indispensable.

Explainable artificial intelligence apparatus—namely the representer theorem and SHAP analysis—was deployed to peer deeper into the model's decision-making. The resulting picture aligns with earlier findings, reaffirming previously sketched relationships between PAC structural traits and their free energy landscape. These takeaways will help cement a trustworthy mapping between molecular space and feature space, a prerequisite for fielding an iterative AL/MD cycle on a larger scale.

In its entirety, this investigation spotlights both the promise and the hurdles of enlisting active learning to sample complex physical spaces, a recurring challenge that cuts across domains as varied as combustion, nuclear reactor safety, chemical reaction networks, and nanoparticle synthesis.

Acknowledgements

None

Conflict of interest

None

Financial support

This research has been funded in part by the National Science Foundation Environmental Convergence Opportunities in Chemical, Bioengineering, Environmental, and Transport Systems (ECO-CBET)[Award Number: AWD024893] and by the Michigan Institute for Computational Discovery and Engineering (MICDE) Research Scholar Program .

Ethics statement

None

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

Received: 11 Apr 2022 Revised: 19 Jun 2022 Accepted: 06 Jul 2022
Published online: 15 January 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Fayyazbakhsh A, Bell ML, Zhu X, Mei X, Koutný M, Zhang Y. Engine emissions with air pollutants and greenhouse gases and their control technologies. J Clean Prod. 2022;376:134260.

2. Thomson MJ. Modeling soot formation in flames and reactors: Recent progress and current challenges. Proc Combust Inst. 2023;39(1):805-23.

3. Perera F, Nadeau K. Climate change, fossil-fuel pollution, and children's health. N Engl J Med. 2022;386(24):2303-14.

4. Motaghian S, Beji T. Soot modeling in the numerical simulation of buoyant diffusion flames and fires-a review. Fire Saf J. 2024;104279.

5. Wang Q, Saldinger JC, Elvati P, Violi A. Molecular structures in flames: A comparison between SNapS2 and recent AFM results. Proc Combust Inst. 2021;38(1):1133-41.

6. Mercier X, Carrivain O, Irimiea C, Faccinetto A, Therssen E. Dimers of polycyclic aromatic hydrocarbons: the missing pieces in the soot formation process. Phys Chem Chem Phys. 2019;21(16):8282-94.

7. Shao C, Wang Q, Zhang W, Bennett A, Li Y, Guo J, et al. Elucidating the polycyclic aromatic hydrocarbons involved in soot inception. Commun Chem. 2023;6(1):223.

8. Lowe JS, Lai JY, Elvati P, Violi A. Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity. Proc Combust Inst. 2015;35(2):1827-32.

9. Elvati P, Dillstrom VT, Violi A. Oxygen driven soot formation. Proc Combust Inst. 2017;36(1):825-32.

10. Johansson KO, Dillstrom T, Elvati P, Campbell MF, Schrader PE, Popolan-Vaida DM, et al. Radical–radical reactions, pyrene nucleation, and incipient soot formation in combustion. Proc Combust Inst. 2017;36(1):799-806.

11. Wright MR. Introduction to chemical kinetics. John Wiley & Sons; 2005.

12. Elvati P, Violi A. Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains. Proc Combust Inst. 2013;34(1):1837-43.

13. Elvati P, Violi A. Homo-dimerization of oxygenated polycyclic aromatic hydrocarbons under flame conditions. Fuel. 2018;222:307-11.

14. Herdman JD, Miller JH. Intermolecular potential calculations for polynuclear aromatic hydrocarbon clusters. J Phys Chem A. 2008;112(28):6249-56.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

15. Elvati P, Turrentine K, Violi A. The role of molecular properties on the dimerization of aromatic compounds. Proc Combust Inst. 2019;37(1):1099-105.

16. Saldinger JC, Elvati P, Violi A. A machine learning framework to predict the aggregation of polycyclic aromatic compounds. Proc Combust Inst. 2023;39(1):1009-17.

17. Saldinger JC, Elvati P, Alrawi K, Violi A. Predicting aggregation rates of polycyclic aromatics through machine learning. Fuel. 2024;364:131031.

18. Ziatdinov MA, Ghosh A, Kalinin SV. Physics makes the difference: Bayesian optimization and active learning via augmented Gaussian process. Mach Learn: Sci Technol. 2022;3(1):015003.

19. Ren P, Xiao Y, Chang X, Huang P-Y, Li Z, Gupta BB, et al. A survey of deep active learning. ACM Comput Surv. 2021;54(9):1-40.

20. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal. 2021;71:102062.

21. van Tilborg D, Brinkmann H, Criscuolo E, Rossen L, Özçelik R, Grisoni F. Deep learning for low-data drug discovery: hurdles and opportunities. Curr Opin Struct Biol. 2024;86:102818.

22. Yuan D, Chang X, Liu Q, Yang Y, Wang D, Shu M, et al. Active learning for deep visual tracking. IEEE Trans Neural Netw Learn Syst. 2023.

23. Luo C, Zhu S-P, Keshtegar B, Macek W, Branco R, Meng D. Active kriging-based conjugate first-order reliability method for highly efficient structural reliability analysis using resample strategy. Comput Methods Appl Mech Engrg. 2024;423:116863.

24. Alegre LN, Bazzan AL, Roijers DM, Nowé A, da Silva BC. Sample-efficient multi-objective learning via generalized policy improvement prioritization. arXiv preprint arXiv:2301.07784. 2023.

25. Sauer A, Gramacy RB, Higdon D. Active learning for deep Gaussian process surrogates. Technometrics. 2023;65(1):4-18.

26. Xie B, Yuan L, Li S, Liu CH, Cheng X, Wang G. Active learning for domain adaptation: An energy-based approach. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 36. 2022. p. 8708-16.

27. Ahn J-H, Ma Y, Park S, You C. Federated active learning (f-al): an efficient annotation strategy for federated learning. IEEE Access. 2024.

28. Jablonka KM, Jothiappan GM, Wang S, Smit B, Yoo B. Bias free multiobjective active learning for materials design and discovery. Nat Commun. 2021;12(1):2312.

29. Flores RA, Paolucci C, Winther KT, Jain A, Torres JAG, Aykol M, et al. Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction. Chem Mater. 2020;32(13):5854-63.

30. Shmilovich K, Mansbach RA, Sidky H, Dunne OE, Panda SS, Tovar JD, et al. Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. J Phys Chem B. 2020;124(19):3873-91.

31. Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1):5966.

32. Zhang Y, Wen C, Wang C, Antonov S, Xue D, Bai Y, et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. Acta Mater. 2020;185:528-39.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

33. Rao Z, Tung P-Y, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning–enabled high-entropy alloy discovery. Science. 2022;378(6615):78-85.

34. Mai H, Le TC, Chen D, Winkler DA, Caruso RA. Machine learning for electrocatalyst and photocatalyst design and discovery. Chem Rev. 2022;122(16):13478-515.

35. Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. Chem Rev. 2021;121(16):10073-141.

36. Pasolli E, Melgani F. Gaussian process regression within an active learning scheme. In: 2011 IEEE international geoscience and remote sensing symposium. IEEE; 2011. p. 3574-7.

37. Burn MJ, Popelier PL. Producing chemically accurate atomic Gaussian process regression models by active learning for molecular simulation. J Comput Chem. 2022;43(31):2084-98.

38. Chen J, Kang L, Lin G. Gaussian process assisted active learning of physical laws. Technometrics. 2021;63(3):329-42.

39. Sun S, Zhong P, Xiao H, Wang R. Active learning with Gaussian process classifier for hyperspectral image classification. IEEE Trans Geosci Remote Sens. 2014;53(4):1746-60.

40. Yue X, Wen Y, Hunt JH, Shi J. Active learning for Gaussian process considering uncertainties with application to shape control of composite fuselage. IEEE Trans Autom Sci Eng. 2020;18(1):36-46.

41. Silveira PR, Naiff DdF, Pereira CM, Schirru R. Reconstruction of radiation dose rate profiles by autonomous robot with active learning and gaussian process regression. Ann Nucl Energy. 2018;112:876-86.

42. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115.

43. Todeschini R, Gramatica P. The WHIM theory: new 3D molecular descriptors for QSAR in environmental modelling. SAR QSAR Environ Res. 1997;7(1-4):89-115.

44. Stanton DT, Jurs PC. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. Anal Chem. 1990;62(21):2323-9.

45. Labute P. A widely applicable set of descriptors. J Mol Graph Model. 2000;18(4-5):464-77.

46. Yan X, Sedykh A, Wang W, Zhao X, Yan B, Zhu H. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. Nanoscale. 2019;11(17):8352-62.

47. Cai W, Zhang Y, Zhou J. Maximizing expected model change for active learning in regression. In: 2013 IEEE 13th international conference on data mining. IEEE; 2013. p. 51-60.

48. Radaideh MI, Kozlowski T. Surrogate modeling of advanced computer simulations using deep Gaussian processes. Reliab Eng Syst Saf. 2020;195:106731.

49. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. The MIT Press; 2005.

50. Bennett PN, Carvalho VR. Online stratified sampling: evaluating classifiers at web-scale. In: Proceedings of the 19th ACM international conference on information and knowledge management. 2010. p. 1581-4.

51. ElRafey A, Wojtusiak J. Recent advances in scaling-down sampling methods in machine learning. Wiley Interdiscip Rev: Comput Stat. 2017;9(6):e1414.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press

52. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-30.

53. Danka T, Horvath P. modAL: A modular active learning framework for python. arXiv preprint arXiv:1805.00979. 2018.

54. scikit-learn developers. Scikit-learn user guide. Release 0.22.2. 2020.

55. Thara D, PremaSudha B. Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. Pattern Recognit Lett. 2019;128:544-50.

56. Sainburg T, McInnes L, Gentner TQ. Parametric UMAP embeddings for representation and semisupervised learning. Neural Comput. 2021;33(11):2881-907.

57. Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: Helmbold D, Williamson B, editors. Computational learning theory. Berlin, Heidelberg: Springer; 2001. p. 416-26.

58. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.

59. Wang Q, Saldinger JC, Elvati P, Violi A. Molecular structures in flames: a comparison between SNapS2 and recent AFM results. Proc Combust Inst. 2021;38(1):1133-41.

© 2023 The Author(s). | Published by Institute of Smart Energy Systems Press