


RESEARCH ARTICLE

Accelerated predictions of the sublimation enthalpy of organic materials with machine learning

Yifan Liu¹  | Huan Tran¹ | Chaofan Huang² | Beatriz G. del Rio^{1,3} | V. Roshan Joseph² | Mark Losego¹ | Rampi Ramprasad¹

¹School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

²H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

³Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, Valladolid, Spain

Correspondence

Yifan Liu.
Email: yliu3034@gatech.edu

Abstract

The sublimation enthalpy, ΔH_{sub} , is a key thermodynamic parameter governing the phase transformation of a substance between its solid and gas phases. This transformation is at the core of many important materials' purification, deposition, and etching processes. While ΔH_{sub} can be measured experimentally and estimated computationally, these approaches have their own different challenges. Here, we develop a machine learning (ML) approach to rapidly predict ΔH_{sub} from data generated using density functional theory (DFT). We further demonstrate how combining ML and DFT methods with active learning can be efficient in exploring the materials space, expanding the coverage of the computed dataset, and systematically improving the ML predictive model of ΔH_{sub} . With an error of ~ 15 kJ/mol in instantaneous predictions of ΔH_{sub} , the ML model developed in this work will be useful for the community.

KEYWORDS

organic material, machine learning, density functional theory (DFT)

1 | INTRODUCTION

Sublimation is a physical process, during which a material directly changes its state from solid to vapor without passing through a liquid state, at a temperature T below the triple point of its phase diagram.^[1–3] Sublimation and its reverse processes, that is, deposition or desublimation, are widely used in techniques involving both solid and gas phases of a material, for example, vapor deposition,^[4] vapor phase infiltration,^[5] substance purification/separation processes,^[6] and designing new solid-state forms of substances that are inaccessible by other methods.^[7] The key thermodynamic parameter characterizing this process is sublimation enthalpy ΔH_{sub} , defined as the amount of heat needed to sublime 1 mol of a material. Additionally, ΔH_{sub} offers insight into inter-molecular interactions, which are primarily nonbonding and electrostatic in nature, given that

sublimation does not involve chemical reactions.^[1–3] Quantifying ΔH_{sub} is important for a range of applications, encompassing hybrid membrane process techniques,^[5] membrane separations, de-sublimation separation,^[8] microelectronics,^[4] crystal engineering,^[9] and smart coatings.^[10] Simple organic molecules containing elements such as C, H, O, N, F, and S are especially significant in these contexts as they form the basis of many fundamental organic compounds, frequently used as building blocks in the advanced materials and industrial processes.

Two main classes of experimental methods used to measure ΔH_{sub} are direct (or calorimetric) and indirect approaches.^[11–14] In the direct approach, ΔH_{sub} is measured using a microcalorimeter at a fixed temperature while indirect approaches exploit the dependency of the vapor pressure measured at different temperatures to derive both the enthalpies and entropies of sublimation. The indirect methods

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Materials Genome Engineering Advances* published by Wiley-VCH GmbH on behalf of University of Science and Technology Beijing.

appear to be more robust and reliable than the direct methods.^[11] In general, ΔH_{sub} measurements are rather difficult, expensive, and time-consuming, especially for low-volatility chemicals.^[13,14] It is also worth noting that inter-laboratory agreement on measured ΔH_{sub} data is not always good, and the intrinsic complexity in measuring this quantity could be a main reason.^[15,16]

Computing ΔH_{sub} using first principles methods such as density functional theory (DFT)^[17,18] is a promising approach.^[19–25] In such a computation, ΔH_{sub} is defined as the difference between the energies computed for the solid phase of the material, typically modeled as a crystal, and that computed for the gas phase of the components that make up the material, typically modeled as an isolated molecule. The DFT-based computational approach has its own challenges, including constructing realistic molecular and solid-state models, adequately accounting for the nonbonding and dispersion interactions, which can be computationally expensive,^[26,27] and attaining chemical accuracy, which is ≈ 5 kJ/mol. These challenges strongly limit the applicability of DFT in computing ΔH_{sub} ,^[19–25] particularly in a reliable high-throughput manner, especially for organic materials.^[28,29]

Efforts to develop rapid quantitative structure–property relationship (QSPR) models for ΔH_{sub} emerged quite early.^[30–36] Methods used to create the QSPR models vary from the natural idea of group contributions,^[33,35] which is a linear regression model of fragment energies of predefined ad hoc fragments, to those employing more sophisticated nonlinear regression algorithms and descriptors supported by cheminformatics software such as RDKit.^[30–32,34,37] In some cases, the errors of these methods in predicting ΔH_{sub} could be as low as 15–20 kJ/mol^[34,35] but with questionable transferability, that is, these models do not perform well on unseen compounds/materials.^[36] This problem can be traced back to the choice of the descriptors, which may not capture well the delicate nature of ΔH_{sub} discussed above, and the training data, which may not be sufficiently large, diverse, and complete.

During the last decade, significant progress has been made in the development of comprehensive descriptors/fingerprints for organic molecules and polymers.^[38–41] Nevertheless, there remains a pressing need for a systematic strategy to generate high-quality data for training predictive ML models for ΔH_{sub} . One such strategy that has gained prominence is active learning, which has been widely used to quickly and efficiently develop and improve ML predictive models for various physical properties of materials.^[42–44] This approach allows for iterative improvement of the model by strategically selecting new data points for experimental measurement or computations, thereby enhancing the model's predictive capabilities while minimizing resource expenditure. By integrating group contribution methods including feature engineering with advanced ML techniques and employing strategies such as active learning, researchers can efficiently improve and push the boundaries of ΔH_{sub} prediction, potentially unlocking new insights into material properties and accelerating the discovery of novel compounds with desired characteristics.

In this work, a DFT ΔH_{sub} dataset for 845 representative organic molecules was first built and validated by comparing it with available experimental values. The molecules were chosen to cover a broad range of chemical variants and functionalities, including various combinations of C, H, O, N, F, and S elements. Then, an ML model was trained on the DFT dataset to predict the sublimation enthalpy for new organic molecules. Feature analysis unveiled key factors affecting sublimation enthalpy, offering insights that can aid in the creation of new materials. We also demonstrate an approach for refining the model using an active learning strategy. This entails enhancing the ML model's efficacy and expanding the chemical space through selective DFT calculations. The resulting ML ΔH_{sub} model is set to broaden its scope, potentially encompassing domains such as metal–organic hybrid materials, while enhancing its reliability and precision, as portrayed by the schematic in Figure 1. This synergy of ML prediction, active learning, and feature analysis presents a potent toolset for material design and broadening chemical space exploration.

2 | METHODS

2.1 | First-principles calculations of sublimation enthalpy

We used *Vienna Ab Initio Simulation Package* (VASP),^[45–48] a plane-wave based DFT code, to compute the sublimation enthalpy ΔH_{sub} of the organic materials. Within our numerical scheme, ΔH_{sub} was computed as the enthalpy difference between the gas model, created by placing a molecule in a big simulation box with a vacuum layer of at least 10 Å around the molecule. The crystal model of the material was obtained from Crystallography Open Database (COD).^[49–51] Having the gas and the crystal models, our DFT computations were carried out using a plane-wave basis set that corresponds to an energy cutoff of 400 eV, the Perdew–Burke–Ernzerhof (PBE)^[52] exchange–correlation (XC) functional, and the van der Waals vdW-DF2 corrections^[26,27] for the nonbonding dispersion interactions. The gamma point was used for the gas model, and a k-point spacing of 0.2 in reciprocal space was employed for the crystal model.

2.2 | Materials features and machine learning algorithm

We first featurized the dataset of 845 organic materials based on their chemical structure. This process, also known as the fingerprinting, involves using the PYMATGEN package^[53] to get the SMILES string^[54] of the organic molecules, and then using descriptors of RDKit,^[37] an open-source cheminformatics toolkit to compute the features from the SMILES string. The goal of this process is to represent the chemical structure and the bonding environment of the materials in a

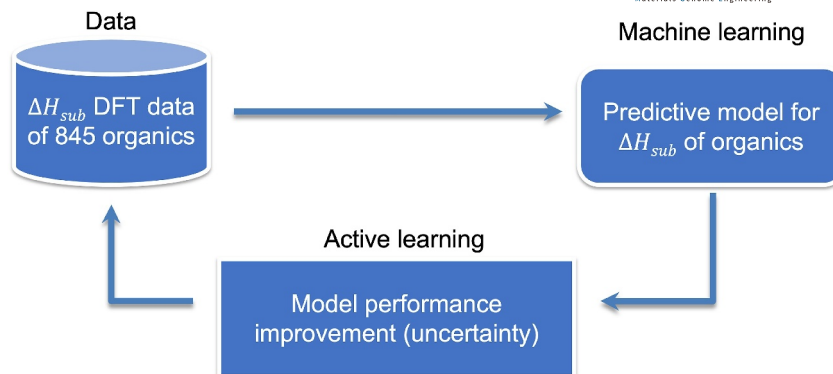


FIGURE 1 Workflow adopted to build the ML model of ΔH_{sub} for organics.

way that is ingestible by the ML algorithms. Ultimately, each material is represented by a set of 208 features. Some of them, which are identically zero, were removed. In Section 3.3, we will use SHAP values (SHapley Additive exPlanations),^[55] a cooperative game theoretic technique, to analyze the importance of the features that influence ΔH_{sub} .

Then, we employed Gaussian process regression (GPR),^[56] using a radial basis kernel and a WhiteKernel function to account for the data noise, to train the ML model on the featurized data. The training process was configured to restart the optimizer five times to ensure robust hyperparameter tuning. The main reason for using GPR in this work is that GPR provides a built-in measure of prediction uncertainty, which plays a central role in the active learning strategy.^[56] A fivefold cross-validation (CV) procedure was used in the training process to regulate potential overfitting, the problem that is critical when learning from small datasets. During the cross-validation process, the training dataset was split randomly into five subsets. A model was trained on the union of four subsets, for example, 80%, and then validated on the remaining subset, for example, 20% of the training set. This process involves examining five models and the hyperparameters of the model with the smallest cross-validation error are selected to train the final model on the whole (100%) training set. Root-mean-squared error (RMSE), mean absolute error (MAE), and the coefficient of determination R^2 are the three metrics that were used to evaluate the performance of the ML models.

2.3 | Active learning workflow

Our active learning approach is depicted in Figure 2 and serves as a test of how future expansion of our dataset can progress. Initially, a seed training set of 299 data points involving only C, H, O, and N, (i.e., F and S were absent) was compiled and their ΔH_{sub} were computed. Starting from the seed dataset, an ML model was developed using GPR. Not surprisingly, the model could not predict ΔH_{sub} for data points containing these chemical species. Then, the ML

model was then used to predict ΔH_{sub} (with uncertainties) of the (test) set of the remaining materials. Based on the ML predictions and uncertainties, a given number of cases in the test set were selected for DFT computations. These newly computed ΔH_{sub} data points were combined with the current training set, allowing for a better ML model to be trained. The key feature of GPR that is at the core of the active learning strategy is the prediction uncertainty, which offers a measure of how much similarity/dissimilarity the examined case (material) shares with all the cases in the training set. Those with high prediction uncertainties are likely far from (or not represented well by) the training data, and thus including them in the training data will improve the ML model (Note, in this work, our DFT calculations have been completed for all the 845 organic materials. During this test of the active learning algorithm, there was no more DFT calculation conducted).

To demonstrate the efficiency of the active learning strategy, we examined two different plans of data selection for iterative augmentation, whose details are also given in Figure 2. In plan A, chosen to be the baseline option, three data points were randomly selected from the test set for inclusion in the training set. In plan B, three new materials with the highest prediction uncertainties were selected. These two plans were performed until all 845 data points had ΔH_{sub} computed and were included in the training set. The performance of these plans is analyzed and discussed in Section 3.2.

3 | RESULTS AND DISCUSSION

3.1 | Creation and validation of computed dataset

A dataset of 845 organic crystal structures was obtained from COD.^[49–51] Each of these materials can be viewed as an infinite lattice of molecules that are packed in an energetically favorable fashion without new primary chemical bonds between the molecules. Therefore, the chemistry of

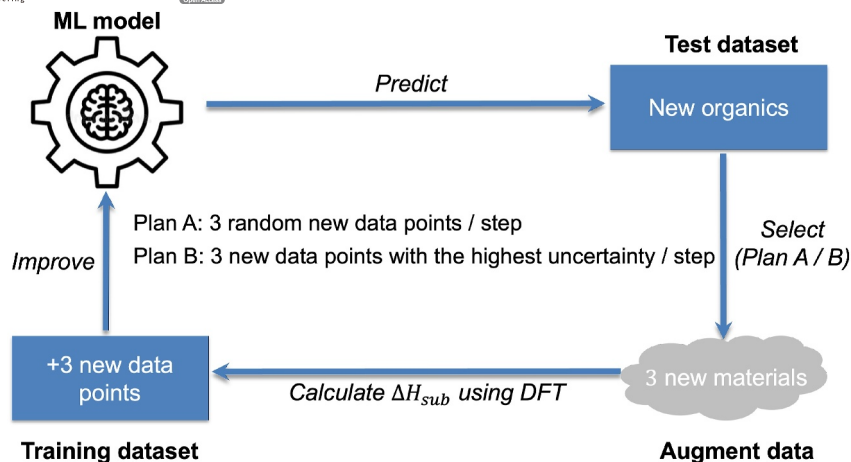


FIGURE 2 Active learning workflow for developing ML models to predict ΔH_{sub} and expanding the training dataset to systematically improve the ML model. Two plans of data augmentation, whose details are discussed in the text, were considered.

each organic crystal in this dataset is specified by the molecule from which the crystal is formed. These molecules, that is, the building block of these organic materials, cover 6 chemical species, including C, H, O, N, F, and S, and span multiple organic chemistries, for example, linear, cyclo-group, and mixed functional-group molecules. The sublimation enthalpy ΔH_{sub} was computed using DFT for all 845 organic crystals.

Among the 845 molecular crystals considered in this work, experimental data of ΔH_{sub} can be found for 28 of them from the National Institute of Standards and Technology (NIST) Chemistry WebBook,^[57] spanning from ≈ 25 kJ/mol to ≈ 190 kJ/mol. Comparing the computed and measured ΔH_{sub} for these 28 materials, the coefficient of determination R^2 is 0.75, establishing reasonable confidence in the DFT computations. For the particular case of solid benzene, the computed ΔH_{sub} is 48.8 kJ/mol, falling well within the reported range of the measured ΔH_{sub} , which is 46.9 – 62.8 kJ/mol.^[57] The satisfactory agreement between the computed and the measured values of ΔH_{sub} demonstrated in Figure 3 serves as a validation of the DFT computational scheme we used in this work.

3.2 | ML model of sublimation enthalpy for organic molecules

In order to understand the effect of training set size on the prediction accuracy before building the final predictive model, various models were generated using increasing training set sizes from 10% to 90% through random selection from the 845 DFT dataset. 10 models were developed for each training set size and the average RMSE, MAE, and standard deviation (error bars) were calculated for all 10 models. The learning curve based on the results from this process is demonstrated in Figure 4a, which evaluates the performance of developed models. As expected, the test

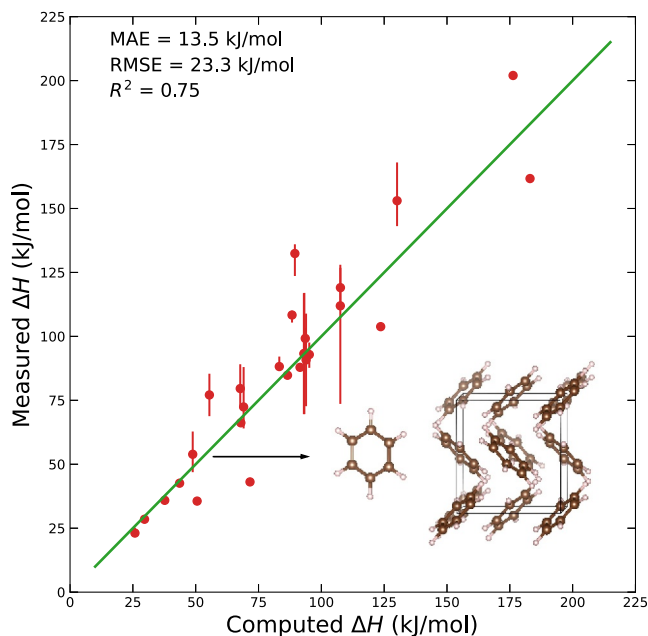


FIGURE 3 Calculated values of ΔH_{sub} , given with respect to the experimental data reported for 28 organic molecular crystals. Error bars are given for cases having multiple reported experimental values.

RMSE and MAE of the ML model decreased with increasing training set size until about 80%–90%.^[58]

Note the gap of about ≈ 5 kJ/mol for RMSE between the training and the test curves, which hints that the features of the materials offered by RDkit may not be sufficient to fully represent the original data (in terms of SMILES). This is, in fact, a common situation in materials informatics, where current methods of data representation are incomplete. Further, it is common in ML to see a drop in performance between training and testing sets, particularly when the training set is relatively small. The data points in the training

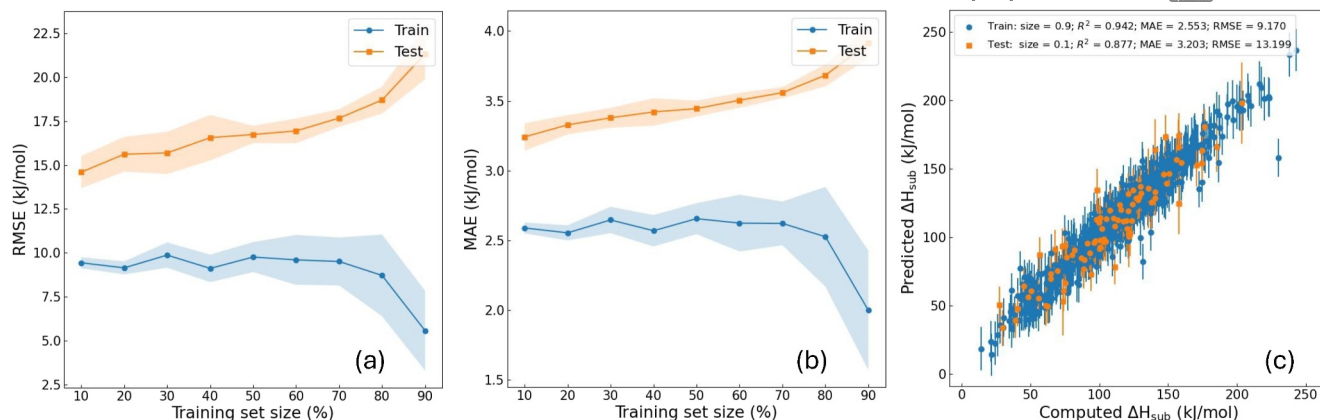


FIGURE 4 Prediction accuracy (a) RMSE and (b) MAE for ML models trained using different training set sizes, averaged over 10 runs. The corresponding test set sizes are equal to the difference between the total training dataset (845) and the training set size. (c) Parity plot obtained from the ML model (208 features) with the train and test size of 90% and 10%, respectively.

set are not fully representative of those in the test set. The ML model provides a test RMSE of 14 kJ/mol (<10% of absolute mean ΔH_{sub}) when 90% of the DFT dataset was used to train the model, which shows that our trained model generalizes to new data points.

A parity plot of DFT-computed ΔH_{sub} versus ML-predicted ΔH_{sub} is shown in Figure 4b. The error bars in the plot represent the GPR uncertainty. The average CV test error (RMSE) of the prediction model that uses the full dataset is 15.24 kJ/mol. In general, Figure 4 indicates that the ML model trained by our DFT dataset accurately predicted the ΔH_{sub} for new organics and could act as a tool for predicting the ΔH_{sub} for novel organic structures.

3.3 | Feature importance analysis

In this section, we used SHAP,^[55] a game theoretic approach, to explain the predictions of ΔH_{sub} in terms of contributions from each feature. Figure 5a ranks the 12 most important features, whereas Figure 5b reveals the way each of them contributes to the predictions of ΔH_{sub} . Five of the 12 features that involve partial charge are “MaxPartialCharge”, “MaxAbsPartialCharge”, “MinAbsPartialCharge”, “MinPartialCharge” and “BCUT2D_CHGHI”. The most important feature is “MaxPartialCharge”, which, according to Figure 5b, has a strong and positive correlation with ΔH_{sub} . This feature contributes strongly and positively to ΔH_{sub} when its value is high, whereas the contribution is low on its low values. Because inter-molecular interactions are electrostatic in nature, high partial charge intensifies these interactions and ultimately raises ΔH_{sub} . Overall, partial charge and any features involving partial charge are critical for ΔH_{sub} .

The second important feature is “FractionCSP3”, which quantifies the fraction of sp^3 hybridized carbon atoms. The rationale behind this observation is that sp^3 hybridized carbons have the lowest electronegativity, so a high fraction of

sp^3 carbon atoms leads to weak inter-molecular interactions, and thus, low ΔH_{sub} . Therefore, “FractionCSP3” correlates negatively to ΔH_{sub} , as revealed by Figure 5b. In addition, “fr_NH1” (number of secondary amines) and “NHOH-Count” (the number of NHs or OHs) are also important features for ΔH_{sub} prediction as higher ΔH_{sub} values contain more of these types of functional groups. The next important feature is “FpDensityMorgan1”, one of three Morgan density fingerprints^[59,60] that show up in the list. “FpDensityMorgan1” measures the local density within the smallest radius used by the Morgan algorithm^[59] while the other two features, that is, “FpDensityMorgan2” and “FpDensityMorgan3”, correspond to the next larger radii. The local density encoded in “FpDensityMorgan1”, “FpDensityMorgan2”, and “FpDensityMorgan3” is important because it directly correlates to the amount of charge that determines the electrostatic (inter-molecular) interactions, according to Coulomb's law.

In summary, the features related to partial charges are critical for determining ΔH_{sub} because of the Coulombic nature of this concept. Going forward, the feature importance analysis presented in this work can be used in the future as precursor design rules for targeted ΔH_{sub} .

3.4 | Active learning

Active learning is an efficient strategy to systematically explore the chemical space, expand and diversify the training set, and ultimately improve the ML model. We show how active learning could be used to enhance our ML model by making use of the dataset described above consisting of 845 organic crystals.

Figure 6 visualizes the training data at three different steps of plan B by projecting them onto the 2D space spanned by PC1 and PC2, the two first principal axes, which are obtained by a principal component analysis (PCA) and capture the most variance of the training data. In this visual

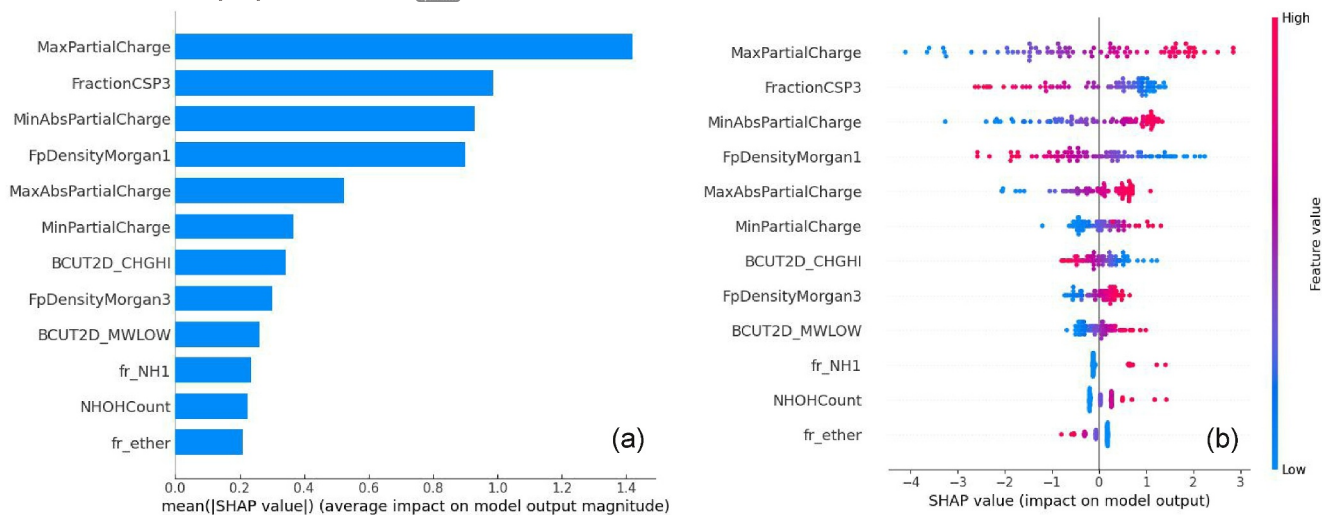


FIGURE 5 (a) Important values assigned to features based on SHAP. The names of the features are shown on the y-axis, while their importance is shown on the x-axis. (b) The contributions of each feature as compared to the average model prediction. The y-axis on the right side indicates the respective feature value being low versus high. Each dot represents one instance in the data.

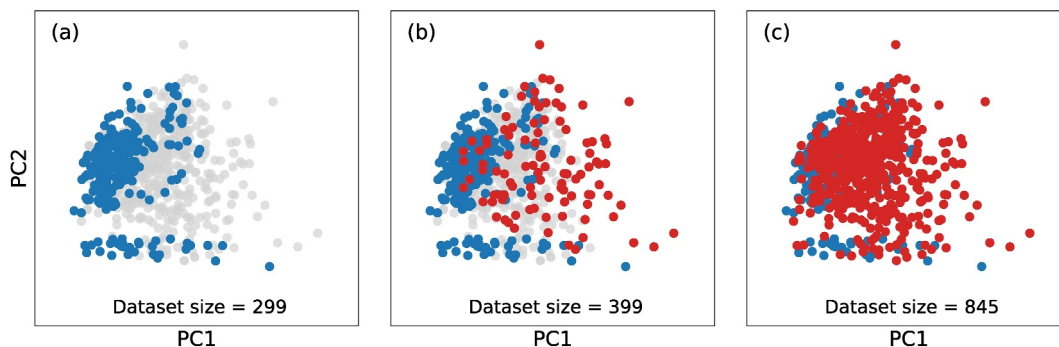


FIGURE 6 (Projections of the training dataset (blue and red circles) at the (a) initial step, (b) an intermediate step, and (c) the final step of plan B onto the 2D manifold spanned by PC1 and PC2, two first principal axes obtained from a principal component analysis (PCA). Red circles highlight the new points that are added during the active learning process. Gray circles represent the entire dataset of 845 organic crystals.

representation, the coverage of the materials space is systematically and efficiently expanded/diversified when the active learning strategy using prediction uncertainty is performed.

The expansion/diversification of the training data as demonstrated in Figure 6 is critically important for improving the ML model. In Figure 7, we quantitatively visualized the performance of the ML models trained at step 0 (whose training set of 299 data points) and step 1 (whose training set of 302 data points) of plan B. At step 0, the training set contains only C, H, N, and O species, thus all the predictions of ΔH_{sub} on S-containing molecules (Figure 7b) are highly erroneous with an RMSE of 131.1 kJ/mol. In fact, most of the predictions for data points containing S or F chemical species return $\Delta H_{\text{sub}} \approx 0$ kJ/mol with very large uncertainties (≈ 80 kJ/mol). At step 1, three new points were added to the training set, one of them has S species and the others have no S (Figure 7c). Figure 7d shows that by having one material that has S species, the RMSE of the predictions

of ΔH_{sub} on S-containing materials is reduced to a third of the initial value to 41.1 kJ/mol, whereas the presence of two new materials without S in the training set also reduces the RMSE of the predictions on the materials involving no S from 35 kJ/mol to 27.2 kJ/mol. Figure 7a–d clearly indicate that augmenting underrepresented cases in the training data is extremely important in expanding the domain of applicability and advancing the power of the ML models. It should be noted that in each step, three data points were selected simultaneously. Although a sequential selection method was also evaluated, its performance was similar to the approach described in Plan B, as discussed in this section.

Active learning is known to be an efficient strategy to augment the training data in an informed manner, targeting specifically underrepresented cases, and ultimately improving the performance of the ML model.^[42–44] In (the baseline) plan A, augmented data were selected entirely randomly without any information on whether they are

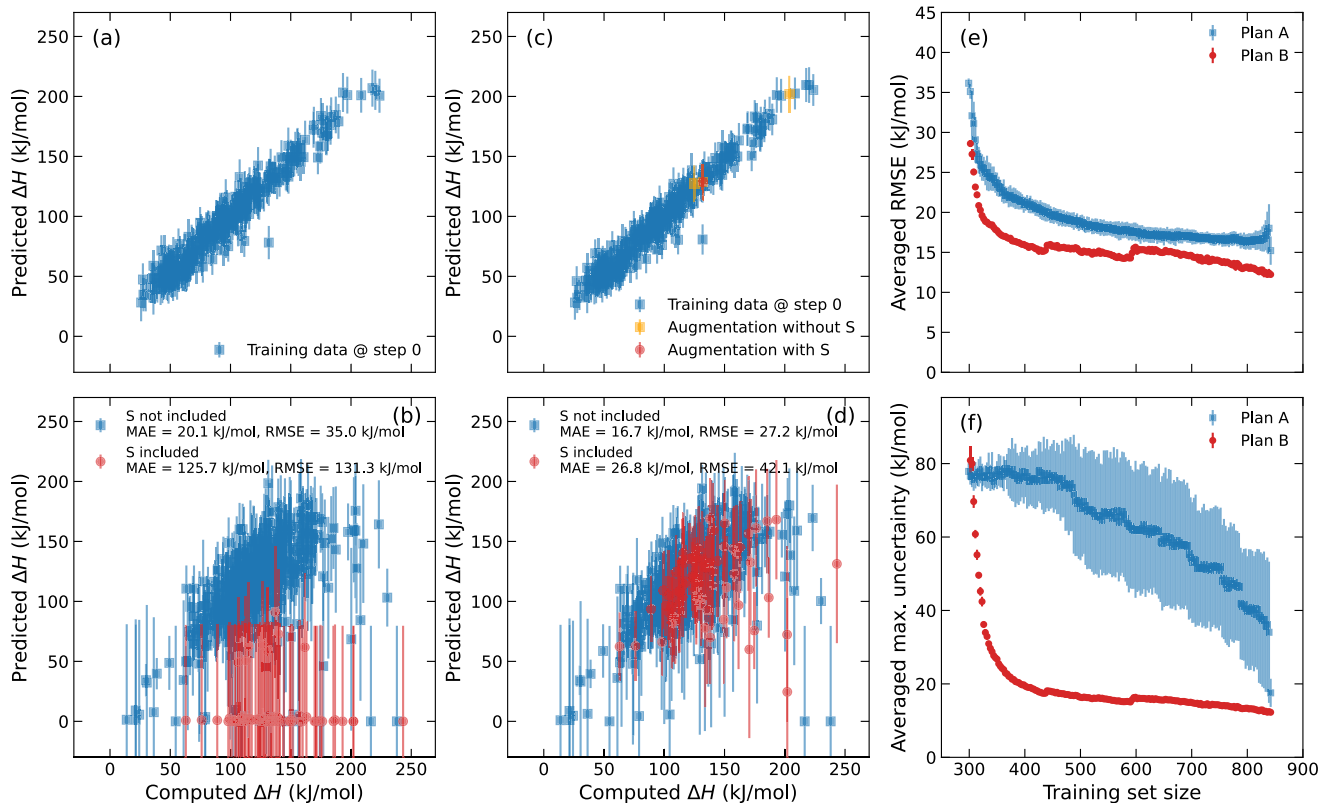


FIGURE 7 ML models of ΔH_{sub} , training at step 0 (a) and (b) and step 1 (c) and (d) of the active learning strategy employing plan B of data augmentation. Panels (a) and (c) visualize the training process while panels (b) and (d) visualize the predictions of the ML models on the test data. In panels (e) and (f), the maximum RMSE and uncertainty of the predictions in the test set are shown for plan A and plan B, respectively. Each data point is obtained by averaging over the results obtained from 10 models trained independently on a given training set.

underrepresented. In contrast, the selection of augmented data in plan B relies on the prediction uncertainties. The rationale of plan B is that a small (or large) value of the prediction uncertainty indicates that the material under consideration is well- (or under-) represented in the training set. Therefore, three data points with the largest uncertainties, that is, they are under-represented in the training data, are selected from the test set to be advanced to DFT computations of ΔH_{sub} .

The superiority of active learning in improving ML models is further demonstrated in Figure 7e,f. For each plan A or B, the RMSE and the maximum uncertainty of the predictions on the test set are shown. Each measure was averaged over 10 models that were trained independently on a given training set. Although the averaged RMSE provides an un-biased assessment of the performance of the ML model trained, the averaged maximum uncertainty signals the ML model's confidence in making the prediction on a new case. Figure 7e,f show that the augmented data points identified by active learning can help to reduce both mean RMSE and maximum uncertainty very quickly, that is, significantly improving the ML models. When the training set size is ≈ 400 , the training data selected by the active learning strategy, which is based on the prediction uncertainty, are spread out very well through the entire materials space considered in this work (see Figure 6). At this training

set size, the averaged RMSE and maximum uncertainty of plan B are saturated at ≈ 15 kJ/mol and ≈ 20 kJ/mol, respectively, the values plan A can only reach when the whole materials space is included in the training set, that is, the training set size is 845.

In summary, Figures 6 and 7 show that active learning is a powerful strategy to quickly select the cases of interest, creating a dataset that is a good sample of the considered materials space, and quickly improve the ML model developed to predict ΔH_{sub} . This observation is important when the materials space becomes big, that is, containing millions of materials, thus minimizing the number of expensive DFT ΔH_{sub} computations while maximizing the performance of the ML model is a critical requirement.

4 | CONCLUSIONS

In summary, we have successfully developed an accurate DFT-based scheme for computing the ΔH_{sub} of organic molecular crystals. Leveraging the computed ΔH_{sub} dataset, we have constructed a predictive ML model capable of accurately predicting ΔH_{sub} with a predictive accuracy of approximately 15 kJ/mol. Although this important parameter remains higher than the chemical accuracy, it will be lowered in the next phase of the model development when more data

of ΔH_{sub} are computed and used in the ML training process. Furthermore, through feature importance analysis, we have identified key features that significantly influence ΔH_{sub} , enabling their direct application in the selection and design of materials.

During the development phase of the computed ΔH_{sub} dataset, we have demonstrated the superiority of active learning in identifying underrepresented species. This approach enables iterative data augmentation through targeted DFT computations, resulting in the systematic and efficient improvement of the ML model. Going forward, active learning will continue to be employed to progressively expand the computed ΔH_{sub} dataset, further enhancing the performance and capabilities of the ML models.

The combination of our accurate DFT-based scheme, the predictive ML model for ΔH_{sub} , and the active learning approach opens up new possibilities for advancing our understanding of sublimation enthalpy and its impact on organic molecular crystals. These findings provide valuable insights for the selection and design of organic precursors and hold great potential for driving advancements in materials science and engineering.

AUTHOR CONTRIBUTIONS

Yifan Liu: Investigation; Writing - original draft; Data curation; Formal analysis; Methodology; Software; Validation. **Huan Tran:** Writing - review and editing; Supervision; Conceptualization. **Chaofan Huang:** Writing - review and editing. **Beatriz G. del Rio:** Writing - review and editing; Supervision. **V. Roshan Joseph:** Project administration; Conceptualization; Writing - review and editing; Supervision; Funding acquisition. **Mark Losego:** Conceptualization; Writing - review and editing; Project administration; Supervision; Funding acquisition. **Rampi Ramprasad:** Funding acquisition; Conceptualization; Writing - review and editing; Project administration; Supervision.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (DMREF-1921873). The DFT dataset was generated with computational support from XSEDE through two allocations DMR080058N and DMR170031. The computed ΔH_{sub} dataset can be accessed at our repository <https://khazana.gatech.edu/dataset/>. The code and data are available on GitHub https://github.com/Ramprasad-Group/Sublimation_enthalpy_model.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Khazana at <https://khazana.gatech.edu/dataset/>.

ORCID

Yifan Liu  <https://orcid.org/0000-0001-5102-0552>

REFERENCES

- Poling BE, Prausnitz JM, O'Connell JP. *The Properties of Gases and Liquids*. McGraw-Hill Education; 2001.
- Acree W, Chickos JS. Phase transition enthalpy measurements of organic and organometallic compounds. Sublimation, vaporization and fusion enthalpies from 1880 to 2010. *J Phys Chem Ref Data*. 2010;39(4):043101.
- Červinka C, Fulem M. State-of-the-Art calculations of sublimation enthalpies for selected molecular crystals and their computational uncertainty. *J Chem Theor Comput*. 2017;13(6):2840-2850.
- Campbell CT, Sellers JRV. Enthalpies and entropies of adsorption on well-defined oxide surfaces: experimental measurements. *Chem Rev*. 2013;113(6):4106-4135.
- Leng CZ, Losego MD. Vapor phase infiltration (VPI) for transforming polymers into organic-inorganic hybrid materials: a critical review of current progress and future challenges. *Mater Horiz*. 2017;4(5):747-771.
- Yurata T, Lei H, Tang L, et al. Feasibility and sustainability analyses of carbon dioxide-hydrogen separation via de-sublimation process in comparison with other processes. *Int J Hydrogen Energy*. 2019;44(41):23120-23134.
- McArdle P, Erxleben A. Sublimation – a green route to new solid-state forms. *CrystEngComm*. 2021;23(35):5965-5975.
- Yurata T, Lei H, Tang L, et al. Feasibility and sustainability analyses of carbon dioxide – hydrogen separation via de-sublimation process in comparison with other processes. *Int J Hydrogen Energy*. 2019;44(41):23120-23134.
- Gharagheizi F, Sattari M, Tirandazi B. Prediction of crystal lattice energy using enthalpy of sublimation: a group contribution-based model. *Ind Eng Chem Res*. 2011;50(4):2482-2486.
- Xia F, Jiang L. Bio-inspired, smart, multiscale interfacial materials. *Adv Mater*. 2008;20(15):2842-2858.
- Almeida AR, Monte MJ. A brief review of the methods used to evaluate vapour pressures and sublimation enthalpies. *Struct Chem*. 2013;24(6):1993-1997.
- Chickos JS, Gavezzotti A. Sublimation enthalpies of organic compounds: a very large database with a match to crystal structure determinations and a comparison with lattice energies. *Cryst Growth Des*. 2019;19(11):6566-6576.
- Fulem M, Růžička K, Červinka C, Rocha MA, Santos LM, Berg RF. Recommended vapor pressure and thermophysical data for ferrocene. *J Chem Thermodyn*. 2013;57:530-540.
- Růžička K, Fulem M, Červinka C. Recommended sublimation pressure and enthalpy of benzene. *J Chem Thermodyn*. 2014;68:40-47.
- Delle Site A. The vapor pressure of environmentally significant organic chemicals: a review of methods and data at ambient temperature. *J Phys Chem Ref Data*. 1997;26(1):157-193.
- Růžička K, Koutek B, Fulem M, Hoskovec M. Indirect determination of vapor pressures by capillary gas-liquid chromatography: analysis of the reference vapor-pressure data and their treatment. *J Chem Eng Data*. 2012;57(5):1349-1368.
- Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev*. 1964;136(3B):B864-B871.
- Kohn W, Sham L. Self-consistent equations including exchange and correlation effects. *Phys Rev*. 1965;140(4A):A1133-A1138.
- Vener MV, Levina EO, Koloskov OA, Rykounov AA, Voronin AP, Tsirelson VG. Evaluation of the lattice energy of the two-component molecular crystals using solid-state density functional theory. *Cryst Growth Des*. 2014;14(10):4997-5003.
- Manin AN, Voronin AP, Manin NG, et al. Salicylamide cocrystals: screening, crystal structure, sublimation thermodynamics, dissolution, and solid-state DFT calculations. *J Phys Chem B*. 2014;118(24):6803-6814.
- Manin AN, Voronin AP, Shishkina AV, Vener MV, Churakov AV, Perlovich GL. Influence of secondary interactions on the structure, sublimation thermodynamics, and solubility of salicylate: 4-hydroxybenzamide cocrystals. combined experimental and theoretical study. *J Phys Chem B*. 2015;119(33):10466-10477.
- Motalov VB, Korobov MA, Dunaev AM, Dunaeva VV, Tyunina EY, Kudin LS. Refined data on the sublimation enthalpy and thermodynamic functions of l-and dl-methionine. *J Chem Eng Data*. 2022;67(6):1326-1334.

23. Levina EO, Chernyshov IY, Voronin AP, Alekseiko LN, Stash AI, Vener MV. Solving the enigma of weak fluorine contacts in the solid state: a periodic DFT study of fluorinated organic crystals. *RSC Adv*. 2019;9(22):12520-12537.
24. Voronin AP, Perlovich GL, Vener MV. Effects of the crystal structure and thermodynamic stability on solubility of bioactive compounds: DFT study of isoniazid cocrystals. *Comput Theor Chem*. 2016;1092:1-11.
25. Tsuzuki S, Orita H, Honda K, Mikami M. First-principles lattice energy calculation of urea and hexamine crystals by a combination of periodic DFT and MP2 two-body interaction energy calculations. *J Phys Chem B*. 2010;114(20):6799-6805.
26. Lee K, Murray ED, Kong L, Lundqvist BI, Langreth DC. Higher-accuracy van der Waals density functional. *Phys Rev B*. 2010;82(8):081101.
27. Woods LM, Dalvit DAR, Tkatchenko A, Rodriguez-Lopez P, Rodriguez AW, Podgornik R. Materials perspective on Casimir and van der Waals interactions. *Rev Mod Phys*. 2016;88(4):045003.
28. Huan TD, Ramprasad R. Polymer structure predictions from first principles. *J Phys Chem Lett*. 2020;11(15):5823-5829.
29. Sahu H, Shen K-H, Montoya J, Tran H, Ramprasad R. Polymer structure predictor (psp): a python toolkit for predicting atomic-level structural models for a range of polymer geometries. *J Chem Theor Comput*. 2022;18(4):2737-2748.
30. Politzer P, Murray JS, Edward Grice M, Desalvo M, Miller E. Calculation of heats of sublimation and solid phase heats of formation. *Mol Phys*. 1997;91(5):923-928.
31. Politzer P, Ma Y, Lane P, Concha MC. Computational prediction of standard gas, liquid, and solid-phase heats of formation and heats of vaporization and sublimation. *Int J Quant Chem*. 2005;105(4):341-347.
32. Byrd EF, Rice BM. Improved prediction of heats of formation of energetic materials using quantum mechanical calculations. *J Phys Chem A*. 2006;110(3):1005-1013.
33. Gharagheizi F, Ilani-Kashkoul P, Acree WE, Mohammadi AH, Ramjugernath D. A group contribution model for determining the sublimation enthalpy of organic compounds at the standard reference temperature of 298 K. *Fluid Phase Equil*. 2013;354:265-285.
34. Liu R, Tang Y, Tian J, et al. QSPR models for sublimation enthalpy of energetic compounds. *Chem Eng J*. 2023;474:145725.
35. Mathieu D. Simple alternative to neural networks for predicting sublimation enthalpies from fragment contributions. *Ind Eng Chem Res*. 2012;51(6):2814-2819.
36. Suntsova MA, Dorofeeva OV. Prediction of enthalpies of sublimation of high-nitrogen energetic compounds: modified Politzer model. *J Mol Graph Model*. 2017;72:220-228.
37. Landrum G. Others RDKit: open-source cheminformatics. 2006. <http://www.rdkit.org>
38. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning and materials informatics: recent applications and prospects. *npj Comput Mater*. 2017;3(1):54.
39. Chen L, Pilania G, Batra R, et al. Polymer informatics: current status and critical next steps. *Mater Sci Eng R Rep*. 2021;144:100595.
40. Tran H, Kim C, Chen L, et al. Machine-learning predictions of polymer properties with Polymer Genome. *J Appl Phys*. 2020;128(17):171104.
41. Kim C, Chandrasekaran A, Huan TD, Das D, Ramprasad R. Polymer genome: a data-powered polymer informatics platform for property predictions. *J Phys Chem C*. 2018;122(31):17575-17585.
42. Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput Mater*. 2019;5(1):21.
43. Kim C, Chandrasekaran A, Jha A, Ramprasad R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun*. 2019;9(3):860-866.
44. Huan TD, Batra R, Chapman J, Kim C, Chandrasekaran A, Ramprasad R. Iterative-learning strategy for the development of application-specific atomistic force fields. *J Phys Chem C*. 2019;123(34):20715-20722.
45. Kresse G, Hafner J. Ab initio molecular dynamics for liquid metals. *Phys Rev B*. 1993;47(1):558-561.
46. Kresse G, Furthmüller J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput Mater Sci*. 1996;6(1):15-50.
47. Kresse G. *Ab initio Molekular Dynamik für flüssige Metalle* Ph.D. thesis. Technische Universität Wien; 1993.
48. Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys Rev B*. 1996;54(16):11169-11186.
49. Vaitkus A, Merkys A, Gražulis S. Validation of the Crystallography open database using the crystallographic information framework. *J Appl Crystallogr*. 2021;54(2):661-672.
50. Quirós M, Gražulis S, Girdzijauskaitė S, Merkys A, Vaitkus A. Using SMILES strings for the description of chemical connectivity in the Crystallography Open Database. *J Cheminf*. 2018;10(1):23.
51. Merkys A, Vaitkus A, Butkus J, Okulič-Kazarinas M, Kairys V, Gražulis S. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *J Appl Crystallogr*. 2016;49(1):292-301.
52. Perdew JP, Burke K, Ernzerhof M. Generalized gradient approximation made Simple. *Phys Rev Lett*. 1996;77(18):3865-3868.
53. Ong SP, Richards WD, Jain A, et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput Mater Sci*. 2013;68:314-319.
54. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci*. 1988;28:31-36.
55. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017.
56. Rasmussen CE, C. K. I. W. *Gaussian Processes for Machine Learning*. the MIT Press: Massachusetts Institute of Technology; 2006.
57. William E, Acree JSC, Jr. NIST chemistry WebBook. In: Linstrom PJ, Mallard WG, eds. *NIST Standard Reference Database Number 69*. National Institute of Standards and Technology; 2022.
58. Joseph VR. Optimal ratio for data splitting. *Stat Anal Data Min ASA Data Sci J*. 2022;15(4):531-538.
59. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem Doc*. 1965;5(2):107-113.
60. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742-754.

How to cite this article: Liu Y, Tran H, Huang C, et al. Accelerated predictions of the sublimation enthalpy of organic materials with machine learning. *MGE Advances*. 2025;e84. <https://doi.org/10.1002/mgea.84>