Machine Learning Transition Temperatures from 2D Structure

Andrew E. Sifain,[†] Samuel H. Yalkowsky,[‡] Betsy M. Rice,[†] and Brian C. Barnes[*,†]

†*CCDC U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA*

‡*Department of Pharmaceutics, College of Pharmacy, University of Arizona, Tucson, AZ 85721, USA*

E-mail: brian.c.barnes11.civ@mail.mil

Abstract

A priori knowledge of melting and boiling could expedite the discovery of pharmaceutical, energetic, and energy harvesting materials. The tools of data science are becoming increasingly important for exploring chemical datasets and predicting material properties. A fundamental part of data-driven modeling is molecular featurization. Herein, we propose a molecular representation with group-constitutive and geometrical descriptors that map to enthalpy and entropy–two thermodynamic quantities that drive thermal phase transitions. The descriptors are inspired by the linear regression-based quantitative structure-property relationship of Yalkowsky and coworkers known as the Unified Physicochemical Property Estimation Relationships (UPPER). Combined with nonlinear machine learning (specifically, eXtreme Gradient Boosting or XGBoost), these concise and easy-to-compute descriptors provide an appealing framework for predicting transition enthalpies, entropies, and temperatures in a diverse chemical space. An application to energetic materials shows that UPPER plus XGBoost is predictive, despite a relatively modest energetics reference dataset. We also report results on public datasets of melting points (*i.e.*, OCHEM, Enamine, Bradley, and Bergström). The

newly proposed representation is determined purely from SMILES string, thus showing promise toward fast and accurate screening of thermodynamic properties.

Introduction

Transition temperatures such as melting and boiling point are fundamental thermodynamic properties that influence applications including the design of pharmaceuticals,[1] melt-casted explosives,[2–4] and energy harvesting materials.[5,6] Discovering materials with acceptable transition temperatures is difficult, in part because they are not known prior to synthesis and measurement. Theoretical prediction of such properties may reduce the chemical space of candidate compounds and expedite discovery.

Atomistic simulations of phase transitions are computationally demanding. Furthermore, in the case of melting, such simulations often require knowledge of crystal structure,[7–10] thus limiting their use during materials discovery. An alternative approach is to utilize surrogate models that map descriptors to reference data. Linear regression-based quantitative structure-property relationships (QSPRs) have had success,[11–16] but are inadequate for finding nonlinear mappings between descriptors and melting point. Nonlinear machine learning (ML) algorithms overcome this shortcoming[17–20] and possess other advantages such as transferability to species outside of the reference dataset and computational efficiency.[21–23] An accurate ML model may help identify target compounds and circumvent expensive atomistic simulations.

ML is becoming an indispensable and versatile tool in the chemical sciences[24] with application to molecular properties,[25–31] spectroscopy,[32–36] and chemical synthesis.[37–40] Model performance depends concomitantly on the learning algorithm, the quality of the reference training set, and the input representation of the chemical system.[41–47] The focus of this paper is on the design of a molecular representation from a microscopic basis in order to predict macroscopic properties such as melting and boiling point.

The molecular representation reflects the level of chemical resolution needed for predicting the target property.[48,49] For quantum molecular properties, geometry (and atom type) are typically chosen as input because even the slightest changes in geometry can affect the wavefunction and its observables.[50–53] On the other hand, macroscopic properties are more robust to higher-level or coarse-gained descriptors.[54–58] In the case of melting, for example, a molecular crystal may be identified by descriptors or a molecular fingerprint[59–62] derived from its repeating structural unit.[20] The mapping of a multi-molecule process from a single molecule is ambiguous however, thus emphasizing the importance of suitable descriptors.

In this paper, we propose descriptors based on the UPPER method of Yalkowsky and coworkers.[63] UPPER, which stands for Unified Physicochemical Property Estimation Relationships, is a comprehensive QSPR based on intuitive and thermodynamic relationships relating phase transition properties to one another including transition enthalpies, entropies, and temperatures. The method's elegance is that properties are related to group-constitutive and geometrical descriptors, determined purely from 2D structure (*i.e.*, Simplified Molecular-Input Line-Entry System or SMILES[64]). While it is generally a challenge to train models to limited experimental data, we find that a concise set of domain-specific descriptors, combined with nonlinear ML algorithms (specifically, gradient boosting[65]), provides an appealing framework for predicting transition enthalpies, entropies, and temperatures in a diverse chemical space. Our software is freely available at https://github.com/USArmyResearchLab/ARL-UPPER.

Methods

We overview the UPPER method for transition properties and the underlying descriptors that will be supplied as input for ML. The addition of heat to a thermodynamic system increases its temperature. When a first-order phase transition occurs, the temperature levels off, remaining constant even as the system continues absorbing heat. Intermolecular binding

forces are overcome as heat converts the state of the system from solid to liquid or liquid to gas. When the two phases of matter are in equilibrium with one another, the Gibbs energy is $\Delta G_{tr} = 0$. The first-order transition temperature can therefore be written as

$$T_{tr} = \frac{\Delta H_{tr}}{\Delta S_{tr}}, \tag{1}$$

where ΔH_{tr} and ΔS_{tr} are enthalpy and entropy of transition, respectively. Here, ΔH_{tr} is the amount of heat absorbed per mole for the transition to take place, while ΔS_{tr} is the change in the system's entropy. UPPER defines analytical forms for ΔH_{tr} and ΔS_{tr} with parameters determined using separate linear regression analysis of composition for ΔH_{tr} and geometry for ΔS_{tr}

Group-Constitutive Descriptors

Within UPPER, enthalpy is computed as a group-constitutive property,

$$\Delta H_{tr} = \sum_i p_i n_i, \tag{2}$$

where p_i is the contribution of the i-th fragment and n_i is the number of i fragments in the molecule. Fragmentation is based on the scheme proposed in Ref. [66], where each fragment consists of the least number of atoms (including all carbons, hydrogens, heteroatoms, and nonbonded electrons) that are not separated by an isolating carbon. An isolating carbon is a carbon that is not doubly or triply bonded to a heteroatom. Such carbons and their attached hydrogens are considered hydrophobic fragments with the remaining groups of atoms being polar fragments. Fragments are represented by their SMARTS (SMiles ARbitrary Target Specification) strings and subsequently assigned to an environmental group (Fig. 1). Environmental groups (Table 1) reflect interactions such as the connectivity and hybridization that each fragment has with its neighboring fragments.

Table 1: Environmental groups.

Group	Description
X	Group bonded to only sp^3 atoms
Y	Group singly bonded to 1 sp^2 atom
YY	Group bonded to 2 sp^2 atoms
YYY	Group bonded to 3 sp^2 atoms
YYYY	Group bonded to 4 sp^2 atoms
Z	Group bonded to 1 sp atom
YZ	Y and Z group
YYZ	YY and Z group
YYYZ	YYY and Z group
RG	Group within an aliphatic ring
FU	Aliphatic bridge-head group
AR	Group within an aromatic ring
BR2	Aromatic carbon shared by 2 rings
BR3	Aromatic carbon shared by 3 rings
BIP	Central carbon in biphenyl substructure

Geometrical Descriptors

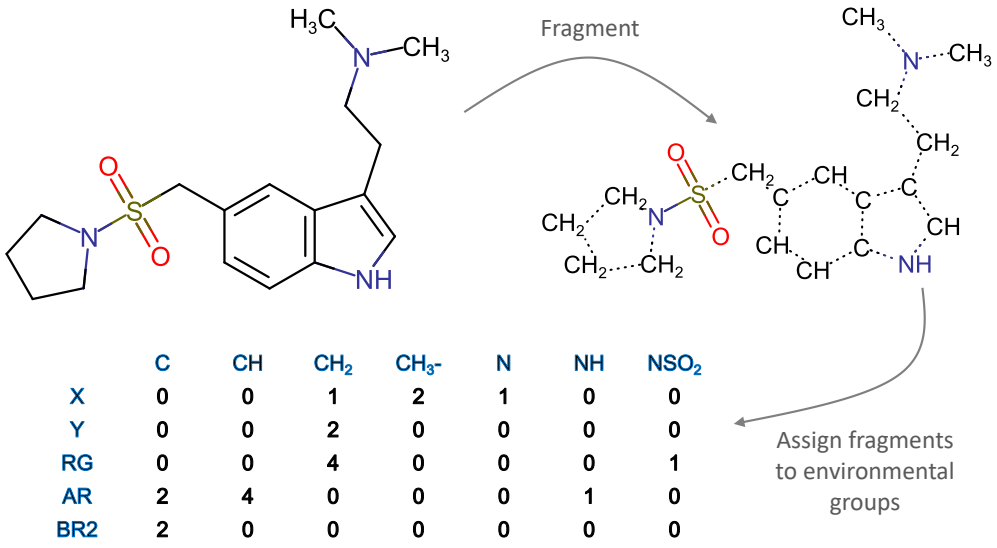
Entropy depends on molecular geometry and encodes translational, conformational, and rotational changes of a molecule that affect properties such as packing efficiency and the likeliness for initial and final states of a phase transition to exist. Entropy is given by

$$\Delta S_{tr} = \Delta S_{tr}^{trans} + \Delta S_{tr}^{conf} + \Delta S_{tr}^{rot}. \tag{3}$$

Components of ΔS_{tr} are computed from geometrical descriptors (descriptions below): eccentricity (ϵ), flexibility (ϕ), and symmetry (σ).

Eccentricity is computed as the sum of atoms in and directly attached to aromatic rings. It is a measure of the packing efficiency of a molecular crystal. Crystals with flat molecules tend to have less than average free volume due to their efficient packing, requiring more energy and a higher temperature to melt. Conversely, crystals made up of spherical molecules pack less efficiently and are more prone to attaining their free rotation.

Flexibility is a measure of the internal conformational freedom of a molecule. Flexible molecules tend to have a greater entropy change during melting than rigid molecules. In



	C	CH	CH_2	CH_3-	N	NH	NSO_2
X	0	0	1	2	1	0	0
Y	0	0	2	0	0	0	0
RG	0	0	4	0	0	0	1
AR	2	4	0	0	0	1	0
BR2	2	0	0	0	0	0	0

Figure 1: 2D structure of almotriptan molecule fragmented according to the isolating carbon method. Dashed lines represent broken bonds. Fragments (actually labeled by SMARTS strings in software to avoid ambiguity) are then assigned to environmental groups.

UPPER, flexibility is calculated by an *ad hoc* expression uniting flexible segments,

$$\phi = 0.3\text{ROT} + \text{LINSP3} + 0.5\,(\text{BRSP3} + \text{SP2} + \text{RING}) - 1 \tag{4}$$

where LINSP3 is the number of nonring, nonterminal, nonbranched sp^3 atoms, ROT is the extra entropy produced by freely rotating sp^3 atoms and is calculated as ROT = LINSP3 − 4, BRSP3 is the total number of nonring, nonterminal, branched sp^3 atoms, SP2 is the number of nonring, nonterminal sp^2 atoms, RING is the number of single, fused, or conjugated ring systems. Compounds with negative ϕ computed using Eq. 4 are assigned ϕ equal to zero.

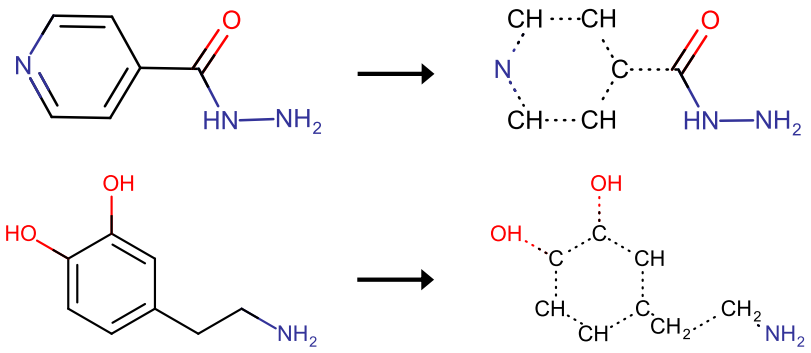
Symmetry affects entropy and in particular the melting point. Symmetric molecules have a higher probability of being in the right orientation for crystallization than nonsymmetrical molecules (of roughly the same weight). As a result, they tend to have a lower entropy of melting and higher melting point. Here, the method to compute σ (see Ref. [67]) operates by locating the center or centers of graphical symmetry and the equivalence classes of atoms connected to those centers.[1] σ is estimated based on a few simple rules determined by the

[1]The centers of graphical symmetry are atoms that are most symmetrical with respect to connections to

hybridization of the graphical center as well as the number of connected atoms and their equivalence classes.

UPPER-Inspired Fingerprint

A combination of the group-constitutive and geometrical descriptors make up the UPPER-inspired fingerprint. The overall size of the fingerprint depends on the molecules in the dataset, as this affects the types of fragments and environmental groups. Figure 2 shows example fingerprints for two molecules.



XNH_2-	XCH_2	YCON HNH_2-	YCH_2	YOH-	ARC	ARCH	ARN	ϕ	σ	ε
0	0	1	0	0	1	4	1	0.5	1	7
1	1	0	1	2	3	3	0	1.5	1	9

Figure 2: Isoniazid and dopamine molecules fragmented and their UPPER-inspired fingerprints consisting of group-constitutive and geometrical (ϕ, σ, and ϵ) descriptors.

Dataset and Learning Algorithms

Our dataset includes experimental transition enthalpies, entropies, and temperatures for both melting and boiling. Compounds of the dataset include open-chain, aliphatic, and aromatic compounds including polyhalogenated biphenyls, dibenzo-p-dioxins, diphenyl ethers, anisoles, and alkanes. There are a wide variety of functional groups such as alcohol,

other atoms.

aldehyde, ketone, carboxylic acid, carbonate, carbamate, amine, amide, nitrile, acetyl, and nitro groups. See Supporting Information and Ref. [68] for more detail.

Tests are carried out with two different models. Original UPPER is the reference model, where coefficients p_i for enthalpy (Eq. 2) are determined using ridge regression. Ridge regression is a variant of linear regression with regularization to reduce overfitting.[2] Entropies are also parameterized using ridge regression. The second model is a variant of Gradient Boosting (GB) called eXtreme GB or XGBoost.[65] A GB model is an ensemble of decision trees where subsequent trees are trained to the residual error of the preceding tree.[69] XGBoost controls overfitting better than GB, giving it strong performance.[70] An added advantage is XGBoost's computational speed.[71] Training details are provided in Supporting Information.

Results and Discussion

The original UPPER method is compared to the new UPPER-inspired fingerprint plus GB approach (denoted UPPERfp+GB). Models for enthalpy (ΔH_{tr}) and entropy (ΔS_{tr}) are randomly split into 90% for training and 10% for testing. Following enthalpy and entropy, we predict transition temperatures (T_{tr}). Finally, UPPERfp+GB is applied to predict the melting points of energetic materials. Prediction errors are quantified by the Root-Mean-Square-Error (RMSE) and the Mean-Absolute-Error (MAE).

Enthalpy

Fig. 3 shows parity plots of enthalpy of melting (ΔH_m) and boiling (ΔH_b). Only group-constitutive descriptors (Table 1) were used in the models of Fig. 3. The results using UPPER and UPPER+GB are comparable, showing that the nonlinear GB algorithm does not provide any added improvement in predicting ΔH_{tr} given group-constitutive descriptors. The prediction accuracy of ΔH_m is not as strong as that of ΔH_b due to missed intermolecular

[2]Original UPPER was tested using linear and ridge regression. Ridge regression was significantly better on the held-out test sets.

interactions in the crystal and liquid phases, such as hydrogen bonding. These interactions are not as significant during a liquid-to-gas transition since molecules are more spatially separated. As a result, ΔH_b is predicted with greater accuracy (*i.e.*, RMSEs of ∼2 compared to ∼4 kJ/mol) using knowledge of only a single molecular unit.

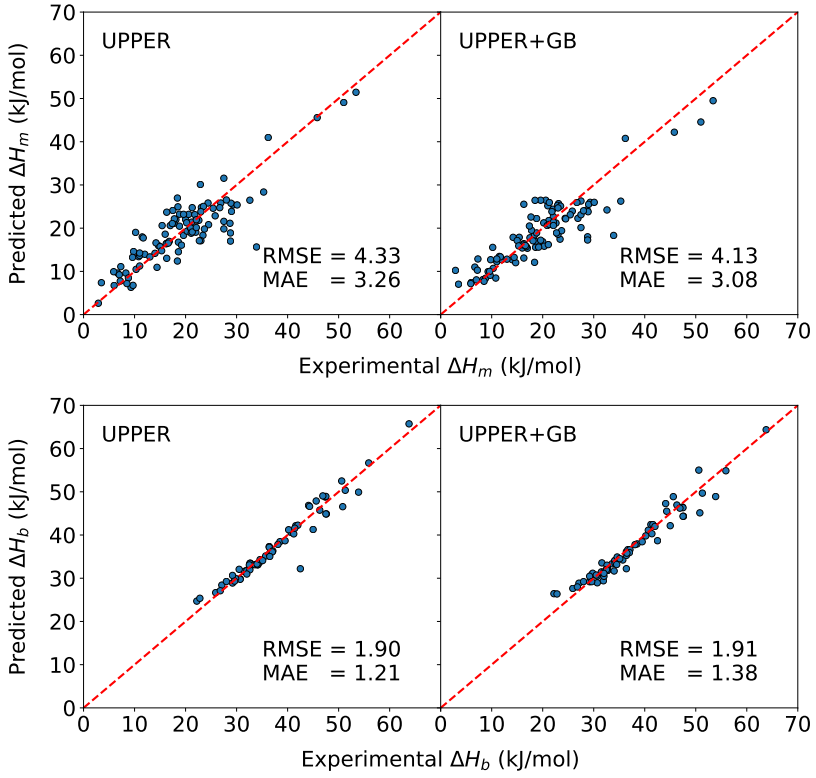


Figure 3: Parity plots of predicted versus experimental ΔH_m (top panels) and ΔH_b (bottom panels) using UPPER (left panels) and UPPER+GB (right panels). Results are of the 10% held-out test set consisting of 108 and 68 molecules for ΔH_m and ΔH_b, respectively. Prediction errors are shown in subpanels. Intermolecular interactions in the crystal-liquid phases (*e.g.*, hydrogen bonding) are not completely accounted for with the group-constitutive descriptors. These missed interactions are likely the cause for the difference in prediction accuracy between ΔH_m and ΔH_b.

Entropy

Parity plots of ΔS_m and ΔS_b are provided in Supporting Information. Similar to ΔH_{tr}, ΔS_{tr} predictions are not significantly improved using the GB model. Trends in the entropy data can be explained by considering physical differences between melting and boiling. For the majority of the data, ΔS_m is smaller than ΔS_b; a consequence of the relative change

of molar volume during a crystal-to-liquid transition versus a liquid-to-gas transition. In particular, boiling produces a volumetric change of usually more than 20 liters per mol, whereas melting produces a smaller change of a few cubic centimeters per mole. Further, ΔS_b data are clumped around $85-90$ J/mol.K (Trouton's rule[3]), whereas ΔS_m shows more variability with a few compounds in the $150-350$ J/mol.K range. The high ΔS_m compounds are long chain-like structures with single bonds that tend to orient themselves in parallel fashion to achieve maximum dispersion in the crystal phase (see Supporting Information). In the liquid phase, their flexible segments have a high degree of conformational freedom. Cross validation results of ΔH_{tr} and ΔS_{tr} are provided in Supporting Information.

Transition Temperatures

Given trained models of ΔH_{tr} and ΔS_{tr}, the ratio of their predictions (Eq. 1) gives T_{tr} (Fig. 4).[4] UPPER+GB slightly outperforms original UPPER, but the overall prediction error is still quite high (RMSEs of $45-55$ K). This result raises the question whether indirectly training to T_{tr} and enforcing Eq. 1 impedes the model's predictive ability. Our new approach (UPPERfp+GB) feeds the entire set of group-constitutive and geometrical descriptors into the GB algorithm and trains the model directly to T_{tr}. In this way, Eq. 1 is not directly enforced, allowing the learning algorithm to choose respective weights over its input features. The parity plots of Fig. 5 suggest that this flexibility is important as RMSEs reduce by about 20 K for T_m and T_b. For reliable evaluation, averaged results over a 10-fold cross validation are provided in Table 2. To identify the added benefit of using a combination of group-constitutive and geometrical descriptors, models were trained solely to group-constitutive descriptors, resulting in slightly increased RMSEs (Table 2).

The UPPERfp+GB approach provides a systematic way of assessing new descriptors. Ref. [72] relates molecular mass (m) to T_m using an expression for atomic vibrations in a

[3]Trouton's rule states that the ratio of the volume of an organic compound as a gas to its volume as a liquid is constant at about 84 J/mol.K.

[4]Due to limited experimental data, ΔH_{tr} and ΔS_{tr} models were trained to all available data. Therefore, results of Fig. 4 are likely biased.

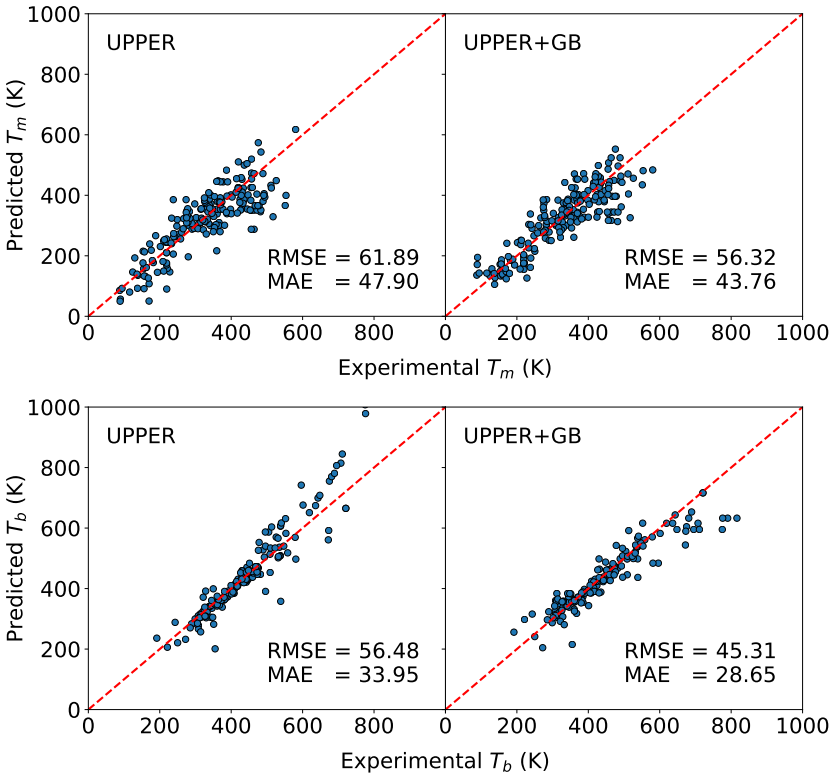


Figure 4: Parity plots of predicted versus experimental T_m (top panels) and T_b (bottom panels) using original UPPER (left panels) and UPPER+GB (right panels). Trained models of ΔH_{tr} and ΔS_{tr} were used to supply predictions of ΔH_{tr} and ΔS_{tr} to compute T_{tr} (Eq.1). Test results consist of 202 and 168 molecules for T_m and T_b, respectively. Prediction errors are shown in subpanels. Despite clear correlations, prediction errors are rather high and could use improvement.

monatomic solid in a thermal environment. The use of m as a descriptor has been shown to improve T_m predictions.[17] Indeed, we find that the cross-validation RMSE of T_m reduces slightly (Table 2). Significant improvement of $\sim$5 K is observed in the case of T_b. Lighter molecules have greater thermal motion than heavier molecules with the same kinetic energy. Thus, lighter molecules boil at lower temperatures, justifying the strong dependence of T_b on m.

The results of UPPERfp+GB are encouraging (Table 2). Nevertheless, the method's descriptors inadequately represent certain compounds. In particular, an example in Supporting Information shows structurally similar compounds with different T_m. Each molecule has an anthracene substructure functionalized by a methyl group, differing only by the methyl's location. Their UPPER-inspired fingerprints are the same, yet T_m of 2-methylanthracene is

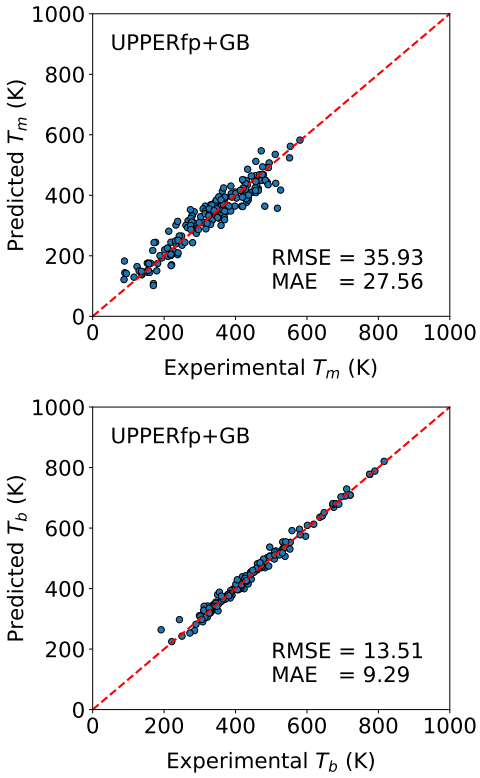


Figure 5: Parity plots of predicted versus experimental T_m (top panel) and T_b (bottom panel) using the new UPPER-inspired fingerprint plus GB approach (UPPERfp+GB). Results are of the 10% held-out test sets consisting of 202 and 168 molecules for T_m and T_b, respectively. Prediction errors are shown in subpanels. Compared to original UPPER and UPPER+GB of Fig. 4, UPPERfp+GB shows significant improvement in predicting T_{tr}.

much larger than 1- and 9-methylanthracene. This significant difference is likely the result of packing arrangement. Unfortunately, packing is difficult to predict from molecular shape, especially 2D structure. We attempted replacing the current 2D eccentricity descriptor with 3D descriptors of eccentricity (ϵ_{3D}) and asphericity (q_{3D}).[73] Cross-validation predictions are slightly improved (Table 2), but while these conformational descriptors help distinguish the methylanthracene compounds, the added information is not sufficient enough for the model to map to their correct T_m's. Thus, new descriptors encoding the effect of molecular shape on intermolecular interactions and subsequent expansion that occurs during melting are needed. The Wiener index–famous for its ability to encode topological information and its strong connection to boiling points of alkanes[74]–was also tested as a descriptor. Prediction errors further improved (Table 2), but overcoming the indistinguishability of the methylanthracene

Table 2: Prediction errors (RMSEs) of T_m, T_b averaged over the 10-fold cross-validation test sets (including standard deviations). Models were trained to combinations of group-constitutive (*gc*), geometrical (*geo*), and mass (m) descriptors. Significant improvement in predictive ability is observed by directly training to T_{tr} (UPPERfp+GB) and adding descriptors that map to T_{tr}.

Method	Descriptors	T_m, T_b
UPPER	$\Delta H_{tr}\,(gc)$, $\Delta H_{tr}\,(geo)$	62.9 ± 4.1, 59.2 ± 6.8
UPPER+GB	$\Delta H_{tr}\,(gc)$, $\Delta H_{tr}\,(geo)$	59.2 ± 5.9, 44.2 ± 5.5
UPPERfp+GB	*gc*	37.9 ± 3.4, 22.3 ± 3.4
UPPERfp+GB	*gc*, *geo*	36.0 ± 3.5, 21.8 ± 4.4
UPPERfp+GB	*gc*, *geo*, m	34.7 ± 2.6, 19.5 ± 5.9
UPPERfp+GB	*gc*, $geo(\epsilon_{3D}, q_{3D})$, m	31.4 ± 2.5, 19.3 ± 4.6
UPPERfp+GB	*gc*, $geo(\epsilon_{3D}, q_{3D})$, m, w	30.6 ± 2.5, 17.2 ± 3.9

compounds remains a challenge (Supporting Information). Quantum-chemical prediction of crystal density[75] relates to packing and may provide useful information. Besides packing, the symmetry descriptor[67] also warrants improvement, as it does not distinguish stereoisomers such as *cis-trans*.

Melting Points of Energetic Materials

Fig. 6 shows an application of the new UPPER-inspired fingerprint to predict T_m of energetic materials containing many nitro groups. The reference data was augmented with energetics, making up about 5% of the entire training set (∼130 compounds). The test set is a diverse set of nitroaliphatic and nitroaromatic compounds including nitropyrimidines and nitropyridines. Prediction accuracies are experimentally informative with overall test set RMSE of 35 K. The model is particularly strong in the case of nitroaliphatic compounds with a RMSE of 25 K. These results give promise toward predictive ML models of exotic energetics given comprehensive datasets.

Melting Points of Public Datasets

Finally, we apply our new approach to train and test on public datasets of T_m (*i.e.*, OCHEM, Enamine, Bradley, and Bergström). Separate models were trained and tested on

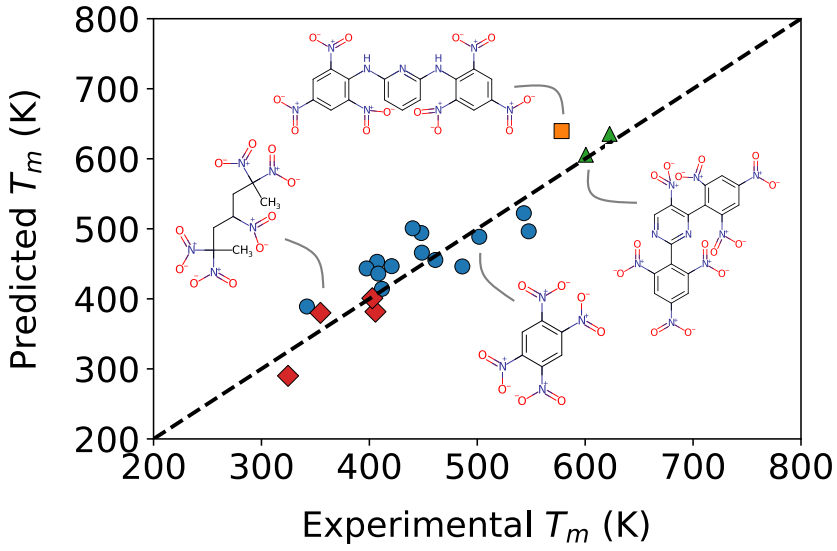


Figure 6: Parity plots of predicted versus experimental T_m of energetic materials. Test set includes nitroaliphatics (◊), nitroaromatics (◯), nitropyridines (□), and nitropyrimidines (△).

each dataset. Due to their relatively small sizes, Bradley and Bergström were combined into one, labeled BradBerg. Further details of the datasets can be found in Refs. [20] and [76]. The top panel of Fig. 7 shows error as a function of temperature for each of the datasets, while the bottom panel shows the distribution of temperatures. Not surprisingly, smallest errors coincide with temperatures that make up the majority of the data. Table 3 reports errors over the middle 50% and 323.15–523.15 K; the latter being a popular range for medicinal compounds.[76] While the 30–40 K RMSE performance is encouraging given the diversity of these datasets, a more thorough curation process would likely benefit applications targeting a specific chemical space. A model's applicable chemical space and accuracy are largely determined by its training dataset.[76] Building a robust model is supported using sufficient and high-quality data. Future work may benefit from advanced sampling techniques such as active learning; a semi-supervised procedure for data generation that interactively queries from a large dataset.[77] As opposed to merely increasing the size of the dataset, active learning finds the right data for the task. It is also important to be mindful of sources of

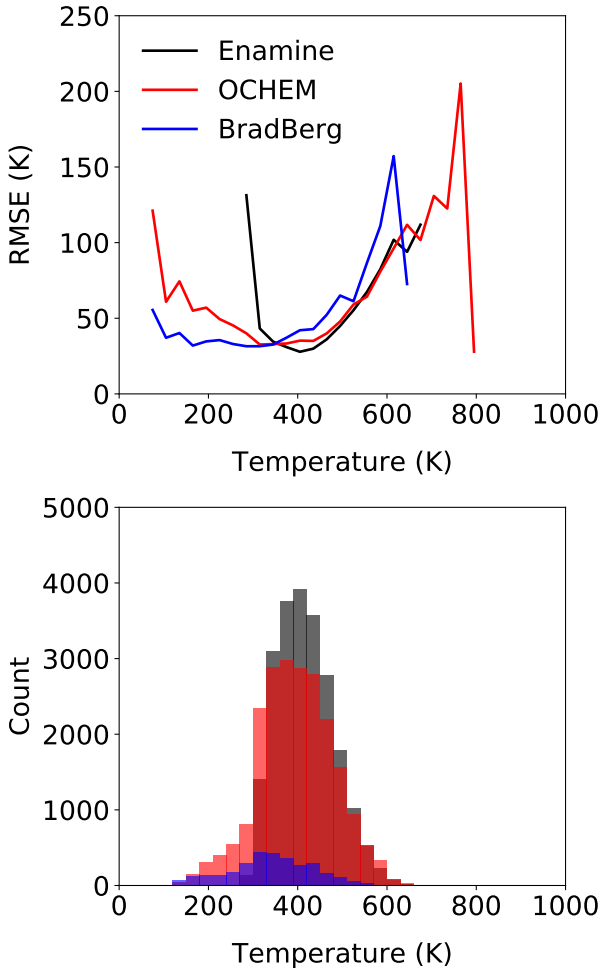


Figure 7: Prediction errors (RMSEs) as a function of melting temperature for each dataset (top panel). Models were trained and tested on each dataset *separately*. Distributions of melting points for each dataset (bottom panel). Temperatures are binned every 30 K. Performance correlates to the density of available data.

error such as experimental conditions, instrumentation, and human error, as trained models cannot overcome this irreducible error.

Conclusion

Our main contribution is a new molecular representation that shows promise toward predicting experimental melting and boiling points of molecular materials. The descriptors originate from a linear regression-based QSPR developed by Yalkowsky and coworkers

Table 3: Prediction errors (RMSEs) of T_m over the 10-fold cross-validation test sets for the middle 50% and 323.15–523.15 K ranges.

Dataset	Size	T_m, middle 50%	T_m, 323.15–523.15 K
Enamine	22381	29.7	34.1
OCHEM	21840	34.4	37.3
BradBerg	3161	34.2	41.8

known as UPPER (see Ref. [63]). UPPER's group-constitutive and geometrical descriptors are used to model enthalpy and entropy; two thermodynamic quantities that drive thermal phase transitions. A notable advantage of UPPER's descriptors is that they are derived purely from SMILES strings. Besides simple structural characteristics such as connectivity and hybridization, there are no numerically intensive calculations necessary. This attribute of the method differs from other molecular representations that use relatively expensive quantum mechanical calculations.[20,76] In this work, we merged an UPPER-inspired fingerprint consisting of group-constitutive and geometrical descriptors with eXtreme Gradient Boosting (denoted UPPERfp+GB).

UPPERfp+GB showed strong predictive ability when tested against a diverse set of compounds. Cross-validation RMSEs of melting and boiling point were found to be 36 and 20 K, respectively (Table 2). Meanwhile, the dataset ranged from about $90 - 700$ K and $150 - 850$ K, comprising a diverse chemical space. The model improved (reducing to 31 and 17 K) with mass, 3D descriptors of eccentricity and asphericity, and topological information using the Wiener index. UPPERfp+GB also provided experimentally informative prediction of melting temperatures of energetic materials, highlighting its transferability to materials containing a significant number of nitro groups compared to the majority of compounds used for training. Our new approach also achieved errors within 30–40 K on melting points of large diverse public datasets.

This work has inspired other projects such as how the new UPPER-inspired fingerprint compares to common molecular fingerprints (see Supporting Information for preliminary calculations). This task goes hand-in-hand with evaluating fingerprints across learning algo-

rithms. Additionally, the intelligent sampling of a training set and more robust descriptors of hydrogen bonding[78] and polymorphs[79] could further improve the model's performance. In its original form, UPPER is a comprehensive QSPR, combining structural information to physicochemical properties including heat of sublimation, solubility, and vapor pressure. Furthering this work could make the UPPER plus ML framework a user-friendly screening tool for the design and discovery of materials in chemistry, physics, and materials science.

Supporting Information Available

Dataset information, training details, model predictions of phase transition properties including transition enthalpies, entropies, and temperatures. An example challenge of the UPPER-inspired fingerprint. Our software is freely available at
https://github.com/USArmyResearchLab/ARL-UPPER

Acknowledgement

Research was sponsored by the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-19-2-0090. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation. This work was supported in part by a grant of computer time from the DOD High Performance Computing Modernization Program at the ARL DoD Supercomputing Resource Center. We thank Brendan Gifford and Jason Morrill for fruitful discussions.

References

(1) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354–357.

(2) Ravi, P.; Badgujar, D. M.; Gore, G. M.; Tewari, S. P.; Sikder, A. K. Review on Melt Cast Explosives. *Propellants Explos. Pyrotech.* **2011**, *36*, 393–403.

(3) Johnson, E. C.; Sabatini, J. J.; Chavez, D. E.; Sausa, R. C.; Byrd, E. F.; Wingard, L. A.; Guzmàn, P. E. Bis (1, 2, 4-oxadiazole) bis (methylene) Dinitrate: A High-Energy Melt-Castable Explosive and Energetic Propellant Plasticizing Ingredient. *Org. Process Res. Dev.* **2018**, *22*, 736–740.

(4) Johnson, E. C.; Bukowski, E. J.; Sabatini, J. J.; Sausa, R. C.; Byrd, E. F.; Garner, M. A.; Chavez, D. E. Bis (1, 2, 4-oxadiazolyl) Furoxan: A Promising Melt-Castable Eutectic Material of Low Sensitivity. *ChemPlusChem* **2019**, *84*, 319–322.

(5) Zalba, B.; Marın, J. M.; Cabeza, L. F.; Mehling, H. Review on Thermal Energy Storage with Phase Change: Materials, Heat Transfer Analysis and Applications. *Appl. Therm. Eng.* **2003**, *23*, 251–283.

(6) Sharma, A.; Tyagi, V. V.; Chen, C.; Buddhi, D. Review on Thermal Energy Storage with Phase Change Materials and Applications. *Renew. Sustain. Energy Rev.* **2009**, *13*, 318–345.

(7) Agrawal, P. M.; Rice, B. M.; Thompson, D. L. Molecular Dynamics Study of the Melting of Nitromethane. *J. Chem. Phys.* **2003**, *119*, 9617–9627.

(8) Zhang, Y.; Maginn, E. J. A Comparison of Methods for Melting Point Calculation using Molecular Dynamics Simulations. *J. Chem. Phys.* **2012**, *136*, 144116.

(9) Brorsen, K. R.; Willow, S. Y.; Xantheas, S. S.; Gordon, M. S. The Melting Temperature

of Liquid Water with the Effective Fragment Potential. *J. Phys. Chem. Lett.* **2015**, *6*, 3555–3559.

(10) Chen, L.; Bryantsev, V. S. A Density Functional Theory Based Approach for Predicting Melting Points of Ionic Liquids. *Phys. Chem. Chem. Phys.* **2017**, *19*, 4114–4124.

(11) Katritzky, A. R.; Lomaka, A.; Petrukhin, R.; Jain, R.; Karelson, M.; Visser, A. E.; Rogers, R. D. QSPR Correlation of the Melting Point for Pyridinium Bromides, Potential Ionic Liquids. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 71–74.

(12) Dearden, J. C. Quantitative Structure-Property Relationships for Prediction of Boiling Point, Vapor Pressure, and Melting Point. *Environ. Toxicol. Chem.* **2003**, *22*, 1696–1709.

(13) Trohalaki, S.; Pachter, R.; Drake, G. W.; Hawkins, T. Quantitative Structure-Property Relationships for Melting Points and Densities of Ionic Liquids. *Energy & Fuels* **2005**, *19*, 279–284.

(14) Yuan, W.; Hansen, A. C.; Zhang, Q. Vapor Pressure and Normal Boiling Point Predictions for Pure Methyl Esters and Biodiesel Fuels. *Fuel* **2005**, *84*, 943–950.

(15) Preiss, U. P.; Beichel, W.; Erle, A. M.; Paulechka, Y. U.; Krossing, I. Is Universal, Simple Melting Point Prediction Possible? *ChemPhysChem* **2011**, *12*, 2959–2972.

(16) Morrill, J. A.; Byrd, E. F. Development of Quantitative Structure Property Relationships for Predicting the Melting Point of Energetic Materials. *J. Mol. Graph. Model.* **2015**, *62*, 190–201.

(17) Godavarthy, S. S.; Robinson, R. L.; Gasem, K. A. An Improved Structure-Property Model for Predicting Melting-Point Temperatures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5117–5126.

(18) Varnek, A.; Kireeva, N.; Tetko, I. V.; Baskin, I. I.; Solov'ev, V. P. Exhaustive QSPR Studies of a Large Diverse Set of Ionic Liquids: How Accurately Can We Predict Melting Points? *J. Chem. Inf. Model.* **2007**, *47*, 1111–1122.

(19) Seko, A.; Maekawa, T.; Tsuda, K.; Tanaka, I. Machine Learning with Systematic Density-Functional Theory Calculations: Application to Melting Temperatures of Single-and Binary-Component Solids. *Phys. Rev. B* **2014**, *89*, 054303.

(20) Jackson, N.; Sanchez-Lengeling, B.; Vazquez-Mayagoitia, A.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. A Diversified Machine Learning Strategy for Predicting and Understanding Molecular Melting Points. 2019; `https://chemrxiv.org/articles/A_Diversified_Machine_Learning_Strategy_for_Predicting_and_Understanding_Molecular_Melting_Points/9914378`.

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

(22) Nebgen, B.; Lubbers, N.; Smith, J. S.; Sifain, A. E.; Lokhov, A.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Transferable Dynamic Molecular Charge Assignment using Deep Neural Networks. *J. Chem. Theory Comput.* **2018**, *14*, 4687–4698.

(23) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495–4501.

(24) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547.

(25) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties:

Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

(26) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine Learning of Accurate Energy-Conserving Molecular Force Fields. *Sci. Adv.* **2017**, *3*, e1603015.

(27) Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579–590.

(28) Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical Modeling of Molecular Energies using a Deep Neural Network. *J. Chem. Phys.* **2018**, *148*, 241715.

(29) Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764–5776.

(30) St John, P.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. Prediction of Homolytic Bond Dissociation Enthalpies for Organic Molecules at Near Chemical Accuracy with Sub-Second Computational Cost. 2019; `https://chemrxiv.org/articles/Prediction_of_Homolytic_Bond_Dissociation_Enthalpies_for_Organic_Molecules_at_near_Chemical_Accuracy_with_Sub-Second_Computational_Cost/10052048`.

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

(32) Ye, S.; Hu, W.; Li, X.; Zhang, J.; Zhong, K.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. A Neural Network Protocol for Electronic Excitations of N-methylacetamide. *Proc. Natl. Acad. Sci.* **2019**, *116*, 11612–11617.

(33) Stein, H. S.; Guevarra, D.; Newhouse, P. F.; Soedarmadji, E.; Gregoire, J. M. Machine Learning of Optical Properties of Materials–Predicting Spectra from Images and Images from Spectra. *Chem. Sci.* **2019**, *10*, 47–55.

(34) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **2019**, *6*, 1801367.

(35) Hu, W.; Ye, S.; Zhang, Y.; Li, T.; Zhang, G.; Luo, Y.; Mukamel, S.; Jiang, J. Machine Learning Protocol for Surface Enhanced Raman Spectroscopy. *J. Phys. Chem. Lett.* **2019**, *10*, 6026–6031.

(36) Kananenka, A. A.; Yao, K.; Corcelli, S. A.; Skinner, J. L. Machine Learning for Vibrational Spectroscopic Maps. 2019; `https://doi.org/10.1021/acs.jctc.9b00698`.

(37) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery using Failed Experiments. *Nature* **2016**, *533*, 73.

(38) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

(39) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, 1–16.

(40) Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A.; Parrilla, P. C.; Pendleton, I. M.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. Robot-Accelerated Perovskite Investigation and Discovery (RAPID): 1. Inverse Temperature Crystallization. 2019; `https://chemrxiv.org/articles/Robot-Accelerated_Perovskite_Investigation_and_Discovery_RAPID_1_Inverse_Temperature_Crystallization/10013090`.

(41) Fourches, D.; Muratov, E.; Tropsha, A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(42) Mitchell, J. B. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.

(43) Zang, Q.; Mansouri, K.; Williams, A. J.; Judson, R. S.; Allen, D. G.; Casey, W. M.; Kleinstreuer, N. C. In Silico Prediction of Physicochemical Properties of Environmental Chemicals using Molecular Fingerprints and Machine Learning. *J. Chem. Inf. Model.* **2017**, *57*, 36–49.

(44) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of Compounds for Machine-Learning Prediction of Physical Properties. *Phys. Rev. B* **2017**, *95*, 144110.

(45) Barnes, B. C.; Elton, D. C.; Boukouvalas, Z.; Taylor, D. E.; Mattson, W. D.; Fuge, M. D.; Chung, P. W. Machine Learning of Energetic Material Properties. 2018; `https://arxiv.org/abs/1807.06156`.

(46) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(47) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M. et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

(48) Huang, B.; Von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 161102.

(49) Collins, C. R.; Gordon, G. J.; Von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148*, 241718.

(50) Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

(51) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(52) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

(53) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962–4967.

(54) Bergström, C. A.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1177–1185.

(55) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.

(56) Sun, Y.; Bai, H.; Li, M.; Wang, W. Machine learning approach for prediction and understanding of glass-forming ability. *J. Phys. Chem. Lett.* **2017**, *8*, 3434–3439.

(57) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying Machine Learning Techniques to Predict the Properties of Energetic Materials. *Sci. Rep.* **2018**, *8*, 9059.

(58) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *ACS Catalysis* **2019**, *9*, 2313–2323.

(59) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63.

(60) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.

(61) Lo, Y.-C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov. Today* **2018**, *23*, 1538–1546.

(62) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(63) Lian, B.; Yalkowsky, S. H. Unified Physicochemical Property Estimation Relationships (UPPER). *J. Pharm. Sci.* **2014**, *103*, 2710–2723.

(64) 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.

(65) Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; pp 785–794.

(66) Leo, A. J. Calculating Log P(oct) From Structures. *Chem. Rev.* **1993**, *93*, 1281–1306.

(67) Walters, W. P.; Yalkowsky, S. H. ESCHER A Computer Program for the Determination of External Rotational Symmetry Numbers from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1015–1017.

(68) Jain, A.; Yalkowsky, S. H. Estimation of Melting Points of Organic Compounds-II. *J. Pharm. Sci.* **2006**, *95*, 2562–2618.

(69) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics, New York, 2001; Vol. 1.

(70) Feng, D.; Svetnik, V.; Liaw, A.; Pratola, M.; Sheridan, R. P. Building Quantitative Structure-Activity Relationship Models Using Bayesian Additive Regression Trees. *J. Chem. Inf. Model.* **2019**, *59*, 2642–2655.

(71) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353–2360.

(72) Austin, J. A Relation Between the Molecular Weights and Melting Points of Organic Compounds. *J. Am. Chem. Soc.* **1930**, *52*, 1049–1053.

(73) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; John Wiley & Sons, 2008; Vol. 11.

(74) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

(75) Rice, B. M.; Byrd, E. F. Evaluation of Electrostatic Descriptors for Predicting Crystalline Density. *J. Comput. Chem.* **2013**, *34*, 2146–2151.

(76) Tetko, I. V.; Sushko, Y.; Novotarskyi, S.; Patiny, L.; Kondratov, I.; Petrenko, A. E.; Charochkina, L.; Asiri, A. M. How Accurately Can We Predict the Melting Points of Drug-Like Compounds? *J. Chem. Inf. Model.* **2014**, *54*, 3320–3329.

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

(78) Alantary, D.; Yalkowsky, S. H. Estimating the Physicochemical Properties of Polysubstituted Aromatic Compounds using UPPER. *J. Pharm. Sci.* **2018**, *107*, 297–306.

(79) Zhang, Y.; Maginn, E. J. Toward Fully in Silico Melting Point Prediction using Molecular Simulations. *J. Chem. Theory Comput.* **2013**, *9*, 1592–1599.

Graphical TOC Entry