A Diversified Machine Learning Strategy for Predicting and Understanding Molecular Melting Points

ganesh sivaraman, Nicholas Jackson, Benjamin Sanchez-Lengeling, Alvaro Vazquez-Mayagoitia, Alan Aspuru-Guzik, Venkatram Vishwanath, Juan de Pablo

Submitted date: 27/09/2019 • Posted date: 30/09/2019
Licence: CC BY-NC-ND 4.0

Citation information: sivaraman, ganesh; Jackson, Nicholas; Sanchez-Lengeling, Benjamin; Vazquez-Mayagoitia, Alvaro; Aspuru-Guzik, Alan; Vishwanath, Venkatram; et al. (2019): A Diversified Machine Learning Strategy for Predicting and Understanding Molecular Melting Points. ChemRxiv. Preprint.

The ability to predict multi-molecule processes, using only knowledge of single molecule structure, stands as a grand challenge for molecular modeling. Methods capable of predicting melting points (MP) solely from chemical structure represent a canonical example, and are highly desirable in many crucial industrial applications. In this work, we explore a data-driven approach utilizing machine learning (ML) techniques to predict and understand the MP of molecules. Several experimental databases are aggregated from the literature to design a low-bias dataset that includes 3D structural and quantum-chemical properties. Using experimental and polymorph-induced uncertainties, we derive a tenable lower limit for MP prediction accuracy, and apply graph neural networks and Gaussian processes to predict MP competitive with these error bounds. To further understand how MP correlates with molecular structure, we employ several semi-supervised and unsupervised ML techniques. First, we use unsupervised clustering methods to identify classes of molecules, their common fragments, and expected errors for each data set. We then build molecular geometric spaces shaped by MP with a semi-supervised variational autoencoder and graph embedding spaces, and apply graph attribution methods to highlight atom-level contributions to MP within the datasets. Overall, this work serves as a case study of how to employ a diversified ML toolkit to predict and understand correlations between molecular structures and thermophysical properties of interest.

File list (1)

ML_of_Melting_Temperatures.pdf (2.64 MiB)
A diversified machine learning strategy for predicting and understanding molecular melting points

Ganesh Sivaraman,†,‡ Nicholas E. Jackson,¶§† Benjamin Sanchez-Lengeling,∥
Álvaro Vázquez-Mayagoitia,⊥ Alán Aspuru-Guzik,#,@∇ Venkatram Vishwanath,†
and Juan J. de Pablo∗,§,¶

†Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL 60439, USA
‡Contributed equally to this work
¶Pritzker Institute of Molecular Engineering, University of Chicago, Argonne National Laboratory, Lemont, IL 60439, USA
§Center for Molecular Engineering, Argonne National Laboratory, University of Chicago, Chicago, IL 60637, USA
∥Department of Chemistry, Harvard University, Cambridge, MA 02138, USA
⊥Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, USA
#Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada
@Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada
△Vector Institute for Artificial Intelligence, Toronto, Ontario M5S 1m1, Canada
∇Fellow, Canadian Institute for Advanced Research, Toronto, Ontario M5G 1Z8, Canada

E-mail: depablo@uchicago.edu
Abstract

The ability to predict multi-molecule processes, using only knowledge of single molecule structure, stands as a grand challenge for molecular modeling. Methods capable of predicting melting points (MP) solely from chemical structure represent a canonical example, and are highly desirable in many crucial industrial applications. In this work, we explore a data-driven approach utilizing machine learning (ML) techniques to predict and understand the MP of molecules. Several experimental databases are aggregated from the literature to design a low-bias dataset that includes 3D structural and quantum-chemical properties. Using experimental and polymorph-induced uncertainties, we derive a tenable lower limit for MP prediction accuracy, and apply graph neural networks and Gaussian processes to predict MP competitive with these error bounds. To further understand how MP correlates with molecular structure, we employ several semi-supervised and unsupervised ML techniques. First, we use unsupervised clustering methods to identify classes of molecules, their common fragments, and expected errors for each data set. We then build molecular geometric spaces shaped by MP with a semi-supervised variational autoencoder and graph embedding spaces, and apply graph attribution methods to highlight atom-level contributions to MP within the datasets. Overall, this work serves as a case study of how to employ a diversified ML toolkit to predict and understand correlations between molecular structures and thermophysical properties of interest.

Introduction

The accurate determination of bulk thermophysical properties of molecules and polymers using only single molecule structure is a topic of critical academic and industrial interest. Historically, machine learning (ML) methods including quantitative activity-structure relationships\(^1\) (QSAR) and quantitative property-structure relationships\(^2\) (QSPR) methods have dominated *in silico* predictive molecular modeling, with considerable success across a broad array of prediction tasks including molecular solubility, biological toxicity, and ther-
mophysical properties. These classes of data-driven, quantitatively predictive models, when incorporated with high-throughput screening and design efforts, can aid in the identification, generation, and characterization of molecular species with applications in drug design, organic electronics, and solar fuels materials, among many others.

A molecular property of interest to a variety of critical industrial applications is the prediction of a molecule’s melting point (MP). Not only does a molecule’s MP define the temperature at which a material transitions from solid to liquid, but it can be correlated with a number of industrially vital material properties. For example, solubilities of candidate drug-like molecules are often estimated using a general solubility equation (GSE) approach, where one of the two inputs is the MP of a molecule. The recent emergence of interest in ionic liquids has made the correlation of MP with ionic liquid structure a critical endeavor, especially as it pertains to their stability. MP can also be well-correlated with a liquid’s viscosity. Molecular and polymeric glass transition temperatures are bounded by the MP of a material, making a priori knowledge of the MP a useful insight regarding anticipated glass transition temperatures and mechanical properties. In any application where high-throughput screening is an avenue for material discovery, accurate MP prediction will determine the scope of practical candidate materials.

Two computational strategies exist for MP prediction: physics-based models and ML approaches. From a physics-based approach, ab-initio or classical molecular dynamics simulations can be utilized to predict the MPs of molecular solids. However, many of these simulation techniques depend strongly upon the challenging feat of predicting the 3D crystal packing of molecular structures a priori, though methods have been devised to circumvent this limitation. Moreover, the accuracy of classical force-fields required to predict MPs is often insufficient to handle the vast diversity of chemical structures and intermolecular interactions present in many materials. Thus, even if 3D crystal structures could be known a priori, errors derived from force-field approximations are difficult to quantify, and molecular dynamics simulations, force-fields, and workflows are not yet easily scalable for
high-throughput screening approaches across chemically diverse structures. Alternatively, ML strategies have been widely employed for the prediction of MPs, with models routinely achieving prediction errors of 35-50K.\textsuperscript{2,13–23} These models rely on the use of experimentally determined MPs, combined with supervised ML, to regress molecular structures to experimental MPs. Within the last five years, the quantity of experimentally available MP data has drastically increased thanks to the efforts of Tetko, with recent work accruing over 200,000 experimental MPs derived from the patent literature\textsuperscript{14,15}.

In this work, we apply data-driven and ML approaches to understand and predict molecular MP, working with the SMILES representations of molecules as input for the ML algorithms. The primary contributions of this work are:

\begin{itemize}
  \item **An integrated experimental and quantum-chemical dataset.** Aggregating from several literature sources, we build a dataset of $\sim$47k molecules with experimental MP, including augmentation with 3D molecular structures and quantum chemical properties. By studying the literature on crystal polymorphs and experimental measurement procedures, we quantify the expected experimental error in MP reporting, providing useful anticipated bounds for MP prediction errors.

  \item **State-of-the-art MP prediction from single-molecule properties.** We employ graph convolution neural networks (GCN) and Gaussian Process Regression (GPR) to obtain MP predictions that approach the underlying experimentally limiting uncertainties in the data sets. Additionally, we assess the role of 3D structural, conformational, and quantum-chemical descriptors and their abilities to improve MP predictions.

  \item **Diversified chemical analysis.** We deploy a variety of unsupervised and supervised tools to characterize the relationship between molecular structure and MP. We build geometric spaces shaped by MP with semi-supervised variational autoencoders and graph embedding spaces, and inspect individual atomic contributions to MP via graph attribution methods. We partition our molecular space into clusters and characterize
fragments in each cluster to better understand the chemical composition and organization of molecules based on MP.

In what follows, we discuss the methods used in this study, for which more technical details can be found in the supporting information (SI). We then present results and discussion around each of the methods. Finally, with the aim of transparency we make our dataset and code available on GitHub. The code repository to build models and make predictions, along with datasets themselves, can be found at https://github.com/argonne-lcf/melting_points_ml.

Methodology

A schematic diagram of the methods employed to study MP is depicted in Figure 1, and each of the components is briefly detailed hereafter. More detailed descriptions (e.g. hyperparameters) can be found in the SI.

Datasets

Four publicly available data sets of experimental MPs were chosen for this study, as outlined by Tetko.\textsuperscript{13} The statistics associated with each of the data sets, including the combined data (labeled “All”), are summarized in Figure 2 a). The Bradley data set is a highly curated, “gold” standard\textsuperscript{24} for MP data sets, and has been double-validated to only contain data with multiple reported measurements within 5 C. The Bergström data set,\textsuperscript{21} which is an order of magnitude smaller in sample size, was also generated via rigorous manual curation. Additionally, most of the compounds reported in the Bergström data set fall well within a subset of the MP range of the Bradley data set. For these reasons, the Bergström and Bradley data sets were merged for this study. The Enamine data set was created by
Figure 1: Diversified ML pipeline for predicting and understanding molecular melting points.

Enamine Ltd,\textsuperscript{25} a chemical supplier. The OCHEM data set was derived from a diverse pool of non-curated data from the Online Chemical Modeling Environment (OCHEM).\textsuperscript{26} Authors interested in further analysis of these datasets beyond our own should consult the excellent work of Tetko.\textsuperscript{13} In this work, we augment the original data sets of Tetko by including a variety of structural and quantum-chemical descriptors, as outlined in the SI.

**MP Prediction**

For MP prediction we use three supervised learning methods: Random forests (RF), GPR, and GCN. Each model encompasses a different type of modelling approach for QSAR applications. RF are an ensembling approach that aggregates several randomized decision trees and pools predictions from each to generate more robust property estimates. RF are used frequently in QSAR applications since they are robust to different modalities of data and
Figure 2: **Experimental Data Sets.** A) MP distributions for experimental data sets. Labels and colors are based on data set source. Black dashed lines indicate drug-like region [323.15, 523.15]. b) Distribution of MP intervals for 119 literature molecules exhibiting multiple crystal polymorphs.

are straightforward to apply. GPR are models that combine features of Bayesian linear regression and kernel ridge regression to generate a distribution of functions that best fit the data based on gaussian assumptions. GPR rely on learning functions (kernels) that use relative distances between data points to make predictions. Predictions on a data point $x$ are reported as the mean of a gaussian distribution and the standard deviation represents the uncertainty bounds for prediction. Both RF and GPR rely on predefined features (e.g. fingerprints, quantum-chemical properties, etc.) to represent molecular structures, whereas GCN\(^{27}\) directly utilize a graph-structured representation of a molecule, with atoms as nodes and bonds as edges of a graph. GCN learn a vectorized representation of a molecule which can be used with another model, such as a multilayer perceptron (MLP), and trained end-to-end. GCN works by iteration; for each node it aggregates neighboring local graph information and transforms it via an MLP to retrieve a new node representation. It then projects all nodes to a graph-level vector which can be thought of as task-optimized fingerprints.\(^{28}\) All graph operations are designed to preserve graph symmetries.
Geometric Spaces

In this work, we construct geometric spaces structured around MP that allow us to better understand how molecules are structured according to MP, as well as serve as a sanity check for when molecules do not follow the distribution of a dataset. Since these latent spaces are high-dimensional, we reduce their dimensionality for visualization purposes using linear principle components analysis (PCA). To construct these geometric spaces, we apply two distinct methods: semi-supervised variational autoencoders (SSVAE) and the penultimate layers of GCNs. SSVAEs are generative models that learn to encode data into a vector representation in a latent space, and then decode the data back to its original representation. Both operations are modeled with neural networks and optimized concurrently.

Bombarelli et al. first demonstrated the usage of VAE with SMILES strings to generate new molecules with drug-like properties using the Zinc dataset. One key result was the ability of the SSVAE to shape the organization of the latent space representation of molecules based on the predicted properties of interest. VAE requires large amounts of data to be able to generalize to new molecules; since our labeled dataset has ~47k, we rely on semi-supervised learning to leverage larger unlabeled datasets. In our network, we have a MP predicting neural network that maps the latent space to predicted MPs. Because we want our latent space to be informed by molecules that have been synthesized and exist on a shelf somewhere in the world, we used a set of 1M purchasable molecules from emolecules to inform the chemical structure of the latent space. We train the VAE with a mix of labeled and unlabeled data: for each batch we mask the loss function that predicts properties. To ensure that we are able to construct grammatically valid SMILES, we use SELFIES. Our geometric space is then the latent space of our SSVAE. This space can also be used for the inverse design of materials due to its built-in decoding capability.

Another means of constructing geometric spaces is by examining the space of activations inside neural networks by analyzing the penultimate layer in a GCN. In the case of regression, the ultimate layer will be a linear model, so if the entire model is accurate, the penultimate
layer can be used to embed molecules, and these molecules should be organized on a gradient since the GCN will have to fit a line across this space in order to predict MP.

**Graph Attribution**

Graph attribution refers to the task of attributing values to elements of a graph. If we care about building interpretable predictions in GCN, then we wish to create a graph attribution method that assigns positive or negative weights to graph elements in relation to their importance for prediction. For this purpose, we utilize grad-CAM with GCNs. These methodologies have been previously explored in the context of drug-like properties. Grad-CAM uses gradient information flowing into the convolutional layer of a GCN to understand the importance of each neuron for a given task. By normalizing this information, we are able to build a heatmap delineating the contributions for each node in a molecular graph. It should be noted that the heatmap for each molecule is a local explanation, that is, the relative weights between different molecular heatmaps are not directly comparable.

**Active Sampling and Clustering**

An important approach in active learning is the exploitation of the cluster structure within the underlying dataset. Our aim is to perform active sampling based on unsupervised learning that can exploit the cluster structure of the MP data sets, thereby reducing the bias arising from passive sampling (i.e. random sampling). Unsupervised clustering involves organizing unlabeled data into groups of common clusters based on their similarity. We apply unsupervised clustering to all of the unlabeled chemical data from our MP data sets with the following two goals: 1) To gain a qualitative understanding of uncertainties in the data sets arising from curation qualities, experimental conditions, and experimental techniques and 2) To create realistic chemical similarity-based test/train splits for use in supervised ML, and to understand the impact of this active sampling on MP prediction.
Results and Discussion

Limits to Prediction

With the curation of large experimental MP data sets, it is necessary to estimate the inherent limitations of MP prediction, specifically as it relates to experimental uncertainties, sample purity, crystal polymorphs, and data parsing and recording. What magnitude of error is expected from the presence of crystal polymorphs? What is the intrinsic MP precision when an experiment is done “perfectly”? When uncertainty is incorporated into the experimentally recorded values of the MP, how does this influence our expectations of the maximum obtainable predictive accuracy? In what follows, we consider three primary contributions to MP error: the presence of crystal polymorphs, experimental errors/uncertainties, and errors in data recording.

Crystal Polymorphs

The presence of crystal polymorphs with distinct MPs for a given molecular structure can impact the predictive accuracy of the ML algorithm. As ML algorithms generally only predict a single value of the MP for a given molecular structure, the predicted MP value likely corresponds to the MP of a single molecular polymorph, the identity of which is usually unknown. Consequently, this polymorph ambiguity introduces an inherent uncertainty in the prediction task. Naively, one can grasp the potential magnitude of such differences in MP by considering a substance such as cocoa butter, and the fact that its industrially relevant polymorphs are known to be separated by nearly 20K. To more quantitatively approach the issue, we have examined the MPs of 119 unique crystal polymorphs gathered from the experimental literature. In this data set, we have computed the size of the MP interval in K for all polymorphs for each molecular structure ($\Delta T_m$), and histogrammed the distribution of MP intervals in Figure 2b). To compute this interval, we take the difference between the maximum and minimum recorded MPs for polymorphs of a given molecule. The data in
Figure 2 b) represents the maximum potential error in MP prediction if we assume the ML algorithm predicts a value somewhere within this interval.

We observe that over half of the examined molecules would exhibit polymorph-related predictive errors of less than 10 K, which is an encouraging result for QSPR predictions. Moreover, over 80 percent of molecules would have their errors bounded by 20 K, though it is somewhat troubling that for some polymorphs MP variations as large as 57 K have been measured. However, the fact that 96 percent of molecules exhibit a MP interval of less than 30 K, along with the fact that only a fraction of the molecular structures in a data set will exhibit multiple crystal polymorphs, suggests that crystal polymorph induced inaccuracies are likely not the sole feature limiting the performance of MP prediction. The first and second moments of the simple crystal polymorph MP distribution correspond to a range of \( \sim 11 - 16K \), which we tentatively use as a lower bound for our expected error due to polymorphs. Note, that the previously obtained intervals of 35-50K MP prediction accuracy are significantly worse than that derived from the simple analysis of Figure 2 b).

**Experimental and Chemical Uncertainties**

MPs are typically measured via either a melting point apparatus (MPA) or differential scanning calorimetry (DSC) experiment. Generally speaking, if properly calibrated and performed, both DSC and MPA (https://www.thinksrs.com/products/mpa100.html) measurements should exhibit reproducibility of \( \sim 0.1 \) C. For the specific case of DSC, one observes a peak in the melting curve from which a specific MP must be derived. ICTAC standards state that one should take the onset of the melt peak as the MP for metals and organics, but that the peak value should be used for polymers. However, even with these considerations, if properly performed, the majority of pure organic materials typically exhibit melting ranges of 1-2 C for a given material.

To properly perform either MPA or DSC experiments, the heating rate must be appropriately chosen. Typical heating rates for both MPA and DSC, depending on the precision
required, are between 0.1 C/min and 20 C/min, with most high precision studies occurring at heating rates less than 1 C/min. Despite the majority of DSC peak widths for small organic molecules being 1-2 C, one can establish a generous upper bound for potential experimental error in MP by examining the literature of the heating-rate dependence of macromolecule MP, where heating-rate effects should be largest. For the case of crystalline polyethylene, the MP decreases approximately 6.5 C when the heating rate is increased from 0.6 C/min to 20 C/min. With this information in mind, if these experiments are performed for pure samples, with appropriate heating rates, and the MP value is taken at the onset of the melt peak, DSC and MPA measurements should yield measurement errors less than 1-2 C, with a polymer-derived maximum bound of roughly 6 C. We emphasize that these arguments are back-of-the-envelope calculations, but believe them to be in agreement with common experimental experience.

Sample purity is an orthogonal issue that can contribute to the inaccuracy of MP measurements. Indeed, in many cases, MP measurements are meant to assess sample purity by identifying an increase in the MP relative to a known pure sample. If a material has degraded in storage or during the experiment, then such purity issues will induce errors in the experimental MP. Tetko\textsuperscript{13} performed analysis of molecular structures exhibiting high MP prediction error and concluded that functional groups capable of decomposition during storage/heating were significantly more represented in the set of outlier compounds relative to the rest of the data set. The magnitudes of these errors are entirely dependent upon the identities of the impurities, and so we refrain from generally speculating on their magnitudes.

**Errors in Data Recording**

Tetko\textsuperscript{13} provided an in-depth analysis of outliers in their 45,000 molecule data set. Specifically, for the OCHEM and Enamine subsets, 394 and 427 outlier compounds were identified, respectively. These outliers corresponded to RMSE prediction errors > 130 C. Their analysis determined that 71 of the outlying compounds exhibited MP of less than room temperature,
and consequently were likely not measured correctly. In the case of the OCHEM subset, three outlying compounds misreported MPs for the salt form of a compound, three cases reported the wrong temperature units, and two cases misrecorded a minus sign. Upon removal of these outliers and comparison to other literature values of MP for questionable data points, this screening improved their predicted RMSE significantly.

**Passive Sampling Data Analysis**

With anticipated experimental uncertainties established, we analyze the statistics of the MP data sets, as shown in Figure 2a). The Enamine data set exhibits a higher mean MP relative to other data sets, resulting in a positive skew as observed by a long tail of the histogram at higher temperatures. Enamine’s mean MP is also closest to that for the drug-like region (i.e. 423 K). As noted by Tetko et al.,\textsuperscript{13} the Enamine data set was generated using identical experimental protocols for all analyzed molecular species. The MP distribution statistics reveal that Enamine also has the smallest standard deviation among the analyzed data sets. The OCHEM data set is an aggregation of a variety of diverse data sources obtained with different experimental protocols and measurements and exhibits a long tail at low temperatures (i.e. negative skew). The large standard deviation of the OCHEM data set relative to Enamine can likely be attributed to the heterogeneity of sources and measurement protocols as reported by Tetko.\textsuperscript{13} The well-curated nature of the BradBerg data set implies that the large standard deviation observed in the distribution in Figure 2a) is due to the inherent diversity of chemical structures and MPs in the data set. Combining all data sets resulted in a MP distribution which has characteristics shifted closer to OCHEM (i.e. negative skew, mean and standard deviations closer to OCHEM).

In supervised ML, an algorithm is trained on a data set and validated on a held-out test data set. The most common way of creating these training and test data sets is via passive sampling, where the original data is randomly split into groups without regard for the underlying statistical nature of the data set. The regression results for different supervised
ML models trained on the passively sampled MP data sets of Figure 2a) are shown in Table 1. We utilize the RF ML method as a benchmark for initially comparing the mean absolute error (MAE) metrics among the individual and combined data sets. The predicted MAE validated on individual test sets follows the trend OCHEM > BradBerg > All > Enamine. We observed a clear correlation between the relative predicted MAE of OCHEM and Enamine and their associated standard deviations. Interestingly, the BradBerg data set exhibits a low MAE and high value of the correlation coefficient, which we attribute to the highly-curated and chemically-diverse nature of the dataset, the latter of which is confirmed by its large standard deviation. These results emphasize the critical importance of having well-curated data sets, as in the cases of data sets that are not well-curated, including more data will not lead to better model performance.

The use of more advanced supervised ML algorithms, including GCN and GPR, on the passively sampled data sets lead to significant improvements in predictive accuracy. Specifically, for both GCN and GPR, MAE below 30K can be achieved for the entire data set using both methods, with MAE of 28.9 K and a correlation coefficient of 0.78 obtained when using the GPR method in conjunction with a feature set containing both 3D and quantum-chemically derived descriptors. If one restricts the performance of the GPR method to only molecules in the ‘drug-like’ interval as described by Tetko, we can obtain a cross-validated MAE of 25.8 K in the drug-like region. It is interesting that the GCN method, which does not include any quantum-chemical or 3D structural information, can obtain MAE below 30 K solely from the details of the graph structure derived from the molecular SMILES strings, a result that is in agreement with recent GCN work. This points to the promise of graph-based techniques that have been described previously, especially provided these methods do not require the additional cost of conformer searches or quantum-chemical analysis to generate ML features. However, we do observe an improvement in predictive performance relative to the GCN when utilizing the GPR methodology and including both 3D structural information and quantum-chemically derived properties (solvation energy plays a reliable
role in reducing the predicted MAE, as described later on). Additionally, the GPR framework provides an assessment of prediction uncertainty, which is desirable for MP prediction, especially if one is unsure of the chemical similarity between a new molecule and the model’s training data set.

Table 1: MP Regression Results for Experimental Data Sets. † Model using only systems with melting temperatures in the drug-like region [323.15, 523.15]K

| Method    | OCHEM MAE (K)/$R^2$ | Enamine MAE (K)/$R^2$ | BradBerg MAE (K)/$R^2$ | All MAE/$R^2$ |
|-----------|----------------------|------------------------|------------------------|--------------|
| GPR       | 30.03(0.01)/0.77     | 28.60(0.00)/0.64       | 25.06(0.03)/0.88       | 28.85(0.01)/0.78 |
| GPR†      | 26.34(0.05)/0.60     | 25.65(0.02)/0.59       | 24.64(0.15)/0.64       | 25.80(0.03)/0.61 |
| RF        | 37.56(0.07)/0.66     | 32.01(0.09)/0.56       | 35.60(0.75)/0.76       | 34.62(0.13)/0.66 |
| GCN       | 31.59(0.83)/0.75     | 29.45(0.55)/0.62       | 28.51(0.80)/0.84       | 29.41(0.26)/0.75 |

**Cluster Analysis**

**Coarse Clustering**

Provided the previous regression results derived from simple statistics and supervised ML using the passively sampled data sets, we now employ a variety of unsupervised learning algorithms to both understand and actively sample the underlying chemical structures of the data sets. First, to understand the intrinsic uncertainties of the employed data sets, we apply Murtagh-Ward clustering using the Tanimoto similarity measure to distribute our ‘All’ data set into seven coarse clusters, as shown in Table 2. The application of the Murtagh-Ward algorithm groups chemically similar species (as determined by the Tanimoto index) into clusters of comparable chemical similarity, all entirely independent from any knowledge of the molecular MPs. To correlate the statistical properties of each cluster with their predictive accuracy via supervised ML, we apply the RF regressor to determine the MAE for MP predictions on each cluster as shown in Table 2.

The Murtagh-Ward algorithm clusters three data sets with more than 10,000 molecules each, and four data sets with less than 4,000 molecules each. We observe a strong correlation between the predicted MAE of a cluster and the cluster’s standard deviation. The two
clusters exhibiting the largest negative skew (i.e. clusters 1 and 6) also exhibit the largest MAE, whereas the clusters with the lowest MAE exhibit significant positive skew. This may suggest an inherent difficult in predicting the MP for low MP chemical structures, whereas higher temperature chemistries may be easier to learn. Since the OCHEM data set is known to be heterogeneous, and thus exhibits a higher predicted MAE in the passively sampled data sets, we chose to examine what fractions of the coarse clusters derived from Murtagh-Ward clustering were composed of molecules belonging to the OCHEM data set. While more than 50% of both clusters 1 and 6 are derived from the OCHEM data set, the other data sets also consist of similar fractions of OCHEM derived molecules. The lack of strong correlations between coarse cluster composition and the fraction of OCHEM molecules suggests that the OCHEM data set is not uniformly difficult to predict, and that a limited chemical subset of OCHEM may be leading to the reduced accuracy.

Table 2: Summary of Clusters Derived from Murtagh-Ward Clustering

| Cluster | Sample Size | Mean | Std | Median | Skew | MAE (K)/R² | fOCHEM |
|---------|-------------|------|-----|--------|------|------------|--------|
| 1       | 10124       | 410.74 | 75.01 | 411.15 | -0.14 | 38.33(0.19)/0.54 | 0.50 |
| 2       | 16155       | 397.91 | 69.56 | 394.15 | 0.24  | 31.83(0.19)/0.63 | 0.45 |
| 3       | 14634       | 422.76 | 71.26 | 419.15 | 0.29  | 35.09(0.26)/0.57 | 0.37 |
| 4       | 1489        | 397.8  | 67.23 | 394.15 | 0.26  | 31.71(0.44)/0.59 | 0.60 |
| 5       | 835         | 412.31 | 68.06 | 406.15 | 0.54  | 35.50(0.89)/0.50 | 0.61 |
| 6       | 3312        | 331.34 | 99.46 | 339.15 | -0.03 | 40.54(0.61)/0.71 | 0.58 |
| 7       | 833         | 267.27 | 65.04 | 270.05 | 0.49  | 32.91(0.94)/0.44 | 0.69 |

A further analysis was performed to unravel parent chemical classes of the compounds reported in each of the clusters. For this purpose, the international chemical identifier key (inchikey) of the compounds from the clusters were extracted and parsed through the ‘ClassyFire’ automated chemical classification server. The resulting parent chemical classes have been visualized in Figure 3. Cluster 1 has been classified to a diverse set of parent chemical classes (≈ 264) with no dominant class. In addition, the majority of compounds in cluster 1 seem to be dispersed to parent classes which are below 5% of the total fraction. This chemical diversity might have lead to the relatively larger standard deviation and
consequently higher MAE predictions. The majority of the chemical compounds reported in
cluster 2 are classified to a dominant parent class (i.e. benzene and substituted derivatives).
Cluster 3 has two dominant parent classes which form as large a fraction as the ‘others’
category. The smaller cluster 4 also has a majority of compounds classified to the benzene
and substituted derivatives parent class (much like the cluster 1). Cluster 6 has the highest
error; this cluster differentiates itself by possessing many small molecules with higher than
average MP and molecules with multiple chlorines and sulfur atoms. This could indicate
that the models have making predictions on smaller molecules and higher atomic number
atoms, where MP is often dominated by electronic phenomena and non-covalent forces.

Figure 3: Chemical classification of Murtagh-Ward clusters. Chemical classes were
obtained with Classyfire. Only parent chemical classes with at least 5% of the total fraction
are shown; parent chemical classes with fractions smaller than 5% were merged in to ‘Others’.
Cluster indices are the same as defined in Table 2.

Fine Grained Butina Clustering

To investigate the ability of active sampling to create a chemically-aware data set for use with
supervised ML, we apply the Butina clustering method to create a new data set (‘Butina
The MPs corresponding to the 33,408 compound data set (13,974 molecules removed) have been visualized in Figure 2a). It is clear that the fine-grained clustering generates a data set whose distribution of MPs is qualitatively similar to that of the parent ‘All’ data set, and selection of the new data set by the unsupervised clustering did not simply prune outlier MPs at the wings of the distribution. The Butina 0.6 data set is composed of 77% BradBerg, 72% Enamine, and 68 % OCHEM.

For training of the supervised ML algorithms, a 70/30 split was applied to the Butina 0.6 data set with the regression results shown in Table. 3. We begin by comparing the RF regressor results for the Butina 0.6 data set with respect to the passively sampled data sets (Table 1). The total regression error falls below that of both OChem and Enamine individually, despite still containing nearly 70% of each data set, without clipping outliers at high or low temperatures. The improved performance of the supervised ML on the actively sampled data set relative to the passively sampled data sets is further supported by the performance of more advanced regression methods, as shown in Table 3. RF exhibits the largest increase in predictive accuracy of $\sim 5.2K$ MAE with respect to OCHEM. Contrastingly, the GPR and GCN exhibit 1.7 K and 2.3 K improvements in predictive accuracy, respectively. The differences in improvements are likely due to the complexity of the supervised learning methods and the differences in the featurizations used. This supports the notion that more complex ML and featurization methods (GPR and GCN) are more effective at extracting relevant details during the learning process, even from the passively sampled data, relative to the simpler RF method. This is further supported by the performance of the Butina data set compared to that of the ‘All’ - for GCN and GPR, predictions are essentially identical, whereas for RF a noticeable 2 K improvement is observed. Consequently, in cases with limited data and or less-sophisticated regression methods, active sampling of chemical space should be a reliable strategy for modest improvements in data sets where chemical space is not designed to be uniformly sampled. For the majority of data sets, particularly those that are experimentally derived, this uniformity of chemical space is not a priori anticipated.
Table 3: MP Regression Results for Butina 0.6 Clustering Data Sets

| Method | MAE (K)   | $R^2$  |
|--------|-----------|--------|
| GPR    | 28.24 (0.02) | 0.75   |
| RF     | 32.31 (0.28) | 0.69   |
| GCN    | 29.26 (0.27) | 0.74   |

Inclusion of 3D Structures and Quantum-Chemical Properties

Following the use of the actively sampled data set relative to the passively sampled data set, we turn our attention to the impact of the inclusion of 3D structural and quantum-chemical descriptors on the performance of the supervised ML methods. To reiterate, we examine three manifestations of 3D structure (Coulomb Matrix, ECFP 3D, and Morse 3D) derived from quantum-chemical geometries, five quantum-chemical properties (total energy, HOMO-LUMO energy gap, SMD Solvation energy, dipole moment, and quadrupole moment), and a selected set of 108 RDKit features (see SI - denoted ‘Features’ in Figure 4 B). Two supervised ML methods (RF and GPR) are applied to assess how these 3D and quantum-chemical geometries impact regression performance (4). Furthermore, we compared the performance in the GPR models using two sources of 3D conformers: geometries provided by DFT (annotated with a -g suffix in Figure 4 B), and geometries obtained from a conformational search algorithm, as described in the Methods.

First, in Figure 4 A), we plot the performance of the RF algorithm as a function of feature sets constructed from ECFP and CM matrix representations, with and without quantum-chemical properties. In all of our studies, it is universally observed that the inclusion of the SMD solvation energy results in a consistent improvement in predicted MAE ($\approx 4K$ for 2D descriptors and $\approx 2K$ for 3D descriptors), however none of the other quantum-chemically derived properties exhibit a significant beneficial effect. Moreover, the ECFP is shown to outperform the CM in all cases, likely due to the extensive size of the CM and the large number of weights that must be trained and can likely lead to overfit models.
In Figure 4 A), we plot the performance of the GPR algorithm as a function of different feature sets. The Morse3D fingerprints result in the worst performance, showing that the inclusion of 3D structure within Morse fingerprints is not an effective feature representation. However, the E3FP fingerprints that also include 3D structure result in a significant improvement relative to the Morse 3D, with the performance of the 3D fingerprints still being comparable to those of the 2D fingerprints used in conjunction with the RF model. Interestingly, we observe that results are similar when using DFT optimized geometries, force-field optimized geometries, or the resultant geometry from a low-energy conformer search. This suggests that knowledge of the single molecule conformation, as well as precise details of intramolecular geometric structure, are less critical to MP prediction than a rough description of the molecular geometry/connectivity. To this end, while the knowledge of the exact crystal structure would likely be critical to predicting polymorph-specific MP, the precise single molecular geometry does not appear to improve MP prediction significantly.
The best performance of all descriptor combinations, including graph-based models, is observed when using a diverse feature set that includes 3D descriptors, quantum-chemically derived data, and the RDKit feature set described in the SI. These lowest MAE values in GPR are observed when combining 2D and 3D descriptors with RDKit features and quantum-chemistry data, and lead to the highest performance - all of these combined is referred to as COMBO in the Figure 4 B); it is worth noting that 2D descriptor plus RDKit features provide the most important contributions for better predictions.

The peak performance observed using GPR and a diverse feature set that includes 3D structure, quantum-chemical descriptors, and RDKit descriptors should be weighed in conjunction with the computational cost of generating such featurizations. In Table 3, GPR results in a $\sim 1K$ reduction compared to graph-based methods, however, the graph-based methods do not require any knowledge of 3D structure or the expense of quantum-chemically derived feature sets. Consequently, while the inclusion of these properties leads to the highest performing models, graph-based ML methods are likely the path forward to obtaining the highest-performing predictions with the least cost for feature set generation.

**MP Structured Geometric Spaces**

Figure 5 a), b) & d) display two geometric spaces shaped by MP, the first two being the graph embedding space of a GCN (a,b) and the final being the latent space of a SSVAE (d). These high dimensional spaces are visualized in 2D with linear PCA as a dimensionality reduction technique. We embed the test set into these spaces and color by temperature.

Figure 5 a) and d) showcases significant temperature gradients, which indicate there are natural directions in these geometric spaces corresponding to temperature changes. When a new molecular structure is provided, it can be embedded in this space with its relative position in this gradient characterizing its expected temperature and whether it corresponds
Figure 5: **Geometric spaces of molecules shaped by melting points.** Each space is constructed from test data sets unless specified. Dimensionality reduction is done with PCA. a) Graph embedding space is colored by temperature. Gray points correspond to newly embedded molecules. b) Graph embedding space for entire dataset is colored by cluster number. Cluster indices are the same as defined in Table 2. c) Selection of randomly sampled molecules. d) Latent space of SSVAE colored by temperature. Gray point corresponds to the average latent vector of molecules with 350-450 K. Molecules in c) are sampled from this point.
to the data distribution of the training set. If a molecule is outside the general cluster of
data points, it could mean the model is observing patterns not contained in the training set.

One feature to note in Figure 5 d) is that the density of molecules is encased in a circular
area. This is due to the prior of the SSVAE; each dimension is assumed to be Gaussian
distributed, which reflects itself as data lying on the surface of a hyper-sphere which when
projected on 2D corresponds to a circle. One notable result of the graph embedding space
is that the notion of chemical clusters is easy to visualize even if this information was not
explicitly provided during training. The SSVAE latent space did not exhibit this feature,
but if provided during training would likely result.

In Figure 5 b) we combine graph embedding with the unsupervised clustering derived
from the previous section to understand how the clusters discovered by the unsupervised
clustering compare to the organization of the latent space observed via graph embedding.
Encouragingly, the chemical distinctions encoded in the unsupervised clustering show up
prominently in different regions of the PCA components for the graph embedding in Figure
5 b), further demonstrating that the structuring of chemical latent space by the MP cor-
responds with distinct chemical classes derived from the use of the unsupervised clustering
algorithms. Consequently, the graph embedding is organizing the latent space not only in a
way that correlates with MP, but in a way in which chemical locality is preserved.

To showcase the usage of these spaces, we sample from the latent space and decode these
vectors into SMILES strings. This decoding process produces valid structures with a rate
of 73% percent. Since the unsupervised component of the SSVAE is trained on purchasable
molecules, we expect the newly sampled structures to resemble plausible molecules. Because
this space is organized around MP, we can sample new structures based on predicted MP. In
Figure 5c) we show six structures sampled from the centroid of a cluster of trained molecules
with MP in the range of 350 to 450 K. When we map these molecules to the graph embedding
space we see the molecules fall in a plausible region for this temperature range. Their distance
from the general density of datapoints gives certainty these molecules come from a similar
distribution of data (drug-like). These models can be augmented with other properties of interest to become a crucial component within a material discovery pipeline.

Figure 6: **Graph attribution of MP on similar molecules.** Colorbar indicates the strength of a node towards a positive temperature. Molecules were chosen to be chemically similar but with a wide array of temperatures.

**Graph Attribution**

Figure 6 showcases a visualization of molecular heatmaps colored by weak and strong contributions to predicted MP over a molecular graph. Each heatmap is a local explanation that highlights the atom-level contributions toward its predicted MP. In particular we picked a set of similar molecules (based on tanimoto distances) that have a large variance of temperatures. Grad-CAM shows that the main differentiating feature between the molecules is their non-ringed members. For example, when comparing molecule numbers 1, 4 and 5, the presence of a OH fragment is observed to strongly increase MP. Other MP increasing trends are the presence of symmetric halogen atoms (Br, Cl) in the center of the molecule. We provide only this limited analysis of chemical trends as a flavor of the utility of the graph
attribution method, and encourage interested readers to consult the codebase available on the GitHub for exploring more in-depth chemical trends.

**Conclusions**

MP prediction represents a classic example of collective multi-molecule property prediction using only knowledge derived from single molecule structure. In this work, we have shed insight on critical features of MP prediction in the context of molecular materials. First, we have performed a literature search that bounds both the experimental and polymorph induced uncertainties in MP determination; we posit that the former should be in the range of $\sim 2 - 3K$, whereas the average value of polymorph induced uncertainties is in the range of $\sim 11 - 16K$. With this knowledge, we have constructed an augmented data set with 3D geometries and quantum-chemical properties, and shown how supervised ML models can push the accuracy of MP prediction to be competitive with experimental uncertainty. We have also assessed the importance of 3D structural and quantum-chemically derived features in improving MP prediction accuracy, and discovered a modest $\sim 1 - 2K$ improvement relative to graph-based methods when using GPR, obtaining predicted MAE in the range of 25-29 K MAE, depending on the temperature interval of interest. However, we suspect that in the future the predictive advantage will continue to trend towards more advanced chemical graph-based techniques.

We have also demonstrated how a diversified ML toolkit can be used to minimize chemical bias in supervised ML training, and provide critical chemical insights into MP correlations. Such considerations are especially of interest in the context of experimental data sets, where the chemical structures of data sets are likely strongly biased towards specific chemical motifs. The use of geometric spaces to examine how molecular space is organized around MP can serve as a useful diagnostic tool to embed new molecules and look at their position with respect to existing data; this is particularly relevant when deploying these models in
real scenarios. We also showcase how local explanations with graph attributions can aid in understanding if the model is making a prediction based on our own notions of chemistry and MP phenomena. Identifying trends in how decreases or increases in MP come about with molecular changes will also aid in the design of new materials.

With these considerations in mind, it is useful to consider the future state of such predictions. In the absence of explicit knowledge of the crystal structures for which MPs are being predicted, we believe it is unlikely to see ML algorithms predict MP with sub-20 K accuracy in the near future. We believe that the ability to obtain predictive MP errors of $\sim 20 - 25K$ MAE should be obtainable in the future, with further improvements coming from more advanced graph-based treatments that incorporate other types of forces besides simple covalent bonds, and do not require extensive feature engineering. Also critical to this end will be the development of larger, well-curated MP data sets analogous to those of Bradley and Bergstrom, as well as the use of unsupervised learning algorithms to generate chemically-aware, low bias data sets. While our current efforts have been focused solely on MP prediction, we believe these techniques can be straightforwardly adapted to other bulk material properties in the future.

**Acknowledgement**

We thank Prof. Lian Yu for his helpful direction to crystal polymorph databases. N.E.J. thanks the Argonne National Laboratory Maria Goeppert Mayer Fellowship for support. G. S. would like to thank Dr. Prasanna Balaprakash for fruitful discussions on unsupervised learning and uncertainty quantification. G. S. also would like to thank Prof. George K. Thiruvathukal, and Dr. Xiao-Yong Jin for help in setting up Apache Spark workflow on Argonne Leadership Computing Facility. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Argonne National Laboratorys work was supported by the
Supporting Information Available

Computational dataset generation

We augment the original data sets of Tetko by including a variety of structural and quantum-chemical descriptors. The generation of these quantities begins with a list of SMILES strings and MPs downloaded from the OCHEM website.\textsuperscript{26} RDKit\textsuperscript{51} was used to convert SMILES strings into 3D structures using random initial coordinates, upon which Hydrogens were added and UFF energy-minimizations were performed. The minimized geometries were then used to seed B3LYP/6-31G\textsuperscript{**} geometry optimizations in Gaussian.\textsuperscript{52} The LANL2DZ pseudopotential and basis set were used for Iodine-containing molecules in the data sets. From the energy-minimized DFT geometries, total energy, HOMO/LUMO energies, SMD Solvation energy,\textsuperscript{53} dipole moment, quadrupole moment, wavefunction extent, and non-electrostatic energies were extracted. SMILES strings were also converted to Morgan Fingerprints. The quantum-chemically derived data sets used are available online as .json files.

In addition to the random coordinate generation and UFF minimization that seeded quantum-chemical calculations, we performed a 3D conformer search exploring rotatable bonds and testing cis and trans isomers.\textsuperscript{54} For each compound we produced and minimized 1500 conformers using RDKit with the MFF94 force field, and the sets of local minima were clustered to obtain a set of diverse and lower energy conformers. These conformers were used to generate 3D Morgan fingerprints for use with the supervised ML methods. A standard suite of 2D descriptors found in RDKit were also computed for all molecules in the data sets, as detailed below.
Grad-CAM with GCN

To obtain importance weights for task $y$, Grad-CAM computes the gradient of $y$ with respect to the activations of a GCN hidden layer which we denote as $A(node_i)$, i.e. $\frac{\partial y}{\partial A(node_i)}$. These gradients flowing back are global pool averaged across all nodes to obtain importance weights $\alpha_k$ for each dimension of $A(node_i) \in \mathbb{R}^K$. Using these weights for a weighted summation across the activations we arrive at an expression for Grad-CAM:

$$Grad - CAM(node_j) = \sum_k^n \alpha_k A(node_j), \quad \text{with } \alpha_k = \frac{1}{Z} \sum_j^n \frac{\partial y}{\partial A(node_j)} \quad (1)$$

To improve the interpretability of the weights, these can be $l^2$ normalized and also passed by a ReLU function to only consider positive values.

Active sampling details

The workflow deployed to achieve the outlined goals is shown in Fig. 1. The only a priori information supplied to the coarse clustering algorithms are the tanimoto similarity matrices $^1$, Ward’s minimum variance method as the metric, and a target number of coarse clusters to output. For this purpose we apply the Murtagh-Ward Clustering method$^{56}$ as implemented in RDKIT. Subsets of the MP distribution are derived from a finite number of coarse grained clusters, and are analyzed to identity and isolate data subsets of high variance. Agglomerate hierarchical clustering belongs to a class of bottom-up unsupervised learning methods,$^{39}$ and works in the following fashion. First, each of the data points are assigned to single clusters. Based on a chosen distance metric, pairs of clusters are recursively merged at each level moving up hierarchically until there is only one large coarse cluster at the highest level of clustering. From hereon we refer to this technique as coarse clustering. The resulting tree like structures can be represented by a dendrogram. Detailed overviews on hierarchical clustering algorithms are available elsewhere.$^{39,56}$ Once all of the compounds in the datasets

$^1$Note that the tanimoto similarity matrices are generated from Extended-Connectivity Fingerprints.$^{55}$
are separated among the finite coarse clusters, in a post processing step MPs are assigned to each of the clusters. Each of the coarse clusters is randomly split in a 70:30 ratio\textsuperscript{2} to create training/test data sets respectively for supervised machine learning (ML). Finally, uncertainty estimates are computed from the MP distributions derived for each of the coarse clusters and are correlated with the results of the ML model predictions.

Similarly, for generating the Butina ‘fine-grained’ clustering, the only \textit{a priori} information supplied are the Tanimoto similarity matrices and a radial cutoff. For this active sampling we used the Butina Clustering method as implement in RDKit\textsuperscript{51}, the details of which are available elsewhere.\textsuperscript{57} This method generates large numbers of ‘fine-grained’ clusters of compounds. For example, in Fig. 1, the hypothetical fine cluster ‘2’ consists of benzene as the cluster centroid grouped along with all the compounds originating from that centroid (within a radial cutoff). Clusters with insufficient data are pruned.

Since our aim is to generate realistic chemical similarity-based training/test data splits corresponding to the 70:30 ratio, only fine clusters with at least 10 compounds were considered. All clusters with insufficient numbers of compounds were pruned. Each of the individual fine clusters were randomly split in a 70:30 ratio and combined separately to create a ‘chemistry aware’ training/test split, ensuring that both training and test datasets have faithful representations of the underlying data distribution. The rationale for following such a procedure is that if the supervised ML algorithm has not seen an entire subclass or family of compounds in the training set, and all of those subclasses/families end up in the test set, then it could lead to property predictions that are unfaithful to the underlying distribution of chemical moieties. Alternatively, if a class of chemical structures are found only in the training set, then the ML algorithm could bias to minimize error associated with those species, resulting in a poor model for the remainder of the data.

\footnote{For consistency, we use this ratio throughout the study.}
Semi-supervised Variational Autoencoder

We base our VAE architecture on the implementation found in the MOSES generative benchmark. The encoder is a single layer GRU with a hidden dimension of 256 and a dropout of 0.25, while the decoder is a three layer GRU with 681 dimensions and 0.25 dropout. Decoding is a harder process than encoding and this is reflected in the complexity of each component. The latent space is of 287 dimensions. For training we utilize a learning rate schedule that cycles between 1e-2 to 1e-7 each 15 epochs. For the semi-supervised component of the network an mlp with two hidden layers was co-trained on the latent space for property prediction. The VAE loss was jointly annealed with the regressor loss, and was trained on the ‘All’ data set for maximum future predictive power. For datapoints for which we did not have labeled MP, we mask the loss to 0, therefore only computing the regression loss on labeled data. A Bayesian optimization approach was used for the tuning of hidden layer dimensions, associated drop outs, and latent space dimensions.

Distributed machine learning random forest regression

Our expanded data set of Tetko consists of the inclusion of minimum energy geometries derived from conformer searches, high-quality DFT relaxed geometries, and quantum-chemical properties. Previous investigations using 3D structures were restricted by the high dimensionality of 3D descriptors such as the Coulomb matrix (CM), which has row vectors of dimension going as the square of the maximum number of atoms encountered in the dataset, $\max(N_{atom})^2$. For example in the QM9 dataset, $\max(N_{atom}) = 29$, which is far smaller than what we encounter in our dataset, $\max(N_{atom}) = 155$. To overcome this challenge we implemented a dedicated distributed computing work flow based on Apache Spark that can exploit leadership class supercomputers. Apache Spark also comes with a native ML library. Given the high dimensionality of the Coulomb matrix, an initial hyperparameter tuning was performed by means of the spark-sklearn module over a small subset of 3000 randomly drawn compounds from the overall dataset. For further tuning of hyperparameters
during individual runs, the work flow can perform the automatic model selections for two ML
algorithms namely: random forests\textsuperscript{65} and gradient-boosted trees.\textsuperscript{66} A total of six descriptors
are available in this workflow, including Extended Connectivity Fingerprints (ECFP) bits,
sorted Coulomb matrices (CM), CM + solvation energy (CMSE), ECFP + solvation energy
(ECFPSE), CM + ECFP + solvation energy (ECFPCMSE), and ECFP + extended DFT
properties (ECFPSEext). High quality DFT relaxed geometries were used for generation of
the CM descriptor. The workflow was benchmarked on the QM9 dataset.\textsuperscript{61,67} By default
the workflow utilizes passive sampling (i.e. random split). For the descriptor that gave
the best result with the passive sampling, a chemistry aware active sampling pipeline was
implemented as shown in Fig. 1.

**Graph Convolution Neural Networks**

SMILES strings are converted to molecular graphs using the molecular graph featurization
implemented in DeepChem. A graph convolutional neural network (GCN) is used to regress
MP to the molecular graph representations. Hyperparameter optimization was performed
for each data set over the number of convolutional layers, number of neurons per inner-atom
representation, number of neurons in the dense output layer, and batch size. A GCN with
two 256 neuron convolutional layers, a dense output layer of 128 neurons, with a batch size
of 32 exhibited the highest 5-fold cross-validation for all training, with different numbers of
training epochs unique to each data set.

**Gaussian process regression**

We performed Gaussian Process Regression (GPR) implemented in GPmol, which is based
on GPflow.\textsuperscript{68} The co-variance matrices in GPR were produced using the Jaccard index as a
distance metric between vectors produced from fingerprints and descriptors. We used a 2D
Morgan circular count (ECFP-c) from SMILES strings produced with a bit size of 2048 and
radius 4. 3D descriptors were created using Morgan 3D fingerprint (E3FP) \textsuperscript{69} and MORSE \textsuperscript{70}
using the Cartesian coordinates from both 1) the B3LYP (E3FPg and MORSEg) and 2) from extensive conformer search (E3FP and MORSE), as described in section Method-Quantum Chemistry and 3D structure section. Additionally, 108 custom bioinformatics features were calculated using the RDKit package and properties derived from DFT simulations, including Total energy, HOMO-LUMO energy gap, dipole and quadrupole moments, and solvation energies. For all supervised learning algorithms, extensive hyperparameter searches were performed in order to determine the optimal inclusion of input features for presentation in the final regression results.

108 RDKit Features Used = NHOHCount, NO Count, NumAliphatic Carbocycles, NumAliphaticHeterocycles, NumAliphaticRings, NumAromaticCarbocycles, NumAromaticHeterocycles, NumAromaticRings, NumHAcceptors, NumHDonors, NumHeteroatoms, NumRadicalElectrons, NumRotatableBonds, NumSaturatedCarbocycles, NumSaturatedHeterocycles, NumSaturatedRings, NumValenceElectrons, qed, TPSA, MolMR, BalabanJ, BertzCT, fr_AlOH, fr_AlOH_noTert, fr_ArN, fr_Ar_COO, fr_ArN, fr_Ar_NH, fr_Ar_OH, fr_COO, fr_COO2, fr_C_O, fr_C_O_noCOO, fr_C_S, fr_OCCCN, fr_Imine, fr_NH0, fr_NH1, fr_NH2, fr_N_O, fr_Ndealkylation1, fr_Ndealkylation2, fr_Nhpyrrole, fr_SH, fr_aldehyde, fr_alkyl_carbamate, fr_alkyl_halide, fr_allylic_oxid, fr_amide, fr_amidine, fr_aniline, fr_aryl_methyl, fr_azide, fr_azo, fr_barbitur, fr_benzene, fr_benzodiazepine, fr_bicyclic, fr_diazo, fr_dihydropyridine, fr_epoxide, fr_ester, fr Ether, fr_furan, fr_guanido, fr_halogen, fr_hdrzine, fr_hdrzone, fr_imidazole, fr_imide, fr_isocyan, fr_isothiocyan, fr_ketone, fr_ketone_Topliss, fr_lactam, fr_lactone, fr_methoxy, fr_morpholine, fr_nitrile, fr_nitro, fr_nitro_atom, fr_nitro_arom_nonortho, fr_nitroso, fr_oxazole, fr_oxime, fr_para_hydroxylation, fr_phenol, fr_phenol_noOrthoHbond, fr_phos_acid, fr_phos_ester, fr_piperidine, fr_piperazine, fr_priamide, fr_prisulfonamd, fr_pyridine, fr_quatN, fr_sulfide, fr_sulfonamd, fr_sulfone, fr_term_acetylene, fr_tetrazole, fr_thiazole, fr_thiocyran, fr_thiophene, fr_unbrch_alkane, fr_urea, MolWt, Mol-LogP

Quantum-Chemical Properties from DFT Calculations = Total energy (TotE), HOMO-LUMO energy gap (gap), SMD Solvation energy (Solv), dipole moment (dipol),
quadrupole (quadp), wavefunction extent, and non-electrostatic energies.

**Dataset Abbreviations**

**props** = [TotE, Solv, gap, dipol, quadpl]

**Features** = RDKit Features

**Combo** = Features + props

**ECFP** = Extended Connectivity Fingerprints bits

**CM** = Sorted Coulomb Matrices

**CMSE** = CM + Solv

**ECFPSE** = ECFP + Solv

**ECFPCMSE** = CM + ECFP + Solv

**ECFPSEext** = ECFP + Quantum-Chemical Properties

This material is available free of charge via the Internet at [http://pubs.acs.org/](http://pubs.acs.org/).

**References**

1. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry* **2014**, *57*, 4977–5010.

2. 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.

3. Szymański, P.; Markowicz, M.; Mikiciuk-Olasik, E. Adaptation of High-Throughput Screening in Drug Discovery—Toxicological Screening Tests. *International Journal of Molecular Sciences* **2011**, *13*, 427–452.

4. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The
Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *Journal of Physical Chemistry Letters* **2011**, *2*, 2241–2251.

(5) Yan, Q.; Yu, J.; Suram, S. K.; Zhou, L.; Shinde, A.; Newhouse, P. F.; Chen, W.; Li, G.; Persson, K. A.; Gregoire, J. M.; Neaton, J. B. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proceedings of the National Academy of Sciences* **2017**, *114*, 3040–3043.

(6) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci* **2001**, *41*, 1208–1217.

(7) Tetko, I. V. Associative neural network. *Methods Mol. Biol.* **2008**, *458*, 185–202.

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

(9) Nikmo, J.; Kukkonen, J.; Riikonen, K. A model for evaluating physico-chemical substance properties required by consequence analysis models. *J. Hazard. Mater.* **2002**, *91*, 43–61.

(10) Zhang, Y.; Maginn, E. J. A comparison of methods for melting point calculation using molecular dynamics simulations. *J. Chem. Phys.* **2012**, *136*, 144116.

(11) Nyman, J.; Day, G. M. Modelling temperature-dependent properties of polymorphic organic crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31132–31143.

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

(13) 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.

(14) Tetko, I. V.; Lowe, D. M.; Williams, A. J. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. *J. Cheminf.* 2016, 8, 2.

(15) Withnall, M.; Chen, H.; Tetko, I. V. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. *ChemMedChem* 2018, 13, 599–606.

(16) Bhhatarai, B.; Teetz, W.; Liu, T.; Oberg, T.; Jeliazkova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E.; Gramatica, P. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Mol. Inf.* 2011, 30, 189–204.

(17) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. Melt- ing point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *J. Chem. Inf. Model.* 2006, 46, 2412–2422.

(18) Krstajic, D.; Buturovic, L. J.; Leahy, D. E.; Thomas, S. Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminf.* 2014, 6, 10.

(19) Sahlin, U.; Jeliazkova, N.; Oberg, T. Applicability domain of dependent predictive uncertainty in QSAR regressions. *Mol. Inf.* 2014, 33, 26–35.

(20) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* 2005, 45, 581–590.

(21) Bergstrom, 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.
(22) 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.

(23) Brown, T. N.; Armitage, J. M.; Arnot, J. A. Application of an Iterative Fragment Selection (IFS) Method to Estimate Entropies of Fusion and Melting Points of Organic Chemicals. *Molecular informatics* **2019**, *38*, 1800160.

(24) Bradley, J.-C.; Lang, A.; Williams, A. Jean-Claude Bradley Double Plus Good (Highly Curated and Validated) Melting Point Dataset. 2014; [https://figshare.com/articles/Jean_Claude_Bradley_Double_Plus_Good_Highly_Curated_and_Validated_Melting_Point_Dataset/1031638/1](https://figshare.com/articles/Jean_Claude_Bradley_Double_Plus_Good_Highly_Curated_and_Validated_Melting_Point_Dataset/1031638/1) (accessed Sep 20, 2017).

(25) ENAMINE Ltd. [http://www.enamine.net](http://www.enamine.net) (accessed Sep 20, 2017).

(26) OCHEM. [http://www.ochem.eu](http://www.ochem.eu) (accessed Sep 20, 2017).

(27) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning* **2017**, *70*, 1263–1272.

(28) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gómez-Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* **2015**, 2215–2223.

(29) 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 central science* **2018**, *4*, 268–276.
(30) Irwin, J. J.; Shoichet, B. K. ZINC A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 177–182.

(31) eMolecules Plus Database Download. https://www.emolecules.com/info/plus/download-database (accessed Jun 26, 2019).

(32) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *2019, arXiv:1905.13741*, arXiv.org e–Print archive, https://arxiv.org/abs/1905.13741 (accessed Sep 4, 2019).

(33) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous molecular design: Then and now. *ACS applied materials & interfaces* **2019**, *11*, 24825–24836.

(34) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. Using Attribution to Decode Dataset Bias in Neural Network Models for Chemistry. *2018, arXiv:1811.11310*, arXiv.org e–Print archive, https://arxiv.org/abs/1811.11310 (accessed Sep 4, 2019).

(35) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. GradCAM: Visual Explanations From Deep Networks via Gradient-Based Localization. *The IEEE International Conference on Computer Vision (ICCV)* **2017**, 618–626.

(36) Preuer, K.; Klambauer, G.; Rippmann, F.; Hochreiter, S.; Unterthiner, T. Interpretable Deep Learning in Drug Discovery. *2019, arXiv:1903.02788*, arXiv.org e–Print archive, http://arxiv.org/abs/1903.02788 (accessed Sep 4, 2019).

(37) Dasgupta, S.; Hsu, D. Hierarchical sampling for active learning. *Proceedings of the 25th international conference on Machine learning* **2008**, 208–215.
(38) Dasgupta, S. Two faces of active learning. *Theoretical computer science* **2011**, 412, 1767–1781.

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

(40) Wille, R. L.; Lutton, E. S. Polymorphism of cocoa butter. *Journal of the American Oil Chemists Society* **1966**, 43, 491–496.

(41) Burger, A.; Ramberger, R. On the polymorphism of pharmaceuticals and other molecular crystals. II. *Microchim. Acta* **1979**, 72, 273–316.

(42) Yu, L. Inferring Thermodynamic Stability Relationship of Polymorphs from Melting Data. *J. Pharm. Sci.* **1995**, 84, 966–974.

(43) Gill, P.; Moghadam, T. T.; Ranjbar, B. Differential Scanning Calorimetry Techniques: Applications in Biology and Nanoscience. *J. Biomol. Tech.* **2010**, 4, 167–193.

(44) Vyazovkin, S.; Chrissafis, K.; Lorenz, M. L. D.; Koha, N.; Pijolat, M.; Roduit, B.; Sbirrazzuoli, N.; Sunol, J. J. ICTAC Kinetics Committee recommendations for collecting experimental thermal analysis data for kinetic computations. *Thermochim. Acta.* **2014**, 590, 1–23.

(45) Hellmuth, E.; Wunderlich, B. Superheating of Linear High-Polymer Polyethylene Crystals. *J. App. Phys.* **1965**, 36, 3039.

(46) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, 57, 1757–1772.

(47) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, 9, 513–530.
(48) Feunang, Y. D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics* 2016, 8, 61.

(49) Kondor, R. N-body Networks: a Covariant Hierarchical Neural Network Architecture for Learning Atomic Potentials. 2018, *arXiv:1803.01588*, arXiv.org e–Print archive, https://arxiv.org/abs/1803.01588 (accessed Sep 4, 2019).

(50) Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. 2018, *arXiv:1802.08219*, arXiv.org e–Print archive, https://arxiv.org/abs/1802.08219 (accessed Sep 4, 2019).

(51) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed Jan 15, 2018).

(52) Frisch, M. J. et al. Gaussian16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(53) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* 2009, 113, 6378–6396.

(54) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *Journal of Chemical Information and Modeling* 2012, 52, 1146–1158.

(55) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* 2010, 50, 742–754.
(56) Murtagh, F.; Contreras, P. Methods of hierarchical clustering. 2011, arXiv:1105.0121, arXiv.org e–Print archive, https://arxiv.org/abs/1105.0121 (accessed Sep 4, 2019).

(57) Butina, D. Unsupervised Data Base Clustering Based on Daylight’s Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. Journal of Chemical Information and Computer Sciences 1999, 39, 747–750.

(58) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Others, Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. 2018, arXiv:1811.12823, arXiv.org e–Print archive, https://arxiv.org/abs/1811.12823 (accessed Sep 4, 2019).

(59) GPyOpt : A Bayesian Optimization framework in python. 2016; http://github.com/SheffieldML/GPyOpt (accessed Dec 10, 2018).

(60) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. Physical review letters 2012, 108, 058301.

(61) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data 2014, 1, 140022.

(62) Zaharia, M.; Xin, R. S.; Wendell, P.; Das, T.; Armbrust, M.; Dave, A.; Meng, X.; Rosen, J.; Venkataraman, S.; et al Franklin, M. J. Apache spark: a unified engine for big data processing. Communications of the ACM 2016, 59, 56–65.

(63) Meng, X.; Bradley, J.; Yavuz, B.; Sparks, E.; Venkataraman, S.; Liu, D.; Freeman, J.; Tsai, D.; Amde, M.; et al Owen, S. Mllib: Machine learning in apache spark. The Journal of Machine Learning Research 2016, 17, 1235–1241.
(64) Scikit-learn integration package for Apache Spark. https://github.com/databricks/spark-sklearn (accessed Mar 5, 2018).

(65) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

(66) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, *1189*-1232.

(67) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Lilienfeld, O. A. V. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **2014**, *1*, 140022.

(68) De, A. G.; Matthews, G.; Van Der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrá, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian Process Library using TensorFlow. *Journal of machine learning research: JMLR* **2017**, *18*, 1–6.

(69) Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *Journal of medicinal chemistry* **2017**, *60*, 7393–7409.

(70) Todeschini, R.; Consonni, V. *Handbook of Chemoinformatics*; John & Sons, Ltd, 2008; pp 1004–1033.
Graphical TOC Entry
