A machine learning workflow for molecular analysis: application to melting points

MACHINE LEARNING
Science and Technology

A machine learning workflow for molecular analysis: application to melting points

Ganesh Sivaraman[1,10], **Nicholas E Jackson**[2,3,10], **Benjamin Sanchez-Lengeling**[4], **Álvaro Vázquez-Mayagoitia**[5], **Alán Aspuru-Guzik**[6,7,8,9], **Venkatram Vishwanath**[1] **and Juan J de Pablo**[3,2]

[1] Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL 60439, United States of America
[2] Center for Molecular Engineering, Argonne National Laboratory, Lemont, IL 60439, United States of America
[3] Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637, United States of America
[4] Department of Chemistry, Harvard University, Cambridge, MA 02138, United States of America
[5] Computational Science Division, Argonne National Laboratory, Lemont, IL 60439, United States of America
[6] Department of Chemistry, University of Toronto, Toronto, ON M5S 3H6, Canada
[7] Department of Computer Science, University of Toronto, Toronto, ON M5S 3H6, Canada
[8] Vector Institute for Artificial Intelligence, Toronto, Ontario, M5S 1m1, Canada
[9] Fellow, Canadian Institute for Advanced Research, Toronto, Ontario, M5G 1Z8, Canada
[10] Contributed equally to this work

E-mail: jacksone@anl.gov **and** depablo@uchicago.edu

Keywords: materials, melting point, machine learning, workflow

OPEN ACCESS

RECEIVED
20 November 2019

REVISED
23 March 2020

ACCEPTED FOR PUBLICATION
17 April 2020

PUBLISHED
22 June 2020

Abstract

Computational tools encompassing integrated molecular prediction, analysis, and generation are key for molecular design in a variety of critical applications. In this work, we develop a workflow for molecular analysis (MOLAN) that integrates an ensemble of supervised and unsupervised machine learning techniques to analyze molecular data sets. The MOLAN workflow combines molecular featurization, clustering algorithms, uncertainty analysis, low-bias dataset construction, high-performance regression models, graph-based molecular embeddings and attribution, and a semi-supervised variational autoencoder based on the novel SELFIES representation to enable molecular design. We demonstrate the utility of the MOLAN workflow in the context of a challenging multi-molecule property prediction problem: the determination of melting points solely from single molecule structure. This application serves as a case study for how to employ the MOLAN workflow in the context of molecular property prediction.

1. Introduction

Efficient computational tools capable of connecting single molecule structure with bulk multi-molecule structure and/or function are crucial to nearly all molecular design efforts. Historically, machine learning (ML) methods including quantitative activity-structure relationships [1] (QSAR) and quantitative property-structure relationships [2] (QSPR) methods have played a primary role in *in silico* predictive molecular modeling, with considerable success across a broad array of prediction tasks including molecular solubility, biological toxicity, and thermophysical properties [1]. These classes of data-driven, quantitatively predictive models, when incorporated with high-throughput screening and design efforts, can aid characterization and generation of molecular species with applications in drug design [3], organic electronics [4], and solar fuels materials [5], among many others.

The recent proliferation of ML techniques for molecular prediction, analysis, and design has led to rapid advances in researchers' abilities to understand and design molecular systems [6, 7]. Emerging graph-based featurization techniques have allowed for unprecedented accuracy in an array of molecular regression and classification tasks [8]. A broad ensemble of inverse design methods, enabled by advances in generative ML techniques, show promise for the de Novo creation of molecular structures with targeted properties [9]. Unsupervised learning techniques have been applied in contexts where complex molecular data sets need to be analyzed and sorted with minimal bias in order to discern underlying molecular mechanisms [10]. With the broad expertise required to both understand and benefit from this diversity of ML techniques, it can be

difficult to apply these tools for practical applications to molecules and materials of interest. In this work, we introduce a workflow known as MOLAN, which applies a diverse set of tools to tackle molecular analysis and design problems using ML. With the aim of transparency and broad applicability, the code repository to build models and make predictions, along with datasets themselves, can be found online [11].

To demonstrate the utility of the MOLAN workflow, we have selected an experimental dataset and topic of interest to a variety of industrial applications: the prediction of a molecule's melting point (MP) solely from its molecular structure. 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 [12, 13]. 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 [2, 14]. MP can also be well-correlated with a liquid's viscosity [15]. In any application where high-throughput screening is an avenue for material discovery, accurate MP prediction will determine the scope of practical candidate materials, and significant progress has been made in this field which allows for a presentation of our results in the context of previous efforts. [2, 16–26]

The outline of the paper is as follows. First, we introduce the motivation for, and steps of, the MOLAN workflow, and provide details regarding its implementation [11]. We then apply the MOLAN workflow to the MP dataset assembled by Tetko [16]. Specific attention is payed to the role of unsupervised learning and the stratified sampling of chemical space, as well as the high-accuracy of MP regression results. A literature search is performed that suggests that the predicted accuracy obtained using the MOLAN workflow is comparable to the fundamental limit derived from a consideration of the underlying experimental uncertainties and the presence of crystal polymorphs. This suggests that future improvements in prediction accuracy will likely be derived from explicit consideration of the 3D molecular structure of the crystals. We conclude with a discussion of the MOLAN workflow, as well as key features learned from its application to the MP problem.

2. Methodology

2.1. Motivation for the MOLAN workflow

Provided the many diverse developments in ML applied in disparate molecular contexts, the MOLAN workflow aims to collect a useful subset of these methodologies that can be applied 'off-the-shelf' to molecular systems of interest. With this aim, we have identified seven components of molecular ML workflows common to a variety of applications: Molecular Featurization, Chemical Clustering and Dataset Analysis, Assessing Intrinsic Dataset Uncertainties, Low-Bias Dataset Generation, Regression Models, Molecular Attribution and Embedding, and Geometric Spaces and Inverse Design. We have constructed the MOLAN workflow with the goal of integrating these components in a fashion that can be applied to any molecular dataset of interest. Of course, given the broad range of molecular ML methods in the literature, what follows is naturally influenced by our own biases, and is not an absolute endorsement of the 'best' methodologies - it is simply a straightforwardly accessible combination of powerful methods that we deem highly useful in any molecular ML application. In figure 1 we illustrate this workflow schematically, by incorporating these seven components into three stages.

The first stage of figure 1 (Molecular Featurization, Clustering and Analysis of Molecular Datasets, Assessing Intrinsic Dataset Uncertainties) concerns exploratory molecular analysis and involves (i) including a variety of common molecular featurizations, (ii) understanding the underlying chemical and property distributions of the dataset, and (iii) characterizing the anticipated limitations to prediction accuracy of the observables trying to be predicted. This exploratory molecular analysis and characterization of the data is critical to all subsequent ML tasks, and its importance cannot be overstated when it comes to enabling high-accuracy regression, attribution, and inverse design models that follow later in the workflow.

The second stage (Low-Bias Dataset Generation, Regression Models) addresses the regression tasks common to many molecular ML applications. Here, we incorporate a low-bias dataset generation step, aimed at uniformly sampling chemical space. As many chemical datasets are highly clustered in chemical space, the bias introduced by this fact can degrade the performance (as well as the transferability) of trained regression models. By applying a uniform stratified sampling of chemical space, the MOLAN workflow directly addresses this bias. Following this task, we then apply three regression models (Random Forests, Graph Convolutional Neural Networks, Gaussian Process Regression) that have proven useful in our previous molecular regression tasks; this is by no means an exhaustive list of high-performing methods and represents our personal bias, with the addition that these methodologies are relatively straightforward to apply. To be explicit, this stage follows the exploratory molecular analysis section which is critical for understanding the

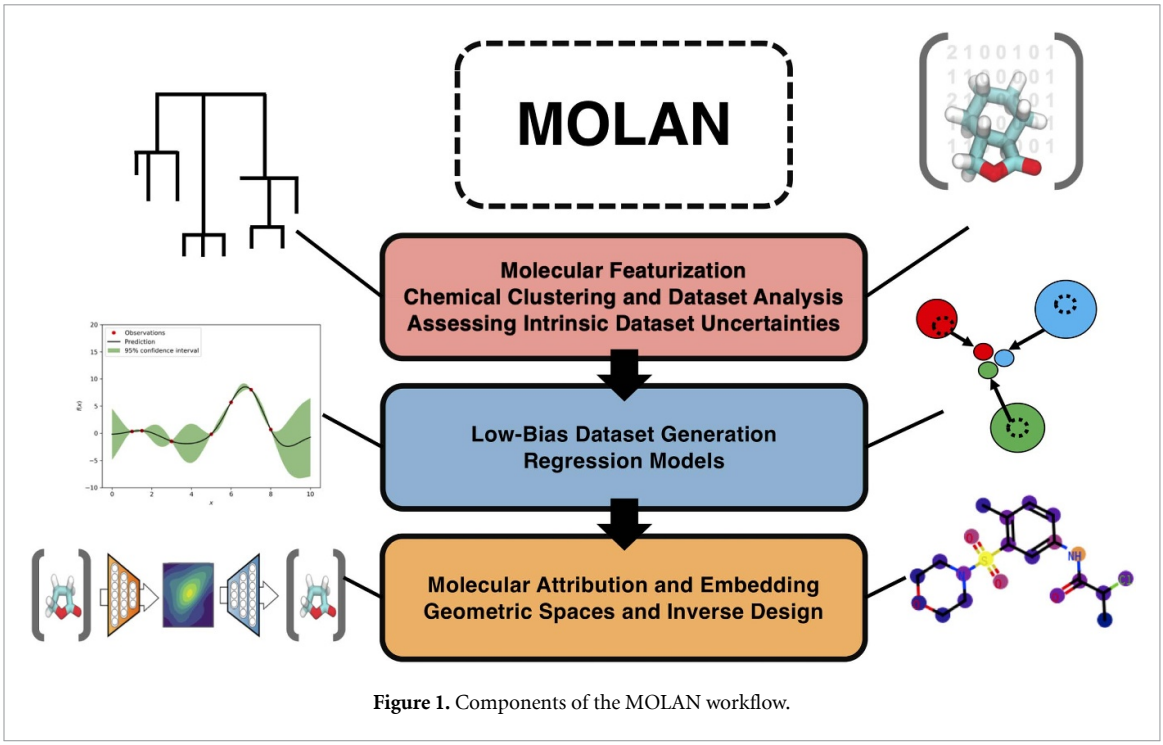

Figure 1. Components of the MOLAN workflow.

statistical properties of the dataset that is to be modeled. It also proves essential to have high-performing regression tools in order to enable the most powerful application based tools that follow in the next stage.

As a third stage (Molecular Attribution and Embedding, and Geometric Spaces and Inverse Design), we integrate techniques aimed at understanding specific chemical contributions to property prediction (graph attribution), as well as the generation of molecular species targeted to specific properties of interest. Given the often small sizes of experimental datasets, and the limitations this potentially imposes on many advanced ML techniques, of particular interest is MOLAN's use of a semi-supervised variational autoencoder (VAE) that leverages the existing emolecules database [27]. By leveraging the chemical structures in this dataset in such a fashion, we can provide enough information regarding the description of the chemical environment which is necessary for inverse design, and then fine-tune it using the smaller property-based dataset of interest.

It is our belief that applying this workflow to molecular datasets will reap large benefits, particularly for beginners who do not possess the requisite ML expertise required to create such a workflow from scratch.

2.2. Molecular featurization

Molecular featurization is the means by which chemical information is translated into a numerical and machine-readable format for use as input into ML algorithms. Due to the high-accuracy of supervised learning techniques using graph-based featurizations [8], as well as the success of physically motivated cheminformatics descriptors, the MOLAN workflow utilizes graph-based molecular featurizations, 2D and 3D Morgan Fingerprints, Coulomb Matrices, as well as a collection of cheminformatics properties derived from RDKit [28] and single molecule electronic structure derived from quantum-chemical calculations. This diverse set of featurizations was selected to (i) cover a broad range of chemical featurizations taken from RDKIT (See SI) relevant to classic cheminformatics melting point predictions [16], (ii) include features recently used for high-performance molecular regression tasks (Coulomb Matrices, graph-based representations, and fingerprinting techniques) [1, 8, 29], and (iii) incorporate quantum-chemically motivated quantities of relevance to intermolecular interactions.

In the context of the specific application to MP demonstrated in this work, we utilized the following quantum-chemically derived quantities (further details regarding quantum-chemical calculations can be found in the SI): total energy (TotE), HOMO-LUMO energy gap (gap), Solvation energy (Solv), dipole moment (dipol), quadrupole moment (quadp), and wavefunction extent. These properties were selected via their anticipated influence on intermolecular interactions, which will be critical to describing MP: valence and conduction band energy levels will influence polarization, and electrostatic multipole moments and wavefunction extent will critically influence intermolecular interaction strength. In the context of solvation energies, we are specifically interested in exploring the question of how this single molecule quantity, which utilizes a continuum description of the solvation environment, can be related to MP predictions.

2.3. Chemical clustering and dataset analysis

One key aspect of the MOLAN workflow is the emphasis on the application of unsupervised clustering techniques for data set analysis and refinement. In this regard, a first step in the MOLAN workflow applies techniques from unsupervised learning to coarsely cluster molecular data sets into finite subsets with similar molecular structures. Specifically, MOLAN uses information supplied via the Tanimoto similarity index matrices, Ward's minimum variance method as a metric (MOLAN uses the Murtagh-Ward clustering method [30] as implemented in RDKit), and a target number of coarse clusters to accomplish this task. Tanimoto similarity was utilized as a featurization for unsupervised clustering due to its use in the original Butina clustering work [31], and was also applied in the context of Murtagh-Ward to maintain consistency in featurization among the unsupervised clustering methods. The SMILES strings are converted into extended connectivity fingerprints (ECFP). Tanimoto similarity matrices are generated from the ECFP [32]. Subsets of the molecular data set can then be drawn from a finite number of coarse-grained clusters and analyzed to identify data subsets of high variance. Murtagh-Ward clustering belongs to a class of bottom-up unsupervised learning methods known as agglomerate hierarchical clustering [33]. Detailed overviews on hierarchical clustering algorithms are available elsewhere [30, 33].

Once all of the compounds in the datasets are separated among the finite coarse clusters, in a post processing step molecular properties are assigned to each of the clusters. Uncertainty estimates are computed from molecular property distributions derived for each of the coarse clusters and are correlated with ML regression model predictions (coarse clusters are randomly split in a 70:30 ratio to create training/test data sets) in order to quantify inherent variance in prediction accuracy, without taking into account explicit chemical identity. In a last step, in order to quantify inherent chemical variance and its relationship to the molecular prediction task, the international chemical identifier key (inchikey) of the compounds from the coarse clusters are extracted and parsed through the *'ClassyFire'* automated chemical classification server [34]. The underlying chemical distributions of the coarse clusters can then be linked to their regression performance to quantify statistical properties of the data set related to specific chemical motifs.

2.4. Assessing intrinsic dataset uncertainties

Another critical feature of the MOLAN workflow is an emphasis on accessing the intrinsic uncertainties of the underlying datasets to be modeled. An important aspect of most molecular ML projects is trying to regress a property of interest as accurately as possible. However, due to the potential for overfitting in many ML methodologies, one often needs a firm grasp of how the dataset was generated, as well as its underlying physics, in order to assess what kind of predictive accuracies can be expected. While there is no prescriptive methodology applicable to every type of dataset, we believe this can be addressed by extensive literature searches that lead to improved physical understandings. In applying the MOLAN workflow to MP prediction, we illustrate this point by conducting an extensive analysis of underlying sources of error for experimentally-derived small molecule MP data, taking into account variable MPs resulting from multiple crystal polymorphs, the underlying accuracies of the experiments themselves, as well as the quality of data recording and chemical stability. This analysis is performed and included in the SI, with its most salient features summarized in the main text.

2.5. Low-Bias dataset generation

One strength of the MOLAN workflow is the application of unsupervised learning to develop low bias chemical datasets. As many molecular datasets can be composed of numerous chemically redundant data points, the ability to actively sample existing data sets in order to generate reduced dataset sizes that have more uniform samplings of chemical space is highly desirable. In this regard, the MOLAN workflow employs a uniform stratified sampling procedure on the chemical space of data sets. An important approach in active sampling in this regard is the exploitation of the cluster structure within the underlying dataset [35, 36]. To this end, the MOLAN workflow employs the Butina clustering algorithm [31] as implemented in RDKIT [28]. The only *a priori* information supplied to the Butina clustering algorithm is the Tanimoto similarity matrices and a radial cutoff. This method generates large numbers of 'fine-grained' clusters of compounds. Clusters with insufficient data are pruned. Once these 'fine-grained' clusters of chemical motifs are established, they can then be uniformly sampled from to generate a smaller data set with reduced chemical bias.

Since the introduction of the MOLAN workflow and its application to MP prediction is the goal of this work, we aim to generate realistic chemical similarity-based training/test data splits corresponding to a 70:30 ratio - due to the large size of our MP data set (~40 k) only fine clusters with at least 10 compounds were considered. All clusters with insufficient numbers of compounds were pruned. Each of the individual fine chemical clusters were randomly split in a 70:30 ratio and combined separately to create a uniform stratified sampled 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.

2.6. Regression models

For regression tasks the MOLAN workflow employs three supervised learning algorithms, all with high-accuracy regression performance. Each model encompasses a different type of modeling approach for QSPR applications. Random forests (RF) are employed as a common and robust regression strategy. Gaussian Process Regression (GPR) is employed as a method capable of including prediction uncertainties, and Graph Convolutional Neural Networks (GCN), a type of Graph Neural Network (GNN), are employed as a state-of-the-art, featurization agnostic regression technique.

2.6.1. Random forests

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 are straightforward to apply. For consistency, we use a training/test split ratio of 70:30 throughout this study. Mean absolute error (MAE) in the prediction is used as the evaluation metric. The standard deviation in the prediction error as derived from 5-fold cross validation are also reported in parenthesis along with the MAE in relevant results tables. A number of 2D and 3D descriptors described in the Molecular Featurization subsection are carefully benchmarked, the results of which can be seen in figure 4(a). The best performing descriptor from this benchmark is chosen for the final regression.

2.6.2. Gaussian process regression

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. The MOLAN workflow uses Gaussian Process Regression (GPR) implemented in GPmol, which is based on GPflow. [38] The co-variance matrices in GPR were produced using the Jaccard index as a distance metric between vectors produced from fingerprints and features. We produced 2D and 3D fingerprints and features provided by the RDkit package. For the final regressions we used a subset of those descriptors after we performed a benchmark to select them, details of which can be seen in the results section surrounding figure 4(b). 2D fingerprints were generated with Morgan circular count (ECFP-c) from SMILES strings producing a vector size of 2048 and radius 4, while 3D descriptors were created using a Morgan-like 3D fingerprint (E3FP) [39] and MORSE [40] using the Cartesian coordinates from both (1) quantum-chemically computed (E3FPg and MORSEg) and (2) from extensive conformer search geometries (E3FP and MORSE), as described in the SI). GPR also utilized 108 custom bioinformatics features calculated using the RDKit package [28] 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 (see figure 4b).

2.6.3. Graph convolution neural networks

GCN [8] utilize a graph-structured representation of a molecule, with atoms as nodes and bonds as edges of a graph, as opposed to both RF and GPR relying on predefined features (e.g. fingerprints, quantum-chemical properties, etc) to represent molecular structures. 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 a 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 [37]. All graph operations are designed to preserve graph symmetries. In the MOLAN workflow, SMILES strings are converted to molecular graphs using the molecular graph featurization implemented in DeepChem. A GCN is used to regress MP using 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 data, with different numbers of training epochs unique to each data set.

2.7. Molecular attribution and embedding

In order to derive chemically specific insights underlying chemical correlations, the MOLAN workflow uses 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. This feature then allows one to directly correlate molecular property prediction with specific aspects of the molecular graph, which we here refer to generally as graph attribution. This space of activations can also be used to directly construct geometric spaces within which to examine molecular structures. One key aspect of the MOLAN workflow is the ability to build interpretable predictions in GCN, which involves assigning positive or negative weights to graph elements in relation to their importance for prediction [41]. For this purpose, we utilize grad-CAM [42] with GCNs. These methodologies have been previously explored in the context of drug-like properties. [43]. Grad-CAM uses gradient information flowing into the convolutional layer of a GCN to understand the importance of each neuron for a given task.

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 α_k for each dimension of $A(node_i) \in R^K$. Using these weights for a weighted summation across the activations we arrive at an expression for Grad-CAM:

$$Grad-CAM(node_j) = \sum_{k}^{n} \alpha_k A(node_j), \text{ with } \alpha_k = \frac{1}{Z} \sum_{j}^{n} \frac{\partial y}{\partial A(node_j)} \tag{1}$$

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. By normalizing the information delivered by Grad-CAM, 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.

2.8. Geometric spaces and inverse design

The MOLAN workflow constructs geometric spaces structured around desired molecular properties (in our case, MP) that allow one to better understand how molecules are structured, as well as serve as a sanity check for when particular 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 principal components analysis (PCA). To construct these geometric spaces, we utilize semi-supervised variational autoencoders (SSVAE). 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* [44] first demonstrated the usage of VAE with SMILES strings to generate new molecules with drug-like properties using the Zinc [45] dataset. One key result was the ability of the VAE to shape the organization of the latent space representation of molecules based on the predicted properties of interest.

One challenge for VAE is their requirement of a large amount of chemical data in order to be able to generalize to new molecules. Since molecular data sets, particularly experimental, are often of limited size, the MOLAN workflow critically relies on semi-supervised learning to leverage larger unlabeled data sets. Because we i) do not want all molecular applications to be limited by small data set sizes and ii) want the latent space in any particular application to be informed by molecules that have been synthesized and exist on a shelf somewhere in the world, the MOLAN workflow uses a set of 1 M purchasable molecules from emolecules to inform the chemical structure of the latent space [27]. The MOLAN workflow then trains the VAE with a mix of labeled and unlabeled data from both the emolecules and the property prediction data set: 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 [46]. Our geometric space is then the latent space of our SSVAE, and its principal components can be examined to analyze and understand the chemical diversity of the underlying data set. Similarly, once this latent space has been constructed and organized according to a specific molecular property, the latent space can be decoded to realize the de Novo generation of new molecular species with targeted properties.

In the context of the MP application in this work, we have a MP predicting neural network that maps the latent space to predicted MPs. We base our VAE architecture on the implementation found in the MOSES generative benchmark [47]. 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 (multi-layer perceptron) 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. The regressor loss was annealed linearly from 0 to 1 and the KL term was annealed cyclically to prevent mode collapse [48]. For datapoints for which we did not have labeled MP, we mask the loss to zero and only compute the regression loss on labeled data. Each batch had a ratio of 20:1 unlabeled/labeled datapoints. A Bayesian optimization [49] approach was used for the tuning of hidden layer dimensions, associated drop outs, and latent space dimensions.

3. Results and discussion

3.1. MP Datasets and experimental uncertainties

Four publicly available data sets of experimental MPs were chosen for this study, as outlined by Tetko [16]. 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 "gold'' standard [50] for MP data sets, and has been double-validated to only contain data with multiple reported measurements within 5 K. The Bergström data set [24], 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 Enamine Ltd [51], a chemical supplier. The OCHEM data set was derived from a diverse pool of non-curated data from the Online Chemical Modeling Environment (OCHEM) [52]. Note that 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 Methods section.

To accurately model molecular data, it is critical to have an assessment of the underlying errors and inherent limits to prediction accuracy for the ML methodology. A thorough characterization of these contributing factors is a key element of the MOLAN workflow. In the context of experimental MP, there are two potentially crucial limitations to prediction accuracy: the underlying experimental error of the MP measurements and the existence of multiple crystal polymorphs (with distinct MP) for a single molecule. In order to quantify both of these contributions, we have performed an extensive error analysis in the Supporting Information. We summarize the most important points of the error analysis here: in figure 2(b) we have computed the size of the MP interval in K (ΔT_m) for 119 experimental polymorphs from the literature. To compute this interval, we take the difference between the maximum and minimum recorded MP for polymorphs of a given molecule - this provides an upper limit on potential prediction errors assuming the ML algorithm predicts a value somewhere in this range. The key result of this effort is that over 80% of molecules in this search exhibit MP distributions for experimental polymorphs bounded by 20 K, with over 96% bounded by 30 K. These points, when combined with fact that only a fraction of molecular structures will exhibit such polymorphs, suggests that polymorph induced inaccuracies are not solely responsible for limiting MP predictions. When these facts are combined with the relatively small (~1–5 K) errors anticipated for experimental errors (see SI), we note that these total errors are substantially less than the typical 35–50 K error often achieved in MP prediction tasks in previous works. Authors interested in further analysis of these datasets beyond our own should consult the seminal work of Tetko [16].

3.2. Regression for passively sampled data

We begin by analyzing the statistics of the MP data sets, as shown in figure 2(a)). 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* [16], 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 [16]. The curated nature of the BradBerg data set implies that the large standard deviation observed in

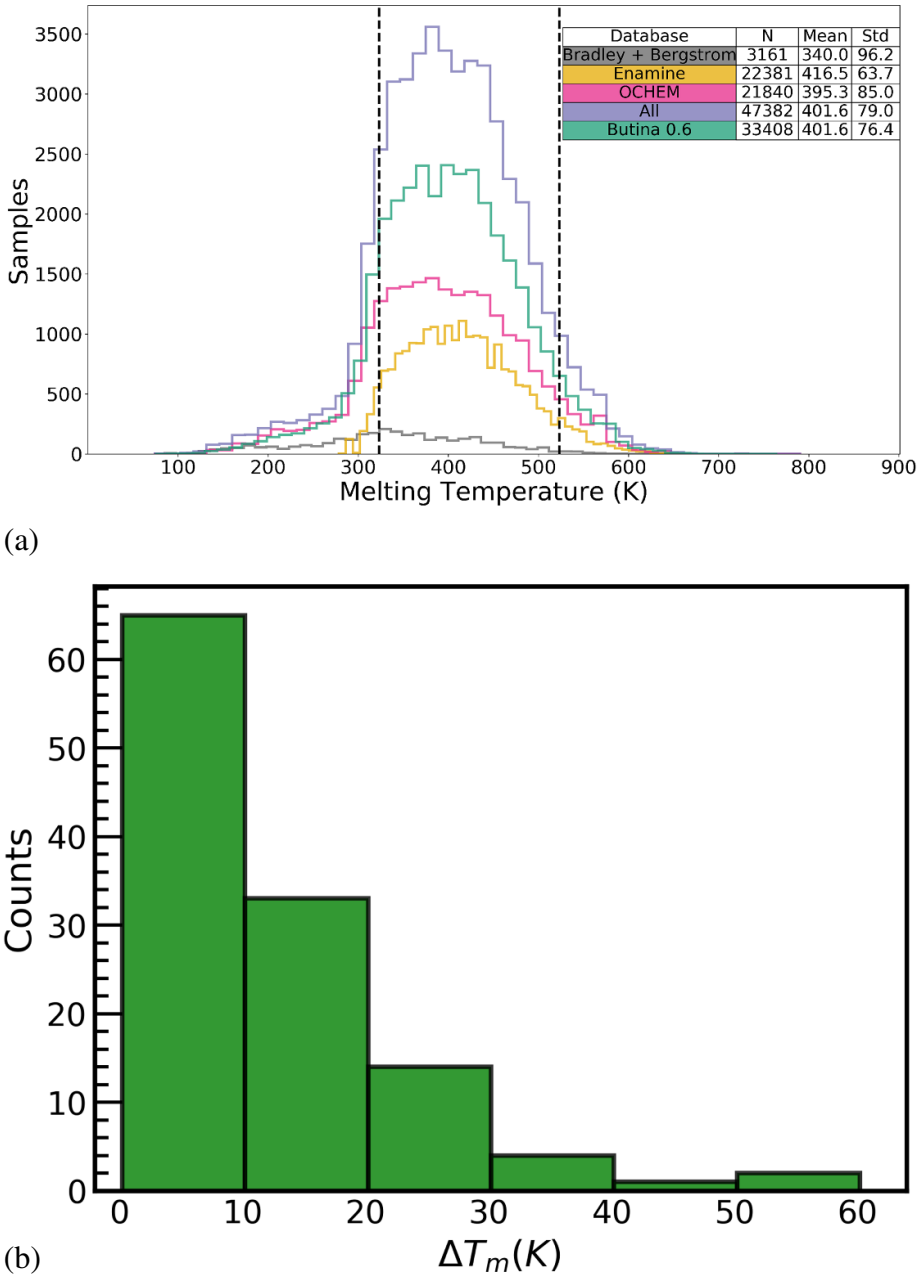

Figure 2. Experimental data sets. (a) MP distributions for experimental data sets. Labels and colors are based on data set source. Black dashed lines indicate drug-like region [323.15, 523.15]. (b) Distribution of MP intervals for 119 literature molecules exhibiting multiple crystal polymorphs.

the distribution in figure 2(a)) 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, i.e. no uniform stratified sampling of the data has occurred. The regression results for different supervised ML models trained on the passively sampled MP data sets of figure 2(a) are shown in table 1. We utilize the RF ML method with the ECFPSE featurization (see SI) 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 curated and chemically-diverse nature of the dataset, the latter of which

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

is confirmed by its large standard deviation (see figure 2a). These results emphasize the critical importance of having curated data sets, as in the cases of data sets that are not curated, including more data will not lead to better model performance.

The use of GCN and GPR on the passively sampled data sets lead to significant improvements in predictive accuracy. Specifically, for both GCN and GPR, MAE below 30 K 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 (see figure 4b). If one restricts the performance of the GPR method to only molecules in the 'drug-like' interval as described by Tetko [16], 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 [53]. This points to the promise of graph-based techniques that have been described previously [8, 54], 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.

3.3. Chemical clustering and classification of mp dataset

Provided the previous regression results derived from simple statistics and supervised ML using the passively sampled data sets, we now employ the unsupervised learning algorithms of the MOLAN workflow to both understand and actively sample the underlying chemical structures of the MP 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 Tanimoto similarity is computed from ECFP with a radius of 2 and 1024 bits. Bit lengths larger than 1024 are not found to provide any additional insights from clustering for the MP datasets. In addition, radius larger than 2 was found to be computationally not feasible for larger dataset sizes. The number of coarse clusters are analyzed for values in the range 3 to 10. Seven coarse clusters are found to be optimal for the statistical analysis of MP datasets. To reiterate, 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	Size	Mean	Std	Median	Skew	MAE (K)/R^2	f_{OCHEM}
1	10 124	410.74	75.01	411.15	−0.14	38.33(0.19)/0.54	0.50
2	16 155	397.91	69.56	394.15	0.24	31.83(0.19)/0.63	0.45
3	14 634	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

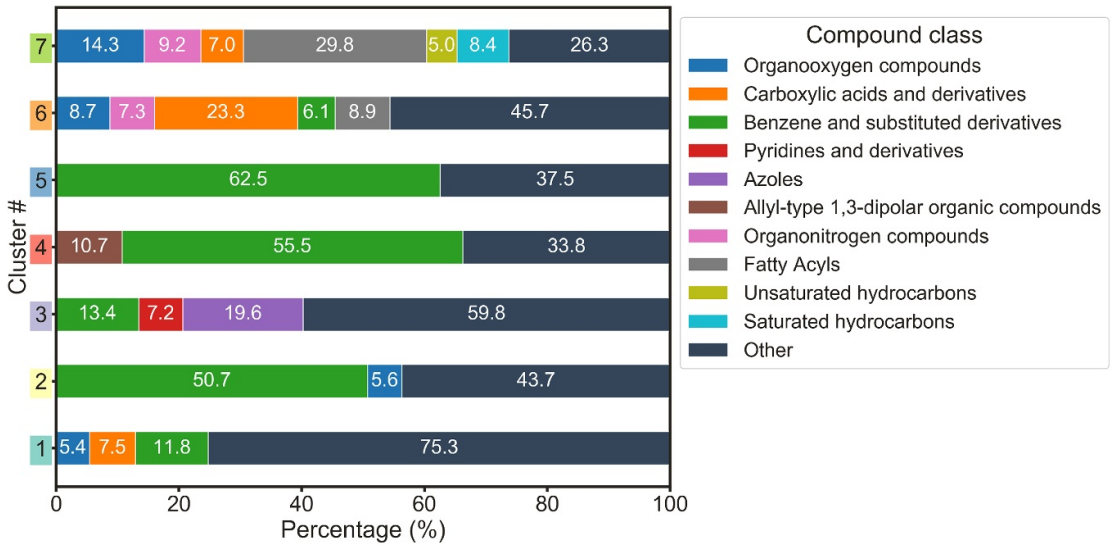

Figure 3. Chemical classification of Murtagh-Ward clusters. Chemical classes were obtained with Classyfire. Only parent chemical classes with at least 5% of the total fraction are shown; parent chemical classes with fractions smaller than 5% were merged in to 'Others'. Cluster indices are the same as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

A further analysis was performed to unravel the parent chemical classes of the compounds reported in each of the clusters using the ClassyFire algorithm. 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 difficulty making predictions on smaller molecules and higher atomic number atoms, where MP is often dominated by electronic phenomena and non-covalent forces.

3.4. Regression on low-bias data sets

To investigate the ability of uniform stratified sampling to create low chemical bias datasets for use with supervised ML, we apply the Butina clustering method to create a new data set ('Butina 0.6'). The MPs corresponding to the 33,408 compound data set (13,974 molecules removed) have been visualized in figure 2 (a). 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

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

outliers at high or low temperatures. The improved performance of the supervised ML on the uniform stratified 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 ~5.2 K MAE with respect to OCHEM. Both the improved performance and the significant reduction in data quantity are a key motivation for the efficacy of the MOLAN workflow, which integrates unsupervised learning techniques as a key element of any ML molecular property analysis task. 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 or less-sophisticated regression methods, uniform stratified sampling of chemical space should be a reliable strategy for modest improvements in data sets where chemical space is not uniformly sampled. For the majority of data sets, particularly those that are experimentally derived, this uniformity of chemical space is not *a priori* anticipated.

3.5. Effect of data set augmentation with 3d descriptors

Following the use of the uniform stratified 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. This is a key point specific to the MP prediction task, as material melting is inherently a 3D, multi-molecule process, and we would like to assess the role of 3D features in improving the accuracy of the prediction task. Our augmented 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) [55], which has row vectors of dimension going as the square of the maximum number of atoms encountered in the dataset, $\max(N_{\text{atom}})^2$. For example in the QM9 dataset, $\max(N_{\text{atom}}) = 29$ [56], which is far smaller than what we encounter in our dataset, $\max(N_{\text{atom}}) = 155$. To overcome this challenge we implemented a dedicated distributed computing work flow based on Apache Spark [57] that can exploit leadership class supercomputers. Apache Spark also comes with a native ML library [58]. Given the high dimensionality of the Coulomb matrix, an initial hyperparameter tuning was performed by means of the spark-sklearn module [59] 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 [60] and gradient-boosted trees [61]. High quality DFT relaxed geometries were used for generation of the CM descriptor. The workflow was benchmarked on the QM9 dataset [56]. By default the workflow utilizes passive sampling (i.e. random split). For the descriptor that gave the best result with the passive sampling, a uniform stratified sampling pipeline was implemented as shown in figure 1.

In figure 4 we examine three manifestations of 3D structure (CM, E3FP, and Morse 3D) derived from quantum-chemical geometries, five quantum-chemical properties (total energy, HOMO-LUMO energy gap, 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 Supporting Information.

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 solvation energy results in a consistent

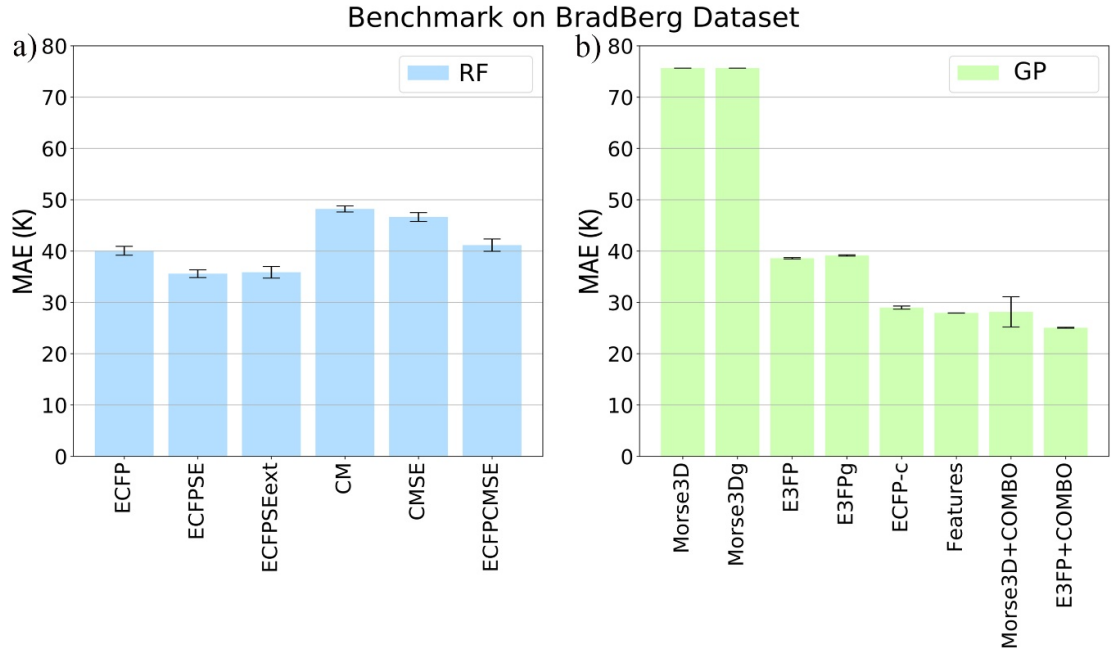

Figure 4. Influence of molecular descriptor choice on BradBerg MP predictions. **(a)** MP predictions using the RF algorithm. **(b)** MP predictions using GPR. Feature set abbreviations described in SI.

improvement in predicted MAE ($\approx$ 4 K for 2D descriptors and $\approx$ 2 K 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(b), 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 Methodology section. 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 ~1 K 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.

3.6. Geometric space construction and graph attribution

Now we turn our attention to discussing the application of the third stage of the MOLAN workflow (figure 1) to the MP data sets. Our intent is to showcase how geometric spaces are shaped by the property of interest (i.e. MP). The SSVAE can be used to generate candidate molecules. The generated molecules will be embedded into the graph embedding space to investigate with more accuracy their performance and possible

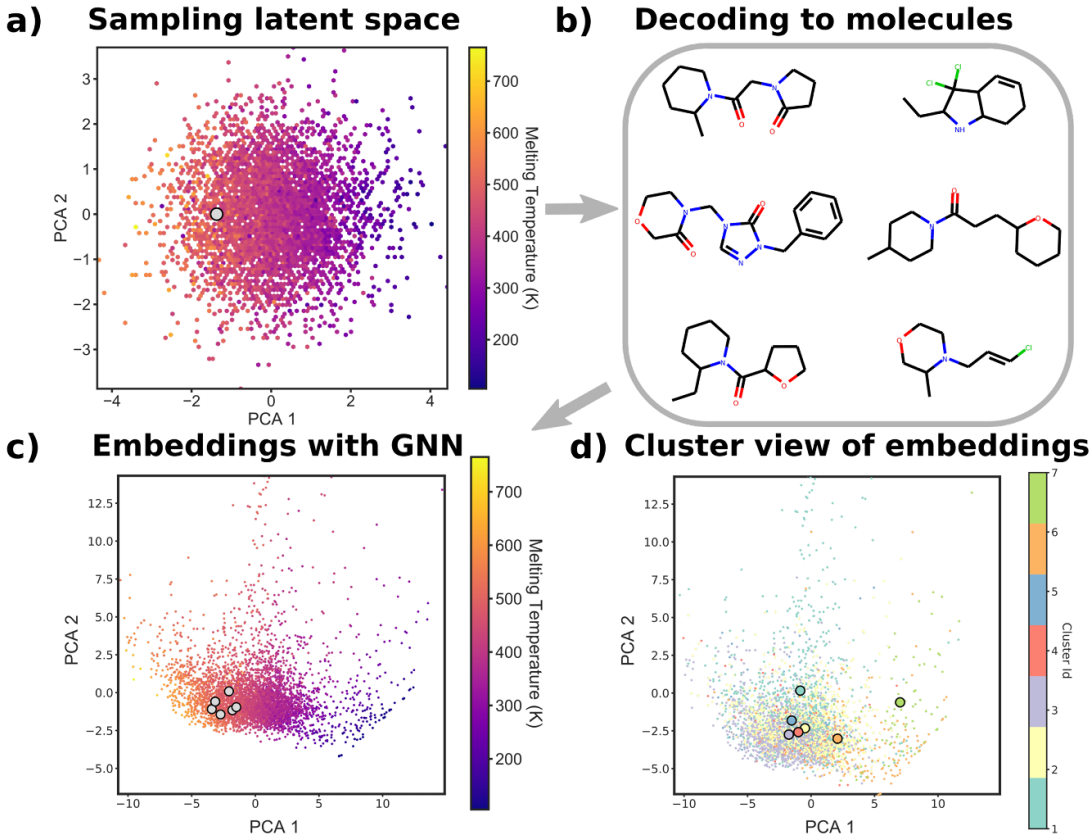

Figure 5. Using geometric spaces of molecules to sample, decode and embed. Each space is constructed from test data sets. Dimensionality reduction is done with PCA. (a) Latent space of SSVAE colored by temperature. Gray point corresponds to the average latent vector of molecules with 350–450 K. (b) Example molecules decoded from the latent vector. (c) Embedding these molecules into the last layer of a GNN. Gray points correspond to newly embedded molecules. (d) Graph embedding space colored by cluster number. Cluster indices are the same as defined in table 2. Points correspond to cluster centroids.

uncertainty (based on distance from test embeddings). Figure 5 displays the usage of geometric spaces derived from a SSVAE (a) and a GNN, specifically a GCN (c,d). The SSVAE is shaped by MP and latent vectors can be decoded to molecular structures which can then be processed by a GNN, to obtain last-layer activations which are a compact MP-optimized representation of a molecule. From this viewpoint the SSVAE represents primarily a generative space biased by MP while the graph embedding space is particularly tuned for prediction. 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 (c) showcase significant temperature gradients, which indicate there are natural tendencies 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 to the data distribution of the training set. The actual direction of this gradient might vary due to architecture and initialization, and so the position with respect to the linear trend is the crucial information. These spaces can be used also to detect out-of-distribution data points [62], data that has irregular activations that do not follow the distribution of the training data are more likely to be outliers or have erroneous estimations of their MP. The relationship between distances in activations of training data and new data with respect to error and uncertainty is a point for future research.

One feature to note in figure 5(a) 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. Meanwhile the graph embedding space does not have this prior and so the actual space varies based on architecture and training. Both spaces present a clear gradient in coloring indicating the space does organize itself from lower to higher temperatures. It should be noted that comparing the actual direction of the gradient between both spaces is not meaningful since small changes in the embeddings or latent vectors can give drastically different results with PCA. One notable result of the graph embedding space is that the notion of chemical clusters derived from unsupervised clustering techniques appear to a certain degree; the cluster centroids are

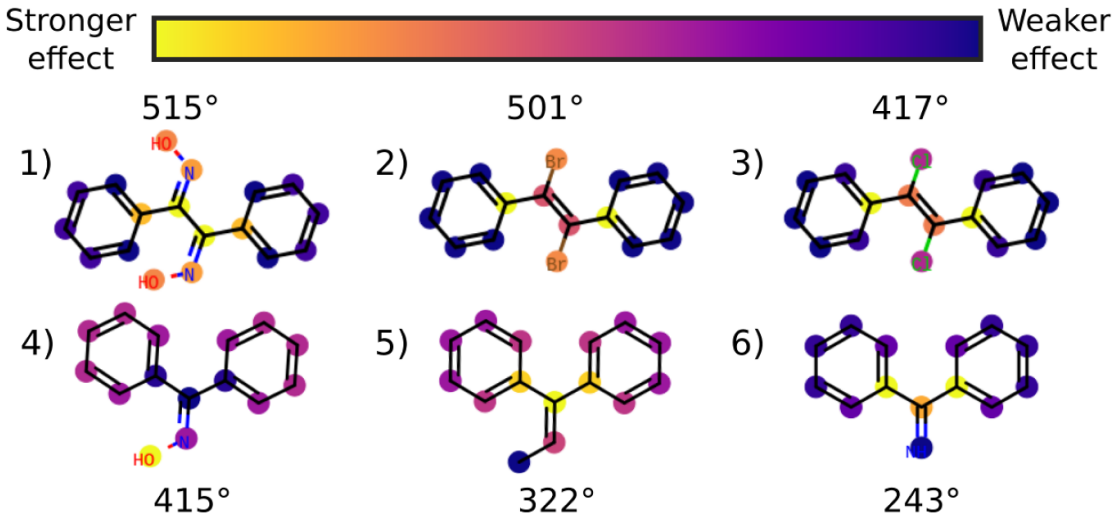

Figure 6. Graph attribution of MP on similar molecules. Colorbar indicates the strength of a node towards a positive temperature. Molecules were chosen to be chemically similar but with a wide array of temperatures.

highlighted and some of them are quite distinct (figure 5(d)). The SSVAE latent space did not exhibit this feature, but if provided during training would likely result.

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 5(c) 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.

Lastly, in figure 6 we apply the graph attribution techniques to visualize molecular heatmaps derived via graph attribution colored by weak and strong contributions to predicted MP. 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. Both OH and halogen pendant groups are consistent with higher MPs due to the potential contributing effects of hydrogen bonding and large polarizabilities, which is also generally consistent with hypotheses from previous works [24]. GradCAM is a local attribution method, it does not look at large-scale trends. The ClassyFire results are one example of semantic analysis on the entire dataset without the use of attributions. 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 online [11] for exploring more in-depth chemical trends. Future work lies in automatically discovering patterns in attributions and relating it to chemical concepts.

4. Conclusions

With the many supervised and unsupervised learning techniques available in the literature, it is often difficult to understand how to best integrate these methods for maximum effect in the context of molecular analysis, design, and discovery. To this end, we have introduced the MOLAN workflow in the context of molecular MP determination, and shown how MOLAN minimizes chemical bias in supervised ML training, provides critical chemical insights into physical property correlations, and provides a pathway for generating new molecular moieties that target specific molecular properties. A prescribed workflow that takes into account the underlying chemical biases present in a data set is 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 prescribed in the MOLAN workflow can help examine how molecular space is organized around MP and 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. Critically, the MOLAN workflow also utilizes some of the highest performing supervised learning techniques, allowing for state-of-the-art predictive accuracy.

With regards to the specific task of MP, 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 applied the MOLAN workflow to shed insight on critical features of MP prediction in the context of molecular materials. First, a literature search was performed to frame our study in the context of experimental and polymorph induced uncertainties in MP determination; we posit that the former should be in the range of ~2–3 K, whereas the average value of polymorph induced uncertainties is in the range of ~11–16 K. With this knowledge, we have applied the MOLAN workflow to construct a low chemical-bias data set augmented 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. Of key interest is the fact that clustering-derived datasets exhibit superior regression accuracy with approximately 70% of the original data quantities. We have also assessed the importance of 3D structural and quantum-chemically derived features in improving MP prediction accuracy, and discovered a modest ~1–2 K 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. Finally, the application of graph attribution techniques identified chemical trends consistent with qualitative concepts of MP molecular structure correlations, and the use of SSVAE provides a mechanism for generatively designing off-the-shelf molecules with targeted MP in the future. In the future, we are excited to see what data sets and applications can be empowered by the prescribed MOLAN workflow for molecular prediction, analysis, and inverse design.

Acknowledgments

This work was supported by the Department of Energy, Basic Energy Sciences, Division of Materials Science and Engineering. N.E.J. thanks the Argonne National Laboratory Maria Goeppert Mayer Fellowship for support. We gratefully acknowledge the computing resources provided by Blues and Bebop, high-performance computing clusters operated by the Laboratory Computing Resources Center at Argonne National Laboratory. G.S. would like to thank Dr. Prasanna Balaprakash for fruitful discussion on unsupervised learning and uncertainty quantification. We thank Prof. Lian Yu for his helpful direction to crystal polymorph databases. G.S. would also like to thank Prof. George K. Thiruvathukal and Dr. Xiao-Yong Jin for help in setting up the Apache Spark workflow at the 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 Laboratory's work was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357. Alan Aspuru-Guzik acknowledges support from the Office of Naval Research under the Vannevar Bush Faculty Fellowship as well as support from the Canada 150 Research Chairs program and Dr Anders G. Froseth.

Data availability statement

The data that support the findings of this study are openly available.

RDKit descriptors

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_Al_OH, fr_Al_OH_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_HOCCN, 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_piperdine, fr_piperzine, fr_priamide, fr_prisulfonamd, fr_pyridine, fr_quatN, fr_sulfide, fr_sulfonamd, fr_sulfone, fr_term_acetylene, fr_tetrazole, fr_thiazole, fr_thiocyan, fr_thiophene, fr_unbrch_alkane, fr_urea, MolWt, MolLogP

Limitations to prediction of melting point

With the large number of diverse data sources contributing to a likely non-uniform experimental MP dataset structure, 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. As a short summary, using 119 literature-derived small molecules and their associated ranges in MP, we conclude that 80% of crystal polymorphs exhibit MP ranges of less than 20 K. When combined with an estimated 1–5 K error derived from the experimental measurements themselves, we argue that this aggregate limitation to prediction error is significantly less than the 35–50 K errors in prediction accuracy obtained in previous MP prediction works.

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 20 K [63]. To more quantitatively approach the issue, we have examined the MPs of 119 unique crystal polymorphs gathered from the experimental literature [64, 65]. In this data set, we have computed the size of the MP interval in K for all polymorphs for each molecular structure (ΔT_m), and histogrammed the distribution of MP intervals in figure 2(b). 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 ~11–16 K, which we tentatively use as a lower bound for our expected error due to polymorphs. Note, that the previously obtained intervals of 35–50 K 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 [66] and MPA (https://www.thinksrs.com/products/mpa100.html) measurements should exhibit reproducibility of ~0.1 K. 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 [67]. However, even with these considerations, if properly performed, the majority of pure organic materials typically exhibit melting ranges of 1–2 K 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 K min^{-1} and 20 K min^{-1}, with most high precision studies occurring at heating rates less than 1 K min^{-1}. Despite the majority of DSC peak widths for small organic molecules being 1−2 K, 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 K when the heating rate is increased from 0.6 K min^{-1} to 20 K min^{-1} [68]. 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 K, with a polymer-derived maximum bound of roughly 6 K. 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 [16] 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 [16] 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 K. 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. In this work, we utilize the versions of these datasets provided directly by these authors at http://ochem.eu/article/55638, which incorporate the filtering of these outliers. Critical to point out in this regard, however, is that the aggregate model predictions compared between the data set with, and without, outliers results in changes of ~1 K MAE, as shown in their work. In this work, we find that changes inclusion/exclusion of ~6 % of our data sets results in negligible changes in the quality of our predictive models, which is likely attributable to our employed unsupervised clustering filtering scheme.

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 [52]. RDKit [28] 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** geometry optimizations in Gaussian [69]. 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 in water (Gaussian16—SCRF = (SMD,DoVacuum), all other relevant parameters set to defaults) [70], 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 [11].

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 [71]. 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 Morgan-like 3D 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.

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

ORCID iDs

Ganesh Sivaraman https://orcid.org/0000-0001-9056-9855
Nicholas E Jackson https://orcid.org/0000-0002-1470-1903
Alán Aspuru-Guzik https://orcid.org/0000-0002-8277-4434

References

[1] Cherkasov A *et al* 2014 QSAR modeling: where have you been? where are you going to? *J. Med. Chem.* **57** 4977–5010
[2] Varnek A, Kireeva N, Tetko I V, Baskin I I and Solov'ev V P 2007 Exhaustive QSPR studies of a large diverse set of ionic liquids: how accurately can we predict melting points? *J. Chem. Inf. Model.* **47** 1111–22
[3] Szymański P, Markowicz M and Mikiciuk-Olasik E 2011 Adaptation of high–throughput screening in drug discovery–toxicological screening tests *Int. J. Mol. Sci.* **13** 427–52
[4] Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera R S, Gold-Parker A, Vogt L, Brockway A M and Aspuru-Guzik A 2011 The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid *J. Phys. Chem. Lett.* **2** 2241–51
[5] Yan Q *et al* 2017 Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment *Proc. Natl Acad. Sci.* **114** 3040–3
[6] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 Machine learning and the physical sciences *Rev. Mod. Phys.* **91** 045002
[7] Dimitrov T, Kreisbeck C, Becker J S, Aspuru-Guzik A and Saikin S K 2019 Autonomous molecular design: then and now *ACS Appl. Mater. Inter.* **11** 24825–36
[8] Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural message passing for quantum chemistry *Proc. of the 34th Int. Conf. on Machine Learning* vol 70 pp 1263–72
[9] Sanchez-Lengeling B and Aspuru-Guzik A 2018 Inverse molecular design using machine learning: generative models for matter engineering *Science* **361** 360–5
[10] Lemmer M, Inkpen M S, Kornysheva K, Long N J and Albrecht T 2016 Unsupervised vector-based classification of single-molecule charge transport data *Nat. Commun.* **7** 12922
[11] MOLAN A 2020 Machine Learning Workflow for Molecular Analysis: Application to Melting Points https://github.com/argonne-lcf/molan
[12] Ran Y, Jain N and Yalkowsky S H 2001 Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE) *J. Chem. Inf. Comput. Sci* **41** 1208–17
[13] Tetko I V 2008 Associative neural network *Methods MOl. Biol.* (Totowa, NJ: Humana Press) vol 458 pp 185–202
[14] Preiss U P, Beichel W, Erle A M T, Paulechka Y U and Is Universal K I 2011 Simple melting point prediction possible? *Chem. Phys. Chem* **12** 2959–72
[15] Nikmo J, Kukkonen J and Riikonen K 2002 A model for evaluating physico-chemical substance properties required by consequence analysis models *J. Hazard. Mater.* **91** 43–61
[16] Tetko I V, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko A E, Charochkina L and Asiri A M 2014 How accurately can we predict the melting points of drug-like compounds? *J. Chem. Inf. Model.* **54** 3320–9
[17] Tetko I V, Lowe D M and Williams A J 2016 The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from patents *J. Cheminf.* **8** 2
[18] Withnall M, Chen H and Tetko I V 2018 Matched molecular pair analysis on large melting point datasets: a big data perspective *ChemMedChem* **13** 599–606
[19] Bhhatarai B *et al* 2011 CADASTER QSPR models for predictions of melting and boiling points of perfluorinated chemicals *Mol. Inf.* **30** 189–204
[20] Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E and Mitchell J B 2006 Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization *J. Chem. Inf. Model.* **46** 2412–22
[21] Krstajic D, Buturovic L J, Leahy D E and Thomas S 2014 Cross-validation pitfalls when selecting and assessing regression and classification models *J. Cheminf.* **6** 10
[22] Sahlin U, Jeliazkova N and Oberg T 2014 Applicability domain of dependent predictive uncertainty in QSAR regressions *Mol. Inf.* **33** 26–35
[23] Karthikeyan M, Glen R C and Bender A 2005 General melting point prediction based on a diverse compound data set and artificial neural networks *J. Chem. Inf. Model.* **45** 581–90
[24] Bergstrom C A, Norinder U, Luthman K and Artursson P 2003 Molecular descriptors influencing melting point and their role in classification of solid drugs *J. Chem. Inf. Comput. Sci.* **43** 1177–85
[25] Zang Q, Mansouri K, Williams A J, Judson R S, Allen D G, Casey W M and Kleinstreuer N C 2017 In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning *J. Chem. Inf. Model.* **57** 36–49
[26] Brown T N, Armitage J M and Arnot J A 2019 Application of an iterative fragment selection (IFS) method to estimate entropies of fusion and melting points of organic chemicals *Molecular informatics* **38** 1800160
[27] eMolecules plus database download 2019 https://www.emolecules.com/info/plus/download-database (accessed Jun 26 2019)
[28] Landrum G 2018 RDKit: Open-source cheminformatics www.rdkit.org (accessed Jan 15 2020)
[29] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[30] Murtagh F, Contreras P 2011 Methods of hierarchical clustering https://arxiv.org/abs/1105.0121 (accessed Sep 4 2019)

[31] Butina D 1999 Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets *J. Chem. Inf. Comput. Sci.* **39** 747–50
[32] Rogers D and Extended-Connectivity Fingerprints H M 2010 *J. Chem. Inf. Model.* **50** 742–54
[33] Hastie T, Tibshirani R and Friedman J 2001 *The Elements of Statistical Learning* Springer Series in Statistics (New York: Springer)
[34] Feunang Y D *et al* 2016 ClassyFire: automated chemical classification with a comprehensive, computable taxonomy *J. Cheminf.* **8** 61
[35] Dasgupta S and Hsu D 2008 Hierarchical sampling for active learning *Proc. of the 25th International Conference on Machine Learning* pp 208–15
[36] Dasgupta S 2011 Two faces of active learning *Theor. Comput. Sci.* **412** 1767–1781
[37] Duvenaud D K, Maclaurin D, Iparraguirre J, Gómez-Bombarell R, Hirzel T, Aspuru-Guzik A and Adams R P 2015 Convolutional networks on graphs for learning molecular fingerprints *Adv. Neural Inf. Process. Systems* 2215–23
[38] De A G, Matthews G, Van Der Wilk M, Nickson T, Fujii K, Boukouvalas A, León-Villagrá P, Ghahramani Z and Hensman J 2017 GPflow: A Gaussian process library using tensorflow *J. Machine Learn. Res.: JMLR* **18** 1–6
[39] Axen S D, Huang X-P, Cáceres E L, Gendelev L, Roth B L and Keiser M J 2017 A simple representation of three-dimensional molecular structure *J. Med. Chem.* **60** 7393–409
[40] Todeschini R, Consonni V 2008 *Handbook of Chemoinformatics* (New York: Wiley) pp 1004–33
[41] McCloskey K, Taly A, Monti F, Brenner M P and Colwell L 2018 Using attribution to decode dataset bias in neural network models for chemistry https://arxiv.org/abs/1811.11310 (accessed Sep 4 2019)
[42] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization *The IEEE Int. Conf. on Computer Vision (ICCV)* pp 618–26
[43] Preuer K, Klambauer G, Rippmann F, Hochreiter S and Unterthiner T 2019 Interpretable deep learning in drug discovery http://arxiv.org/abs/1903.02788 (accessed Sep 4 2019)
[44] Gómez-Bombarelli R *et al* 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Central Sci.* **4** 268–76
[45] Irwin J J and Shoichet B K ZINC 2005 A free database of commercially available compounds for virtual screening *J. Chem. Inf. Model.* **45** 177–82
[46] Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A 2019 SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry https://arxiv.org/abs/1905.13741 (accessed Sep 4 2019)
[47] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V and Veselov M 2018 Others, molecular sets (moses): a benchmarking platform for molecular generation models https://arxiv.org/abs/1811.12823 (accessed Sep 4 2019)
[48] Fu H, Li C, Liu X, Gao J, Celikyilmaz A and Carin L 2019 Cyclical annealing schedule: a simple approach to mitigating kl vanishing https://arxiv.org/abs/1903.10145
[49] GPyOpt : A Bayesian optimization framework in python 2016 http://github.com/SheffieldML/GPyOpt (accessed Dec 10 2018)
[50] Bradley J-C, Lang A, Williams A 2014 Jean-Claude bradley double plus good (highly curated and validated) melting point dataset (https://figshare.com/articles/Jean_Claude_Bradley_Double_Plus_Good_Highly_Curated_and_Validated_Melting_Point_Dataset/1031638/1) (accessed Sep 20 2017)
[51] ENAMINE Ltd (www.enamine.net) (accessed Sep 20 2017)
[52] OCHEM (www.ochem.eu) (accessed Sep 20 2017)
[53] Coley C W, Barzilay R, Green W H, Jaakkola T S and Jensen K F 2017 Convolutional embedding of attributed molecular graphs for physical property prediction *J. Chem. Inf. Model.* **57** 1757–72
[54] Wu Z, Ramsundar B, Feinberg E N, Gomes J, Geniesse C, Pappu A S, Leswing K and Pande V 2018 MoleculeNet: a benchmark for molecular machine learning *Chem. Sci.* **9** 513–30
[55] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[56] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 Quantum chemistry structures and properties of 134 kilo molecules *Sci. Data* **1** 140022
[57] Zaharia M, Xin R S, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J and Venkataraman S *et al* 2016 Franklin M J Apache spark: a unified engine for big data processing *Commun. ACM* **59** 56–65
[58] Meng X, Bradley J, Yavuz B, Sparks E, Venkataraman S, Liu D, Freeman J, Tsai D and Amde M *et al* 2016 Owen S. Mllib: Machine learning in Apache Spark *J. Machine Learn. Res.* **17** 1235–41
[59] Scikit-learn integration package for Apache Spark https://github.com/databricks/spark-sklearn (accessed Mar 5 2018)
[60] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
[61] Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* 1189–232
[62] Speakman S, Sridharan S, Remy S, Weldemariam K, McFowland E 2018 Subset scanning over neural network activations
[63] Wille R L and Lutton E S 1966 Polymorphism of cocoa butter *J. Am. Oil Chem. Soc.* **43** 491–6
[64] Burger A and Ramberger R 1979 On the polymorphism of pharmaceuticals and other molecular crystals. II *Microchim. Acta* **72** 273–316
[65] Yu L 1995 Inferring thermodynamic stability relationship of polymorphs from melting data *J. Pharm. Sci.* **84** 966–74
[66] FGill P, Moghadam T T and Ranjbar B 2010 Differential scanning calorimetry techniques: applications in biology and nanoscience *J. Biomol. Tech.* **4** 167–93
[67] Vyazovkin S, Chrissafis K, Lorenz M L D, Koha N, Pijolat M, Roduit B, Sbirrazzuoli N and Sunol J J 2014 ICTAC kinetics committee recommendations for collecting experimental thermal analysis data for kinetic computations *Thermochim. Acta.* **590** 1–23
[68] Hellmuth E and Wunderlich B 1965 Superheating of linear high-polymer polyethylene crystals *J. App. Phys.* **36** 3039
[69] Frisch M J *et al* 2016 Gaussian16 Revision C.01. Gaussian Inc. Wallingford CT
[70] Marenich A V, Cramer C J and Truhlar D G 2009 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 **113** 6378–96
[71] Ebejer J-P, Morris G M and Deane C M 2012 Freely available conformer generation methods: how good are they? *J. Chem. Inf. Model.* **52** 1146–58