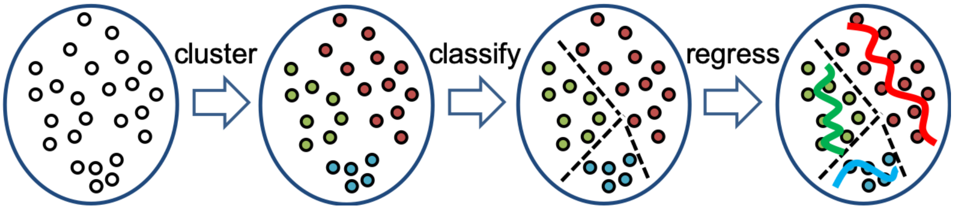
Thinking Globally, Acting Locally: On the Issue of Training Set Imbalance and the Case for Local Machine Learning Models in Chemistry

Mojtaba Haghighatlari,[1, *] Ching-Yen Shih,[1] and Johannes Hachmann[1, 2, 3, †]

[1] *Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States*
[2] *Computational and Data-Enabled Science and Engineering Graduate Program, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States*
[3] *New York State Center of Excellence in Materials Informatics, Buffalo, NY 14203, United States*

The appropriate sampling of training data out of a potentially imbalanced data set is of critical importance for the development of robust and accurate machine learning models. A challenge that underpins this task is the partitioning of the data into groups of similar instances, and the analysis of the group populations. In molecular data sets, different groups of molecules may be hard to identify. However, if the distribution of a given data set is ignored then some of these groups may remain under-represented and the sampling biased, even if the size of data is large. In this study, we use the example of the Harvard Clean Energy Project (CEP) data set to assess the challenges posed by imbalanced data and the impact that accounting for different groups during the selection of training data has on the quality of the resulting machine learning models. We employ a partitioning criterion based on the underlying rules for the CEP molecular library generation to identify groups of structurally similar compounds. First, we evaluate the performance of regression models that are trained globally (i.e., by randomly sampling the entire data set for training data). This traditional approach serves as the benchmark reference. We compare its results with those of models that are trained locally, i.e., within each of the identified molecular domains. We demonstrate that local models outperform the best reported global models by considerable margins and are more efficient in their training data needs. We propose a strategy to redesign training sets for the development of improved global models. While the resulting uniform training sets can successfully yield robust global models, we identify the distribution mismatch between feature representations of different molecular domains as a critical limitation for any further improvement. We take advantage of the discovered distribution shift and propose an ensemble of classification and regression models to achieve a generalized and reliable model that outperforms the state-of-the-art model, trained on the CEP data set. Moreover, this study provides a benchmark for the development of future methodologies concerned with imbalanced chemical data.



I. INTRODUCTION

Machine learning (ML) is in the process of revolutionizing several aspects of chemical (and materials) research. ML approaches illuminate underlying patterns in chemical data, they facilitate efficient predictions in the characterization and behavior of chemical systems, and they augment conventional processes of decision making in chemical research [1, 2]. One application of ML is the creation of data-derived surrogate models that accelerate the intensive process of molecular discovery, design, and development by orders of magnitude [3, 4]. The application of ML on the results of virtual high-throughput screening (HTPS) studies has been one of the earliest and most successful approaches for the large-scale exploration of molecular space [5–8].

A majority of methodological advancements for ML in the chemical domain have so far focused on improving the performance of data-derived prediction models for desirable materials properties. Research on other pertinent questions of chemical data mining and modeling has received less attention. Examples of these issues are: (i) diversity/sparsity of the molecular structures, (ii) applicability domains of trained ML models, and (iii) learning from imbalanced data. While these issues require expertise from domain sciences [9], they are ultimately interconnected and share common solutions across disciplines. In this work, we address these challenges on a well-known molecular data set.

A data set is imbalanced if it can be partitioned into groups of similar instances (e.g., molecules), but the count of instances per group differs significantly [10]. By that means, even if the data set is large, some groups

[*] mojtabah@buffalo.edu
[†] hachmann@buffalo.edu

may remain under-represented (i.e., minority groups). Typically, ML approaches fail to capture the characteristics of the minority groups because they are less exposed to the instances they contain. In the case where little is known about the unique or even rigorous criteria to discover the distribution of an unknown molecular library, chemical intuition or a direct mapping between structural features and groups (e.g., *via* clustering, i.e., an unsupervised ML approach) can mitigate this limitation.

The current trend in chemical and materials studies is around developing models that are as general as possible, thus they are transferable to a broader molecular space [11]. However, there are also studies that argue against the pursuit of universal prediction models. For instance, a study by Goldsmith *et al.* proposed the method of subgroup discovery to cluster the crystal structures of semiconductors [12]. The results of their work show that local models – trained on subgroups of the entire compound data – can significantly outperform a global model that is trained on the natural distribution of the overall data set. More recently, Kailkhura *et al.* introduced an ML framework that successfully addresses the skewed distribution of materials by partitioning the range of target properties [13]. While this approach has similar objectives to our work, we discuss the imbalance in data as a consequence of the sparse structural features than target properties. Thereby, it needs to be addressed as an unsupervised ML approach.

To exemplify, visualize, and address the challenge of imbalanced data in a molecular system, we focus on the Harvard Clean Energy Project (CEP) data set [14]. CEP is one of the prominent HTPS efforts to find organic photovoltaics for their application in solar cells. This data set has been widely used in several modeling and method development projects [15–18], mainly to recover the rigorous and deterministic quantum chemical mapping from the structure/topology of a molecule to its properties. However, the resulting ML models are only trained on the natural distribution of the data, and ignore the sparsity of the data set. Taking advantage of the imposed constraints in the combinatorial exploration of molecular space, we investigate the imbalance in the CEP data set. We show that the results of our pattern recognition efforts lead to reliable prediction across the range of data, and thus enhance the applicability domain of trained models. In a broader perspective, the applied approaches attempt to emphasize the issue, and establish efficient practices to develop generalized predictive models on organic molecular data sets.

II. BACKGROUND, METHODS, AND COMPUTATIONAL DETAILS

A. Motivation

As described by Hachmann *et al.* [19], the CEP molecular library is generated using 26 distinct building blocks that react with each other and create new fragments based on two types of reactions (i.e., linking or fusion). The maximum number of building blocks per molecule is limited to five. Although all the 26 building blocks are prevalent substructures for the photovoltaic applications, the type of reaction between them results in very different molecular moieties that might not be feasible to synthesis. The motivation for this work is based on the feedback from experimental collaborators regarding the ease of synthesis for molecules that are only a product of linking reactions. The synthesis condition for fused fragments are often harsh and may negatively result in a ring expansion or contraction [20].

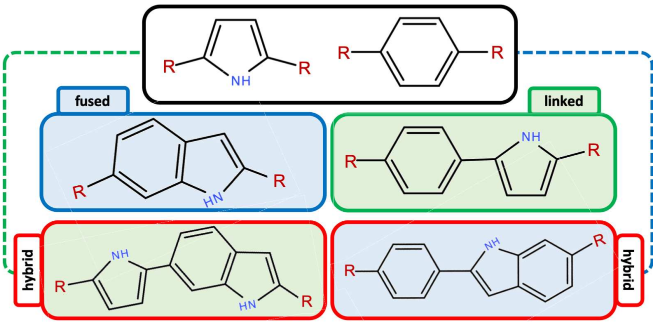


FIG. 1: Examples of fused, linked, and hybrid molecules initiated from two building blocks (i.e., benzene and pyrrole) with two reaction sites ('R's represent chemical handles) per each. The product of reaction also has two chemical handles to participate in the next reactions and create larger molecules. A hybrid molecule is a product of linking a building block to a fused molecule, or *vice versa.*

Thus, the initial goal of this study is to interpret the generation of each molecule in the CEP library based on the combination rules and types of building blocks. Note that the reaction scheme has not been captured initially along with the original CEP library generation. One immediate solution to this problem is to search for all possible combinations of building blocks in the molecules. However, this is an intractable approach due to the large number of possible combinations. Therefore, we develop an algorithm to divide CEP molecules to three groups of fused, linked, or hybrid. As it is illustrated in Fig. 1, a hybrid molecule is a result of both linking and fusion reaction between constituent building blocks. The details of the algorithm will be discussed in Sec. II B.

After extracting the substructure information, our motivation for the rest of the study is two-fold. First, we investigate the distribution of top candidates in the subgroups of the molecules. We train ML regression models on each subgroup separately (i.e., develop local models) and compare their performance with the model that is trained on a random sample of the entire CEP data set (i.e., a global model). Second, we utilize the extra information regarding the distribution of clusters in the data set to improve the performance of the global model. For this purpose, we oversample under-represented subgroups and create a uniform distribution of clusters in the training set. We observe that the main challenge for any further improvements in the global model is the shift in the distribution of the feature representation for each class of molecules. Therefore, we apply classification and feature transformation methods to demonstrate the impact of the distribution mismatch in the training sets. This approach leads us to an ensemble of classification and regression models to avoid bias towards majority subgroups. Thus, the central claim of this study is to provide the most accurate and generalized ML model for predicting photovoltaic properties of the molecules in the CEP data set. In summary, the contribution of this paper is as follows:

- We propose an algorithm to exploit the structure of the CEP data set. This approach is based on the reaction scheme that molecules have undertaken during the library generation.

- When the reaction scheme is identified, we partition the entire CEP data set based on the synthesis feasibility, thereby, we achieve benchmark data with identified subgroups to assess the underrepresentation of similar molecular structures.

- We next investigate the effect of the imbalanced classes on the performance of ML models that are developed using pure random sampling of the entire data set. Subsequently, we achieve computationally efficient ML models that outperform the state-of-the-art predictive models for the CEP data set.

- Finally, we automatize the entire approach by redesigning the ML training sets and training a classification model to alleviate the problem of distribution shift for regression models.

B. CEP Data Set and Molecular Characterization

The CEP data set contains more than 2.3 million organic photovoltaic molecules that are candidates for donor materials in solar cells. The target property in this data set is the power conversion efficiency (PCE) that is a measure of the performance of solar cells. PCE values are approximated using Scharber model [21] and electronic properties of donor molecules (i.e., molecules in the CEP data set). The electronic properties are calculated at BP86/def2-SVP [22–24] level of the Kohn-Sham density functional theory [25, 26]. All the molecules are represented using SMILES strings, which provide 2D information of the molecular structures, that is, atom type and connectivity.

We propose an algorithm based on the molecular graphs to characterize the unique combination scheme that has been undertaken for the generation of each molecule in the CEP library. The algorithm is specific to the CEP data set and takes advantage of the heterocyclic structure of the building blocks. In the following we describe the overall pseudo-algorithm:

1. represent each molecule as a graph of nodes and edges and keep track of their corresponding chemical labels (i.e., atom and bond types).

2. identify all the cycles in the molecular graphs with size less than 6. They correspond to the 5- and 6-membered molecular rings in the structure of building blocks.

3. look for nodes that are shared between the rings. These types of nodes represent the fusion reaction between rings.

4. look for edges that are not involved in any of the rings. These types of edges form the linking connection between rings.

5. discover type of building blocks and their connections based on atom and bond types and the unique combination of rings in building blocks.

The second step of the algorithm requires an efficient code for finding exact length of paths between two atoms in a closed loop. The available algorithms for this step, e.g., Johnson's algorithm [27], are computationally expensive. We use the built-in function available in the OpenBable [28], which can efficiently perform the first two steps of the algorithm. The final result of this substructure analysis represents the exact type, order, and symmetry of building blocks based on the position of reaction sites in each molecule. The obtained substructure information could be further utilized to focus on a group of molecular moieties that may be of higher importance for the rational design of materials.

Note that we use the term cluster to distinguish fused, linked, and hybrid molecules in the CEP data set. The choice of this terminology is not entirely arbitrary; it tends to differentiate this task from the supervised classification approach. More importantly, it has an analogy with unsupervised clustering approach that will be discussed in our following publications as an automated and generalized fashion to distinguish organic molecules. Thus in this study, we undertake a clustering approach based on the chemical intuitions from and only for CEP data set.

C. Machine Learning Details

The ML task in this study is a supervised learning type due to the availability of the labeled data. A supervised learning approach can be considered as a function that maps the input features to the target outputs [29]. If the output labels are numerical (and continuous), the supervised learning task is a regression problem, and if the labels are categorical, the problem is classification. In this study, we train a regression model to predict the scalar PCE values. We also train a classification model to label each molecule as a fused, linked, or hybrid type. The classification problem is carried out to evaluate the distribution shift in the input features of the three clusters. We also apply principal component analysis to reduce the dimension of the feature space for the visualization of the distribution shift.

We use deep neural networks (DNN) for both of the regression and classification tasks [30, 31]. The DNN consists of neurons that behave as a simple feature transformation unit. Originally, each neuron sends out the result of an activation function acting on the total sum of the weighted inputs that receives from all connected neurons. The neurons are organized in consecutive layers and may be partially connected. Thus, the entire DNN model is able to transform the input features to the latent space, where the mapping to the target output becomes linear. The only difference between our classification and regression models is the choice of activation function for the last layer, which is the Softmax function for classification instead of a linear function for regression. We train a fully-connected standard architecture of DNN with three hidden layers in this work. We optimize other hyper-parameters (e.g., activation functions, regularization parameter, learning rate, etc.) using the 10-fold cross-validation approach on the 90% of the data as the training set. The remaining 10% are held out for the final evaluation of the model. We use our implementation of genetic algorithm to efficiently search and optimize the hyper-parameter space [32]. In addition to the cross-validation, we carry out two additional approaches to avoid over-fitting: (1) the regularization term to penalize the parameters that are biased to the noise [33], and (2) the early stopping approach, which stops training iterations when the model improvement is negligible. Both of these methods avoid unnecessary model complexity.

We perform this ML workflow using *ChemML* [34, 35], our program package for machine learning and informatics in chemical and materials research. In this work, *ChemML* employs the Keras library [36] with Tensorflow as backend [37] to develop the DNN models. The scikit-learn library provides tools for data preprocessing and model evaluation [38]. To plot learning curves, we select five different subset sizes spaced uniformly over the range of the training set size. We next repeat training and evaluation five times for each of the subset sizes and plot the average performance. The main evaluation metric for regression models is the mean absolute error (MAE). MAE represents the absolute deviation of the predicted values from the target properties. For classification models, we use the ratio of correct predictions to evaluate the accuracy of our classifiers.

D. Feature Representation

In the cheminformatics and materials informatics, the input features to an ML model are called descriptors [39]. Descriptors provide a numerical representation of the molecules and are the most important aspect of the ML models. Recent studies emphasize that substructure-based descriptors provide essential representation to predict several properties of molecules [40–42]. This type of descriptors, which are also known as the molecular fingerprint (FP), indicate the presence or absence of particular substructures in the molecule. In this study, we use 2048-bit, radius 3 Morgan FP [43, 44], from RDKit cheminformatics library [45], as previously recommended in similar studies (concerning the structure and property of molecules) [15].

In addition to Morgan FP, we employ neural fingerprint (NFP) as the state-of-the-art neural network architecture, i.e., graph convolutional networks, originally developed and tested on the CEP data set [16]. Previous research has established that NFP provides a more comprehensive representation of the structural makeups than Morgan FP. The NFP takes advantage of the embedded atom features in a molecular graph. It can be considered as a stand-alone ML model by addition of fully connected layers. Therefore, for the purpose of comparison, we add the same standard neural network that was described in Sec. II C to complete the NFP model. Various versions of similar deep learning architectures have been developed recently [18, 46, 47]. Given the better computational scaling of the NFP, it is a sufficiently complex and accurate model to serve as the state-of-the-art technique in this study. However, these models are computationally more demanding than the standard DNN by two orders of magnitude in terms of minute calculation on the same computation resource. We ultimately consider this point in all aspects of our conclusions.

III. RESULTS AND DISCUSSION

A. Statistical Analysis of the Clusters

The probability distribution of PCE values in the CEP data set, and separately in each of the clusters is presented in Fig. 2. The heavier tails of the distribution for linked molecules praise the idea that their structural makeup is more favorable for the photovoltaic applications. A closer look to the molecules with PCE>8% (the inset violin plot) reveals the prevalence of linked molecules among top candidates, specificaly those with PCE>10%. This is an important finding, which strengthens our initial motivation for the study of molecules based on their synthetic feasibility. Thereby, a focus on the linked molecules not only makes the synthesis easier but potentially leads towards more important candidates in photovoltaic materials. Note that the negative PCE is the artifact of the Scharber model and does not present any physical meaning. However, we keep the entire range of the PCE values for the purpose of the training and do not change them to zero. This way we preserve the continuity of the values, which allows us to learn the actual Scharber model.

Table. I summarizes the statistical analysis of the three discovered clusters in the CEP data set.The analysis of positive PCE values among each cluster clearly shows that linked molecules with average of 3.24% have the highest PCE compared to the other two clusters. In regard to the population of each group, we find that the three clusters are not uniformly distributed. The fused and linked molecules only occupy 3% and 23% of the entire data set, respectively. Based on the ML literature, these two clusters are known to be under-represented in comparison with the hybrid cluster with 74% of the total data [48]. Therefore, a global ML model trained on a random sample of the CEP data should be mostly biased to the hybrid molecules. The other point is that distinguishing the three clusters from each other based on their PCE is not possible since their distributions approximately overlap (see Fig. 2 again). This point emphasizes that sometimes molecular structures are indistinguishable based on their target properties and thus, other criteria are required to partition them.

Since our goal for the rest of the study is to compare the performance of three local models with respect to the global model's accuracy, we randomly select 50k data from each cluster and the entire data set, making four samples for training, validation, and final evaluation as described in the methods section. The resulting four samples are referred to as fused, linked, hybrid and random in the rest of the study. The 50k sample size is motivated by the size cap of the smallest cluster (i.e., fused molecules), and the size of the training data in the recent ML studies on the CEP data set.

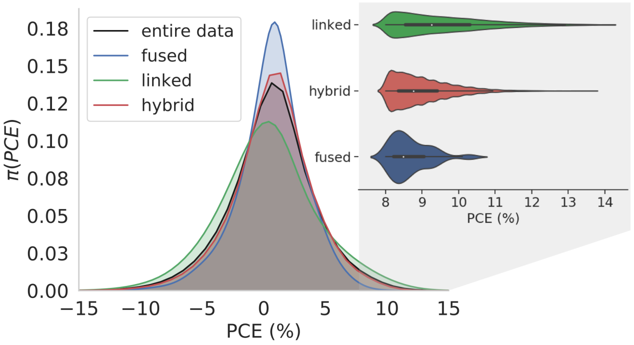


FIG. 2: The histogram showing the distribution of Scharber power conversion efficiency (PCE) values in the entire and each cluster of the CEP data set. The inset violin plot presents the mean, standard deviation and 25/75th percentile over candidates with PCE>8.0%.

TABLE I: Statistical analysis for each cluster in the CEP data set. We obtain the population of each cluster in the entire set of 2.3 million molecules and in a subset of the data with positive PCE. We then compute the average and mean absolute deviation (MAD) for each subset. The population is in million and the avg and MAD of PCE are in %.

	all	$PCE > 0$		
	population(m)	population(m)	avg(%)	MAD(%)
fused	0.07 (3%)	0.04	2.18	1.34
linked	0.54 (23%)	0.27	3.24	2.09
hybrid	1.72 (74%)	1.07	2.61	1.60
all	2.33	1.38	2.72	1.70

B. Predicting PCE of Organic Solar Cells

Following the statistical analysis of the clusters, now we have four samples to serve as our data sets for the training and testing of our models. Thus, each data set provides one training and one test set. We train standard DNN regression models with Morgain FP and NFP as their input on each of the training sets and evaluate the model on all the four test sets. The model that is trained on the random data set serves as our global model because it can potentially predict the PCE for molecules from any of the clusters. Fig. 3.a shows the distribution of absolute errors of each model, evaluated on the same four out-of-sample test sets. We find that each local model performs better on the test set from the same category of training data. The highest errors belong to the fused and linked models when they are evaluated on test sets from each other's test sets. The poor performance of the local models on other clusters, although they share same building blocks, is highly surprising. All three local models outperform the random model for prediction on their own type of molecules. The closest performance to the random model belongs to the hybrid model since hybrid molecules are over-represented in the random sample. For a similar reason, the performance of the random model on the linked and fused test sets are approximately 2 and 5 times worse than their local models.

In regard to the structural similarities, we find that the fused model performs better on the hybrid test set compared to the linked model. This point can also be confirmed based on the comparison between the performance of the hybrid model on the fused and linked test sets. These results determine that hybrid molecules have more in common with fused molecules rather than linked, which is expected based on the reaction scheme that is used for the generation of CEP data [49]. Although more than 70% of the random sample consists of hybrid molecules, the random model still performs slightly worse than the model that is trained on 100% hybrid molecules (i.e., hybrid model). This point clearly highlights the impact of the imbalance in the data on the training of the ML models.

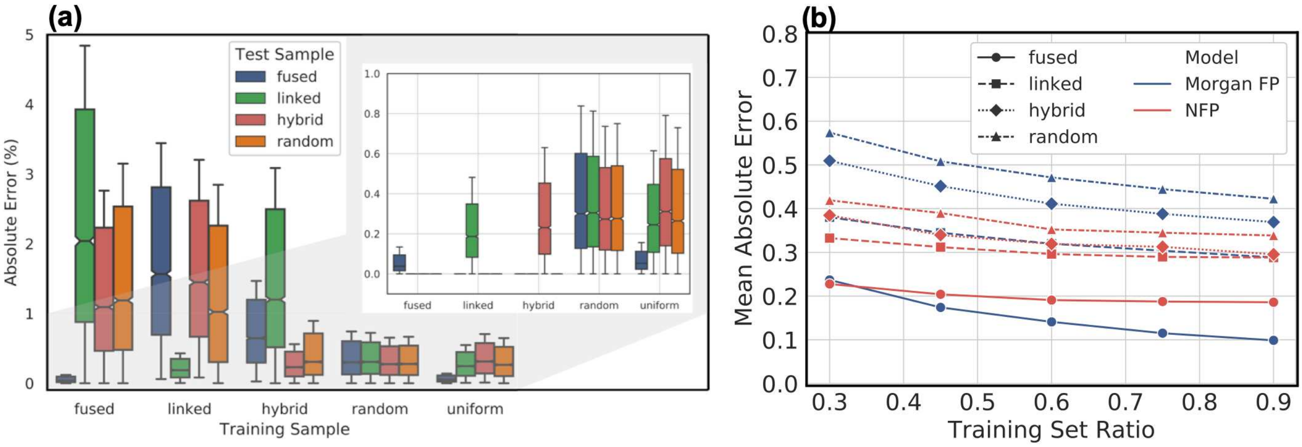


FIG. 3: The performance of deep neural network (DNN) regression models for the prediction of power conversion efficiency values in each data set. **(a)** The box plot showing the distribution of the absolute error of predicted PCE values in terms of median, 25/75th percentile and confidence interval. The inset plot highlights all the box plots that lie in the range from 0 to 1%. **(b)** The learning curves show the dependency of developed models to the training set size. The blue curves belong to the standard deep neural network trained on Morgan fingerprint (FP) representation. The red curves belong to the neural fingerprint (NFP) model. The mean absolute error is calculated based on the evaluation of each model on the test set from the same sample.

In addition, Fig. 3b presents the learning curves that are trained and evaluated on the same data set separately. The dependency of the models to the size of the training set is generally accepted in any ML efforts, and here we see the same trend for all the four samples, as well. The performance of the local models also follows a similar relative trend on the entire range of the training set size. we also note that, in case of Morgan FP, even 30% of the fused (or linked) data can train a better model than 90% of the random data for the prediction of PCE for fused (or linked) molecules. The comparison of the learning curves for two types of models (Morgan FP and NFP) on the random sample shows more than 20% improvement in the prediction accuracy of NFP model across the random test set. This is the main reason that graph convolutional networks have truly increased the excitement in this field. However, this does not mean that the state-of-the-art models can handle the imbalanced data. We still see the same trend as Morgan FP models in terms of the performance of local and global NFP models with respect to each other. The random sample is by far the worse model and after that we have hybrid, linked, and fused models. The underlying point for the current research is that the significantly more complex models sound ineffective for the prediction of the properties of linked molecules. The graph convolutional networks require higher computational complexity than standard DNN models, which is ultimately not tenable for large training data. However, we show that a good clustering approach can significantly mitigate the computational cost and also improve the prediction quality.

All these results confirm that training a model on a random sample is mostly in favor of hybrid molecules that are over-represented in the data set. Note that a better performance of the local models compared to the random model is generally expected. It is well known that ML models intrinsically have better performance for the interpolation tasks rather than extrapolation. Therefore, developing a local model on the portion of data that are of similar characteristics, results in a better performance. However, this point is subject to the clustering approach that lend confidence to the similar characteristics of the subgroups of the molecules. In other words, an arbitrary clustering of a data set that does not emphasize compelling characteristics of the structures may not reproduce the same results.

Furthermore, we show the distribution of the prediction errors across the range of PCE values for each of the test sets, when they are evaluated on their own models (see Fig. 4). The quality of the predictions for the linked and fused models are evenly distributed over the range of the PCE values. This is another advantage of the suggested clustering scheme that leads to molecular candidates, which are: (i) feasible to synthesis, (ii) more desirable with respect to the target property, (iii) homogeneously represented and thus are easier to model, and (iv) approximately equally representative of the remoter but more desirable range of PCE values.

C. Uniform Oversampling of the Training Data

One immediate solution to address the issue with imbalanced data is oversampling of the under-represented clusters. Since enough training data is available in each of the data sets, we simply stack their training data to

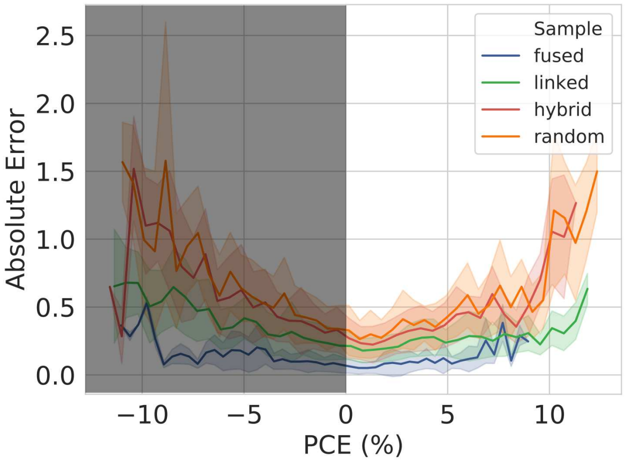


FIG. 4: The distribution of the regression model prediction errors with respect to the PCE values for each of the four samples. The negative PCE values are artifact of the Scharber model, and thus, are faded out.

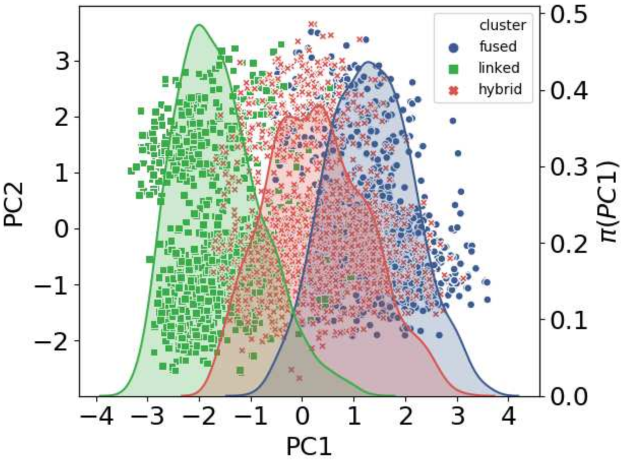


FIG. 5: The scatter plot shows the first and second principal components of a uniform distribution of three clusters specified with color codes. The distribution of first principal component is also plotted accordingly for each cluster. The density of the probabilities are on the right y-axis.

create three times bigger set, but with a uniform distribution of clusters. We develop a Morgan FP model on the uniform training set and evaluate it on the same out-of-sample test sets as before. The result is illustrated in Fig. 3a as the uniform training sample. We observe that the model performs better on all test sets compared to the random sample, except for the hybrid test set. However, the performance of the uniform model is not better than the models purely trained on each of the samples. We can explain this point better with the concept of the distribution shift in the feature representation of the data. For doing so, we first transform the high-dimensional representation (i.e., 2,048 FP vector) to two reduced features by applying principal component analysis (PCA) on 3,000 randomly selected instances of the uniform training set. The scatter plot of the data points with their color-coded clusters is shown in Fig. 5. The population of the data for each cluster clearly illustrates the distribution shift between them. The results show that the linked and fused clusters are very different in terms of the fingerprint representation. Moreover, the distribution of the fused and hybrid principal components has a more significant overlap than the linked and hybrid distributions. All these outcomes are along with our discussions in Sec. III B as well.

A focus on the performance of the models on the hybrid test set suggests that combination of the clusters do not help our model (i.e., the standard DNN) with respect to the choice of descriptor (i.e., the Morgan FP). This point was also observed by comparing the performance of random and hybrid models on the hybrid test set. We also note that creating three times bigger training set should result in a better performance for the model. However, the performance of our models on the hybrid test set decreases from the hybrid model (trained on 100% hybrid data) to random model (trained on 70% hybrid data), and finally to the uniform model (trained on 33% hybrid data). In totality, the uniform model outperforms the random model, and particularly on the fused and linked test sets. Because a focus on the linked molecules is the primary goal of this work, providing a global model that performs better on the linked cluster is a successful outcome.

D. Ensemble Learning by Combining Regression and Classification Models

As we discussed in the previous section, even linear PCA is able to capture the main structural difference between the three clusters. Thereby, our choice of feature representation can linearly distinguish the three suggested categories of molecules. Thus, for developing a global model that has similar performance with our best local models, we propose an ensemble method by merging the classification and regression models. In this section, we train a classifier on the same uniform sample that was created in Sec. III C. The classifier should be able to label an unseen molecule as fused, linked, or hybrid category. We then use our best local regression models accordingly to predict the PCE that corresponds to that molecule.

We optimize a standard DNN model to classify the three clusters. The uniform training set is used for developing the model because it is a fairly balanced data set. We also change the training set size to assess the performance of the classifiers on the very small portions of the data (e.g., 0.001 ratio that corresponds to 135 data points). All the models are evaluated on the same test sets of the initial four samples, and the resulting learning curves are shown in Fig. 6. The figure shows that even with less than 2,000 uniform sample of the CEP data, we can get a model with 98% accuracy to classify three clusters of molecules. The performance of the model on the test sets deteriorates in the order of linked, fused, and then hybrid test sets. These results are also compatible with Fig. 5, because the feature representation of hybrid molecules is distributed between two other clusters and has a more considerable overlap with fused molecules. The underlying point is that the linked cluster is highly distinctive from the other two clusters. Thus, the accuracy of classification model is also a measure of distribution shift between feature representation of clusters.

We further extend the work by merging the classification model and our best local regression models to develop a global predictive model for the CEP data set. The green curve in Fig. 7 presents the learning curve for the ensemble model, evaluated on the random test set. The MAE for ensemble model is lower than both of the Morgan and NFP models across the training set ratio. Table. II summarizes the lowest MAE for the three models. The deep ensemble learning approach presents 31% and 15% improvement compared to the Morgan fingerprint and NFP, respectively. It should be noted that the accuracy of classification model for the linked molecules is close to 100% and thus, the performance of the ensemble model for the linked cluster is very similar to the linked models.

TABLE II: The prediction error of three regression models in terms of mean absolute error (MAE) ± standard deviation. The table summarizes Fig. 7 at 90% training set ratio.

	$MAE(\%)$		
	Morgan FP	NFP	Ensemble
linked	0.288 ± 0.003	0.288 ± 0.002	-
random	0.423 ± 0.004	0.340 ± 0.004	0.290 ± 0.005

IV. CONCLUSIONS

In the work presented here, we introduced a structure-based partitioning scheme for molecular data sets that allows us to identify different domains in compound space. We showed the benefits of creating local ML models that take advantage of the distinct nature of these domains compared to a single global model that does not account for their differences. The improvements in performance and efficiency are considerable, and even standard ML models outperform the most advanced (and correspondingly demanding), state-of-the-art ML approaches. Another attractive feature of our study revealed that local models exhibit a more uniform performance across the spectrum of target property values, including the desirable extremes, for which global models tend to degrade. However, the principal bottleneck for developing local models is the size cap of the minority domains.

We also advanced the use of a classification model and the idea of ensemble learning to achieve the performance of local models but for global predictions. Our statistical

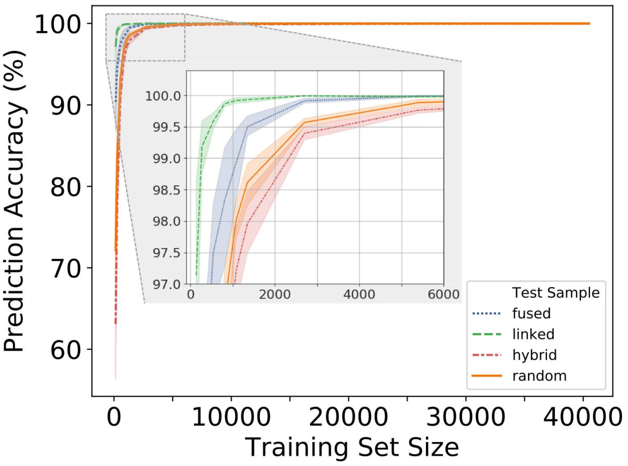


FIG. 6: We develop a model to classify three clusters of the molecules based on the training on the uniform sample of the data. The learning curves show the prediction accuracy of the classifier on four test sets separately. The inset plots focus on the turning point of the learning curves.

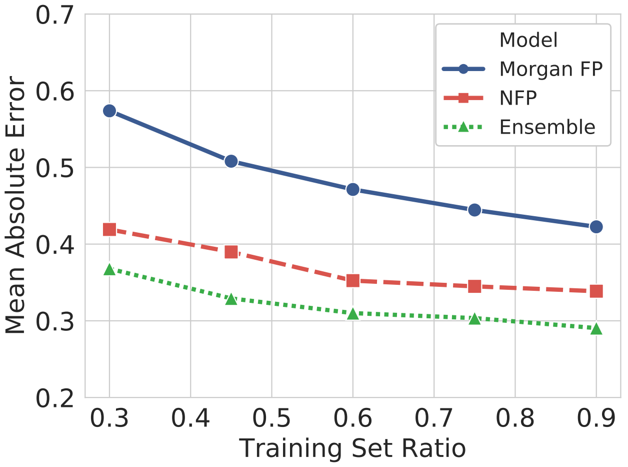


FIG. 7: The learning curves show the decay in mean absolute error of three machine learning (ML) models by increasing the training set size. The blue and red curves belong to the standard neural network and neural fingerprint (NFP) models that are trained on random sample, respectively. The green learning curve presents the proposed ensemble method by merging classification and regression models that are trained and evaluated on three local samples.

analysis of the data set and its imbalance suggests that the latter should be tackled by focusing on the choice of the feature representation, as the sparsity in a feature space can adversely affect a regression task. We propose to resolve this issue by breaking down the imposed sparseness using clustering or classification techniques. Beyond to cluster-aware regression approach presented in this paper, we are currently pursuing an automated process that includes the utilization of unsupervised ML techniques, along with the incorporation (or extraction) of physical priors.

SUPPLEMENTARY MATERIAL

Electronic supplementary material accompanies this paper and is available through the journal website free of charge. It provides statistical analysis of all data sets that are used in this study (Table S1), and tuned hyperparameter values for trained models (Table S2). We also give a link to the repository that data sets are deployed.

COMPETING FINANCIAL INTERESTS

The authors declare to have no competing financial interests.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) CAREER program (grant No. OAC-1751161), and the New York State Center of Excellence in Materials Informatics (grants No. CMI-1140384 and CMI-1148092). Computing time on the high-performance computing clusters '*Rush*', '*Alpha*', '*Beta*', and '*Gamma*' was provided by the UB Center for Computational Research (CCR). The work presented in this paper is part of MH's PhD dissertation [50]. MH gratefully acknowledges support by Phase-I and Phase-II Software Fellowships (grant No. ACI-1547580-479590) of the National Science Foundation (NSF) Molecular Sciences Software Institute (grant No. ACI-1547580) at Virginia Tech [51, 52]. Part of this research was performed while MH was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the NSF. MH also thanks UB Mark Diamond Research Fund (MDRF) for funding part of the research expenses related to this work.

We gratefully acknowledge the use of data from the Clean Energy Project Database, which was generate with generous support from the IBM World Community Grid (WCG) and the computing time donations from its volunteer participants. Additional information about the project as well as links to join the WCG and the Clean Energy Project can be found in Refs. [53] and [54].

[1] Johannes Hachmann, Mohammad Atif Faiz Afzal, Mojtaba Haghighatlari, and Yudhajit Pal, "Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space," Molecular Simulation **44**, 921–929 (2018).

[2] Johannes Hachmann, Theresa L. Windus, John A. McLean, Vanessa Allwardt, Alexandra C. Schrimpe-Rutledge, Mohammad Atif Faiz Afzal, and Mojtaba Haghighatlari, *Framing the role of big data and modern data science in chemistry*, Tech. Rep. (2018).

[3] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[4] Mojtaba Haghighatlari and Johannes Hachmann, "Advances of machine learning in molecular modeling and simulation," Curr. Opin. Chem. Eng. **23**, 51–57 (2019).

[5] Geoffroy Hautier, "Finding the needle in the haystack: Materials discovery and design through computational ab initio high-throughput screening," Computational Materials Science **163**, 108 – 116 (2019).

[6] Roberto Olivares-Amaya, Carlos Amador-Bedolla, Johannes Hachmann, Sule Atahan-Evrenk, Roel S. Sánchez-Carrera, Leslie Vogt, and Alán Aspuru-Guzik, "Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics," Energy & Environmental Science **4**, 4849–4861 (2011).

[7] Mohammad Atif Faiz Afzal, Mojtaba Haghighatlari, Sai Prasad Ganesh, Chong Cheng, and Johannes Hachmann, " Accelerated Discovery of High-Refractive-Index Polyimides via First-Principles Molecular Modeling, Virtual High-Throughput Screening, and Data Mining ," The Journal of Physical Chemistry C **123**, 14610–14618 (2019).

[8] Mohammad Atif Faiz Afzal, Aditya Sonpal, Mojtaba Haghighatlari, Andrew J. Schultz, and Johannes Hachmann, "A deep neural network model for packing density predictions and its application in the study of 1.5 million organic molecules," Chemical Science **10**, 8374–8383 (2019).

[9] Mojtaba Haghighatlari, Gaurav Vishwakarma, Mohammad Atif Faiz Afzal, and Johannes Hachmann, "A Physics-Infused Deep Learning Model for the Prediction of Refractive Indices and Its Use for the Large-Scale Screening of Organic Compound Space," ChemRxiv , 1–9

(2019).

[10] Nathalie Japkowicz, “Learning from Imbalanced Data Sets: A Comparison of Various Strategies,” Proceeding of Association for the Advancement of Artificial Intelligence **1**, 111–117 (2000).

[11] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik, “Inverse molecular design using machine learning: Generative models for matter engineering,” Science **361**, 360–365 (2018).

[12] Bryan R. Goldsmith, Mario Boley, Jilles Vreeken, Matthias Scheffler, and Luca M. Ghiringhelli, “Uncovering structure-property relationships of materials by subgroup discovery,” New Journal of Physics **19** (2017).

[13] Bhavya Kailkhura, Brian Gallagher, Sookyung Kim, Anna Hiszpanski, and T. Yong-Jin Han, “Reliable and Explainable Machine Learning Methods for Accelerated Material Discovery,” Prepr. https//arxiv.org/abs/1901.02717 , 1–24 (2019).

[14] Johannes Hachmann, Roberto Olivares-amaya, Sule Atahan-Evrenk, Carlos Amador-Bedolla, Roel S. Sanchez-Carrera, Aryeh Gold-Parker, Leslie Vogt, Anna M. Brockway, and Alán Aspuru-Guzik, “The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid,” Journal of Physical Chemistry Letters **2**, 2241–2251 (2011).

[15] Edward O. Pyzer-Knapp, Kewei Li, and Alán Aspuru-Guzik, “Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery,” Advanced Functional Materials **25**, 6495–6502 (2015).

[16] David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, “Convolutional Networks on Graphs for Learning Molecular Fingerprints,” in *proceeding of Advances in Neural Information Processing Systems* (2015) pp. 2224–2232.

[17] Hanjun Dai, Bo Dai, and Le Song, “Discriminative Embeddings of Latent Variable Models for Structured Data,” Proc. 33rd International Conference on Machine Learning **48** (2016).

[18] Truong Son Hy, Shubhendu Trivedi, Horace Pan, Brandon M. Anderson, and Risi Kondor, “Predicting molecular properties with covariant compositional networks,” Journal of Chemical Physics **148** (2018).

[19] Johannes Hachmann, Roberto Olivares-Amaya, Adrian Jinich, Anthony L. Appleton, Martin A. Blood-Forsythe, László R. Seress, Carolina Román-Salgado, Kai Trepte, S. Atahan-Evrenk, Süleyman Er, Supriya Shrestha, Rajib Mondal, Anatoliy Sokolov, Zhenan Bao, and Alán Aspuru-Guzik, “Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry-the Harvard Clean Energy Project,” Energy and Environmental Science **7**, 698–704 (2014).

[20] Matthias D'hooghe and Hyun-Joon Ha, eds., *Synthesis of 4- to 7-membered Heterocycles by Ring Expansion* (Springer International Publishing, 2016) p. 367.

[21] Markus C. Scharber, David Mühlbacher, Markus Koppe, Patrick Denk, Christoph Waldauf, Alan J. Heeger, and Christoph J. Brabec, “Design rules for donors in bulk-heterojunction solar cells – towards 10% energy-conversion efficiency,” Advanced Materials **18**, 789–794 (2006).

[22] John P. Perdew, “Density-functional approximation for the correlation energy of the inhomogeneous electron gas,” Physical Review B **33**, 8822–8824 (1986).

[23] A. D. Becke, “Density-functional exchange-energy approximation with correct asymptotic behavior,” Physical Review A **38**, 3098–3100 (1988).

[24] Florian Weigend, Reinhart Ahlrichs, K. A. Peterson, T. H. Dunning, R. M. Pitzer, and A. Bergner, “Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy,” Physical Chemistry Chemical Physics **7**, 3297 (2005).

[25] R.G. Parr and Y. Weitao, *Density-Functional Theory of Atoms and Molecules*, International Series of Monographs on Chemistry (Oxford University Press, 1994).

[26] Wolfram Koch and Max C Holthausen, *A chemist's guide to density functional theory* (John Wiley & Sons, 2015).

[27] Donald B. Johnson, “Efficient algorithms for shortest paths in sparse networks,” Journal of the ACM **24**, 1–13 (1977).

[28] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison, “Open Babel: An open chemical toolbox,” Journal of Cheminformatics **3**, 33 (2011).

[29] Matthias Rupp, O. Anatole Von Lilienfeld, and Kieron Burke, “Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry,” Journal of Chemical Physics **148** (2018), 10.1063/1.5043213, arXiv:1806.02690.

[30] Jrgen Schmidhuber, “Deep learning in neural networks: An overview,” Neural Networks **61**, 85 – 117 (2015).

[31] Adam C. Mater and Michelle L. Coote, “Deep Learning in Chemistry,” Journal of Chemical Information and Modeling **59**, 2545–2559 (2019).

[32] Gaurav Vishwakarma, Mojtaba Haghighatlari, and Johannes Hachmann, “Towards Autonomous Machine Learning in Chemistry via Evolutionary Algorithms,” ChemRxiv , 9782387 (2019).

[33] Federico Girosi, Michael Jones, and Tomaso Poggio, “Regularization theory and neural networks architectures,” Neural Computation **7**, 219–269 (1995).

[34] Mojtaba Haghighatlari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava Urala Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann, “ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data,” ChemRxiv , 8323271 (2019).

[35] Mojtaba Haghighatlari, Gaurav Vishwakarma, Doaa Altarawy, Ramachandran Subramanian, Bhargava Urala Kota, Aditya Sonpal, Srirangaraj Setlur, and Johannes Hachmann, “*ChemML* – A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data,” (2019).

[36] François Chollet, “Keras, available at https://keras.io,” (2015).

[37] Abadi Martin, Agarwal Ashish, Barham Paul, Brevdo Eugene, Chen Zhifeng, Citro Craig, Corrado Greg S., Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Goodfellow Ian, Harp Andrew, Irving Geoffrey, Isard Michael, Yangqing Jia, Jozefowicz Rafal, Kaiser Lukasz, Kudlur Manjunath, Levenberg Josh, Mané Dandelion, Monga Rajat, Moore Sherry, Murray Derek, Olah Chris, Schuster Mike, Shlens Jonathon, Steiner Benoit, Sutskever Ilya, Talwar Kunal, Tucker Paul, Vanhoucke

Vincent, Vasudevan Vijay, Viégas Fernanda, Vinyals Oriol, Warden Pete, Wattenberg Martin, Wicke Martin, Yu Yuan, and Zheng Xiaoqiang, "{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems, available at https://www.tensorflow.org," (2015).

[38] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research **12**, 2825–2830 (2011).

[39] O. Anatole Von Lilienfeld, "Quantum machine learning in chemical compound space," Angewandte Chemie International Edition **57**, 4164–4169 (2018).

[40] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim, "Machine learning in materials informatics: recent applications and prospects," npj Computational Materials **3**, 54 (2017).

[41] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande, "MoleculeNet: A benchmark for molecular machine learning," Chemical Science **9**, 513–530 (2018).

[42] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay, "Analyzing Learned Molecular Representations for Property Prediction," Journal of Chemical Information and Modeling **59**, 3370–3388 (2019).

[43] Harry L Morgan, "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service." Journal of Chemical Documentation **5**, 107–113 (1965).

[44] David Rogers and Mathew Hahn, "Extended-Connectivity Fingerprints," Journal of Chemical Information and Modeling **50**, 742–754 (2010).

[45] Gregory Landrum, "RDKit: Open-source cheminformatics, avaialble at http://www.rdkit.org," (2006).

[46] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl, "Neural Message Passing for Quantum Chemistry," in *Proceeding of 34th International Conference on Machine Learning*, Vol. 70 (JMLR.org, 2017) pp. 1263–1272.

[47] Kristof. T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller, "Schnet : A deep learning architecture for molecules and materials," The Journal of Chemical Physics **148**, 241722 (2018).

[48] Nicolae C. Iovanac and Brett M. Savoie, "Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment," The Journal of Physical Chemistry A **123**, 4295–4302 (2019).

[49] Carlos Amador-Bedolla, Roberto Olivares-Amaya, Johannes Hachmann, and Alán Aspuru-Guzik, "Organic Photovoltaics," in *Informatics for materials science and engineering: Data-driven discovery for accelerated experimentation and application*, edited by Krishna Rajan (Amsterdam: Butterworth-Heinemann, 2013) Chap. 17, pp. 423–442.

[50] Mojtaba Haghighatlari, *Making Machine Learning Work in Chemistry: Methodological Innovation, Software Development, and Application Studies*, Ph.D. thesis, University at Buffalo (2019).

[51] Anna Krylov, Theresa L. Windus, Taylor Barnes, Eliseo Marin-Rimoldi, Jessica A. Nash, Benjamin Pritchard, Daniel G.A. Smith, Doaa Altarawy, Paul Saxe, Cecilia Clementi, T. Daniel Crawford, Robert J. Harrison, Shantenu Jha, Vijay S. Pande, and Teresa Head-Gordon, "Perspective: Computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science," Journal of Chemical Physics **149**, 180901 (2018).

[52] Nancy Wilkins-Diehr and T. Daniel Crawford, "NSF's inaugural software institutes: The science gateways community institute and the molecular sciences software institute," Computing in Science & Engineering **20**, 26–38 (2018).

[53] http://cleanenergy.molecularspace.org (accessed July 3, 2013).

[54] http://www.worldcommunitygrid.org (accessed July 3, 2013).