Artificial Intelligence Chemistry 2 (2024) 100050

Contents lists available at ScienceDirect

Artificial Intelligence Chemistry

journal homepage: www.journals.elsevier.com/artificial-intelligence-chemistry

Applying graph neural network models to molecular property prediction using high-quality experimental data☆

Chen Qu *, Barry I. Schneider *, Anthony J. Kearsley, Walid Keyrouz, Thomas C. Allison

National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA

ARTICLE INFO

Keywords:
Kováts retention index
Boiling point
Mass spectrum
Graph neural network
Deep learning

ABSTRACT

Graph neural networks have been successfully applied to machine learning models related to molecules and crystals, due to the similarity between a molecule/crystal and a graph. In this paper, we present three models that are trained with high-quality experimental data to predict three molecular properties (Kováts retention index, normal boiling point, and mass spectrum), using the same GNN architecture. We show that graph representations of molecules, combined with deep learning methodologies and high-quality data sets, lead to accurate machine learning models to predict molecular properties.

1. Introduction

In recent years, the adoption and efficacy of machine learning (ML) and artificial intelligence (AI) have advanced rapidly due to more powerful compute hardware and software as well as the availability of large data sets. Breakthrough applications include AlphaGo [1], large language models (such as the Generative Pre-training Transformer, GPT), and self-driving cars. ML methodologies have been applied to molecular sciences and chemistry, such as material design and drug discovery [2–9], synthesis planning and reaction optimization [10–13], protein structure prediction [14,15], and to a wide range of theoretical/computational chemistry targets [16–20]. Among these applications, predicting molecular properties stands out as a key component in drug and materials design, and this is the subject of this paper.

Data play a central role in ML. Unfortunately, experimental data in the physical sciences are often scarce and costly to assemble. (Data sets presented in this article are products of decades of curation and even longer to measure.) There are several strategies to address data scarcity. The first is to use theory and computation to generate data. For example, density functional theory can produce large sets of data rapidly (in some cases) to augment or supplement measurement data. However, accurate and reliable computations are still costly; the cost of "gold-standard" coupled-cluster theory scales as N^7, where N is the size of the electronic space. This scaling makes the coupled cluster method prohibitively expensive for molecules with more than tens of atoms. An alternate strategy accepts the scarcity of relevant data and trains a model with small data sets. For example, active learning [21–23] can guide experiments (or expensive computations) in selecting the most informative data points, thereby reducing the size of the data set needed for a successful model. Transfer learning [24–26] uses a pre-trained model that is based on abundant data for a related problem, and then tunes the model with a small amount of data for the target problem. The quantity of experimental data is not the only concern; the quality of experimental data usually varies due to dependence on experimental conditions, instrument precision, sample purity, and even the instrument operator.

Once a suitable problem has been identified and a dataset has been selected for training the ML model, it is necessary to select the features (i.e., input data) that will be used in the training process. Well-chosen features may lead to a robust model, whereas other features may have little effect on the performance of a model. For molecular property prediction, a key question can be the representation of the structure of the molecule or crystal [27]. When the 3D coordinates of the atoms in the molecules are available (e.g., for applications in theoretical and computational chemistry), some widely used molecular representations include Coulomb matrix [28], bag-of-bonds [29], atom-centered symmetry functions [30], and smooth overlap of atomic positions [31]. When 3D coordinates are not available, which is often true for experimental data, the simplified molecular input line entry system (SMILES) representation [32] or molecular fingerprints [33] can be used as the input to various machine learning models.

We employed a 2D topological molecular graph as the representation for molecular systems and a graph neural network (GNN) as these are

☆ Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

* Corresponding authors.
E-mail addresses: chen.qu@nist.gov (C. Qu), barry.schneider@nist.gov (B.I. Schneider).

https://doi.org/10.1016/j.aichem.2024.100050
Received 28 September 2023; Received in revised form 15 December 2023; Accepted 11 January 2024
Available online 19 January 2024

more closely aligned with the internal structure of chemical compounds where chemical bonds play a prominent role in the properties of these compounds. Intuitively, it is easier for a GNN to learn from a chemical compound that is itself represented as a graph. One advantage of a graph representation is that it is a natural and expressive representation of a molecule: nodes in the graph correspond to atomic centers and edges of the graph correspond to chemical bonds. For each node and edge, a feature vector containing attributes of the corresponding atom or bond, such as the atom type, hybridization, or bond order provides the input data for the model. In the model described in this article, these feature vectors are updated during training using information from neighboring atoms and bonds exchanged via message passing [34,35] between the nodes and edges of the molecular graph, giving the model additional flexibility to predict the molecular property of interest. Additional details of the GNN model are given in Section 3.

Here we present applications of the GNN model trained on high-quality experimental data sets curated by NIST [36–38]. In particular, we predict Kováts retention indices [39], normal boiling points, and mass spectra. Before presenting the details of our GNN models and the data libraries used to train these models, we first give some background information on Kováts retention indices, normal boiling points, and mass spectra.

Gas chromatography (GC) is an important analytical technique for the separation and identification of chemical compounds. In a GC experiment, a mixture of target substances, often unknowns, in a gaseous state and a carrier gas is passed through a chromatography column. The time elapsed before the unknown compound passes through the column is indexed against the elution times of known compounds; this index is called the retention index. It has been demonstrated [39] that the retention index can be made independent of many experimental factors such as column length, column diameter, and film thickness. This results in a dimensionless quantity known as the Kováts retention index.

GC is frequently used in combination with mass spectrometry (GC/MS) as a means of enhancing the accuracy of identifying chemical compounds. In this context, matching the retention index can significantly improve the confidence in results generated by library searching versus use of the mass spectrum alone [40]. Therefore, accurate prediction of Kováts retention indices has considerable value and many techniques have been employed to predict retention indices [41–45].

In addition to combining library searching with the retention index, augmenting existing mass spectral libraries is another strategy to better identify unknown compounds [46,47]. In 2019, the Google Brain team reported results from their attempt to predict mass spectra using machine learning. Their machine learning methodology, a multi-layer perceptron (MLP) model trained with the NIST mass spectral library (2017 version) achieved reasonably good results [48]. However, as noted by the authors, further improvements were possible. Recently, GNNs have also been used to accurately predict the mass spectra [49,50]. Zhang et al. employed a similar GNN approach, which inputs a whole molecular graph and outputs a vector representing the intensity at each integer mass-to-charge ratio (m/z), and this model achieved an accuracy comparable to our model [49]. The other model developed by Zhu and Jonas uses a different approach: it first enumerates possible fragment formulae or possible subsets of atoms, and uses a GNN to predict the probability that fragment formula or subset occurs in the experiment. The latter is the best model available at present, not only due to its accuracy, but also because of two additional advantages over our model: first, the peaks (positions and intensities) due to isotope substitution can be naturally obtained by using the fragment formulae or subsets and their corresponding probability; second, this methodology can be generalized easily for high-resolution mass spectra, while our GNN model only applies to integer m/z, and needs to be retrained for a different resolution.

Normal boiling point, albeit a relatively simple property, is still routinely used as a measure of the purity of chemical substances. As such, predicting normal boiling points using a variety of methods continues to be of significant interest. Determination of the normal boiling point dates back hundreds of years [51]. Quantitative Structure-Property Relationship (QSPR) models have been used extensively to predict the normal boiling points [52,53], and recently, machine learning methodologies have been applied [54–58].

Having set the stage for the importance of predicting these three properties, we now describe our data set, model, and results with an emphasis on the importance of the graph representation and the quality of the data sets.

2. Data sets

NIST has been collecting data from various sources such as literature and research labs across the world, and more importantly, has critically evaluated the data for their reliability, resulting in high-quality libraries that are suitable for ML applications. The two data libraries described below are used in our work.

2.1. NIST 20 mass spectral library

The 2020 release of the NIST/EPA/NIH Mass Spectral Library [59] is a critically evaluated [60] collection containing 306,869 molecules with their corresponding mass spectra. Specifically, the library contains a 2D representation of each molecule (in Molfile format), its mass spectrum, a measured (i.e., experimental) value of the Kováts retention index (available for about 112,000 molecules), a predicted value of the Kováts retention index based on the model of Stein et al. [41], an estimated uncertainty on the predicted retention index value, as well as other data and metadata on the chemical compound.

The mass spectrum of a compound in this library consists of a set of m/z, rounded to integer values, and the corresponding intensities (which are proportional to the abundance of ions with those m/z ratios). The intensity of the highest peak is set to 999 for each compound. This is because the absolute intensities depend on many experimental factors, and relative intensities are what we use in spectral-matching algorithms.

The primary purpose of the library is for use in matching unknown chemical compounds to aid in mass spectral identification via direct matching of the measured mass spectrum to a library spectrum. By matching the retention indices first, the number of searches for a matching mass spectrum can be greatly reduced [41]. This makes collecting Kováts retention indices useful in some mass spectral matching schemes.

2.2. TRC

The NIST Thermodynamics Research Center (TRC) SOURCE Data Archival System has captured 22,935 determinations of normal boiling points of different molecules from numerous literature sources. In many cases, there are multiple measurements of the normal boiling point for a single molecule, permitting the selection of a consensus value. In total, the data set consists of about 4000 molecules and corresponding normal boiling points.

2.3. Preprocessing of the data sets

The sets of molecules in both libraries were processed prior to training the ML model to ensure that the data set has sufficient information to adequately represent chemical functionalities and corresponding properties. Each of the filters described below was used to ensure that the data set contains a sufficient number of molecules so that the training procedure is able to learn from a larger number of molecules as opposed to fitting many disparate cases.

First, the number of occurrences of a given atom type in the molecules in the data set was counted. If a particular atom occurred only in a small fraction (<1 %) of molecules, molecules containing that atom

were excluded from the set. For example, among the 306,869 molecules in the NIST 20 Mass Spectral Library, 3160 of them contain one or more I atoms. The next frequent atom, B, only appears in 1456 molecules. We decided to exclude molecules containing B atoms because the library does not have sufficient samples for a good representation of B-containing molecules. For both datasets, we ended up with molecules containing only the following atoms: C, H, O, N, Cl, F, Br, S, Si, P, and I.

Next, we examined histograms of molecular mass and molecular properties (i.e., Kováts retention indices and normal boiling points), and removed molecules with extreme values, because the data at the extreme values are too scarce to produce a meaningful predictive model. As an example, Fig. 1 shows the distribution of molecular mass for molecules in the NIST 20 library. In the histogram plot, it can be seen that there are few entries with molecular mass smaller than 50 amu or larger than 850 amu, so the mass range was set to these bounds when training the model to predict the mass spectra of molecules. Based on this filter, only 307 molecules were eliminated.

In the case of normal boiling points, where multiple determinations of the boiling point for the same molecule are available, a single consensus boiling point is determined as the mean of the set of values. For sets with three or more measurements, the Grubbs 2-tail outlier test [61] was applied to filter out the data points that significantly deviate from the remaining data in this set.

3. Method

Our machine learning model is based on the materials graph network (MEGNet) approach developed by Chen et al. [36,62]. The MEGNet methodology incorporates a graph network architecture that captures molecular structure in a very natural way, providing a powerful framework for machine learning of chemical properties.

It is worth mentioning that MEGNet has been extended to M3GNet, which incorporates 3-body interactions (i.e., angular information) in addition to the distance information [63]. It is shown that models using angular information [64–67] or incorporating equivariant message passing[68–70] can outperform distance only models. Nevertheless, we still employed the MEGNet model because we only used 2D connectivity information of the molecule.

3.1. General architecture of the MEGNet model

In the present graph neural network (GNN) model, atomic centers correspond to nodes in the graph and chemical bonds correspond to graph edges. The input data to the model is obtained from a 2D Molfile representation of the molecule. This format contains information about the atoms and their chemical bonding, but does not provide any 3D structural information.

The MEGNet methodology captures molecular information at the level of atoms, chemical bonds, and the whole molecule, with the chemical structure contained in the structure of the graph representation. This model has been tested for a variety of chemical properties on both molecular and crystalline systems and found to perform well [36].

The GNN model used in this study typically incorporates 5 atomic features, 3 bond features (more strictly, pair features, see below), and 3 global features. These features are summarized in Table 1; they were selected via a trial and error process, considering a large number of features and testing the sensitivity of the model to including/removing various features. Among these features, ring sizes and atomic mass are only used for the mass spectrum, while the remaining features are used in all three problems. Note that for bond (pair) features, we actually considered atom pairs, not just those that are formally chemically bonded, so "no bond" is also a possible value in bond type. The graph distance is defined as the smallest number of *chemical bonds* between the atoms in the pair. The model does not encode the complete graph of the molecule as pairs of atoms with a graph distance greater than 5 are excluded, leading to a considerable reduction in the memory required to train the model. All these features were computed using RDKit [71].

We only used 2D information of the molecule—only connectivity, without 3D structural information such as bond length and angles. The main reason for not using 3D information is that 3D structures are not available in these library. We attempted to use 3D structure information from conversion of 2D structures and force field (molecular mechanics) minimization, but abandoned this approach because we did not see significant improvement. In addition, our goal is to facilitate rapid prediction of certain molecular properties; using Molfile input which is easily created with chemical structure drawing software or obtained via name to structure conversion allows us to develop an efficient workflow for property prediction.

The feature vectors are inputs to "MEGNet blocks," which are composed of two layers of densely-connected multi-layer perceptrons. We used (128, 64) units in these two layers for retention index, (64, 32) units for boiling point, and (256, 128) units for mass spectrum. These layers are input to a "message-passing" block where the atomic, bond, and global attributes are successively updated. The MEGNet blocks (we used 3 blocks for retention index and boiling point, and 6 blocks for mass spectrum) are followed by a "Set2Set" readout function in which the output of the atomic and bonding attributes are mapped to the appropriate vector quantities. This is followed by a concatenation step and a few densely-connected layers before the final output. We used (64, 32) units in these densely-connected layers for retention index, (32, 16) for boiling point, and (2000, 2000, 2000) for mass spectrum. These hyperparameters mentioned above were tuned for each system. The rectified linear unit ("ReLU") activation function is used in all layers

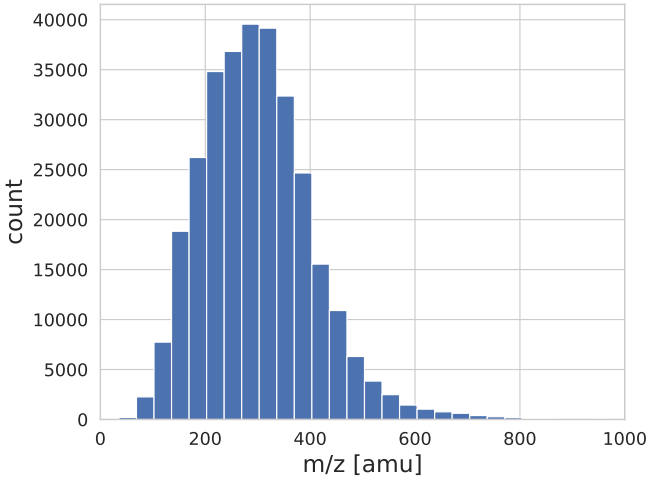

Fig. 1. Distribution of molecular mass for molecules in the NIST 20 library.

Table 1
Features used in the GNN model for mass spectrum.

Feature	Feature Type	Length	Meaning
Atom type	atomic	11	one-hot encoding for 11 possible atoms
Atomic mass	atomic	1	scalar
Hybridization	atomic	6	one-hot encoding for [s, sp, sp^2, sp^3, sp^3d, sp^3d^2]
Formal charge	atomic	1	scalar
Bond type	bonding	5	one-hot [no bond, single, double, triple, aromatic]
Same ring	bonding	1	0/1 whether two atoms are in the same ring
Graph distance	bonding	1	scalar
Number of bonds	global	1	scalar
Molecular mass M	global	1	scalar
Number of non-H atoms	global	1	scalar

except the final output layer, where different activation functions are used for different problems, as described in detail below.

3.2. Adaptation for different problems

3.2.1. Mass spectra

Prediction of the mass spectrum is a more challenging problem than its Kováts retention index or normal boiling point counterparts, because both retention indices and boiling points are scalar quantities, requiring the prediction of a single quantity by the model. In the case of mass spectra, a spectrum, which consists of the positions (m/z) of spectral signals and their corresponding amplitudes, must be predicted. The first step is to find an appropriate numerical representation of the mass spectrum. The experimental mass spectrum of a molecule, as obtained from the NIST library, is given as a set of m/z ratios (corresponding to molecular fragments) with the corresponding relative abundance (i.e., intensity) of the fragment. Both quantities are given as integers. Therefore, we convert a mass spectrum to a vector of length 1000, with each entry of the vector representing an integer value of m/z up to 1000, with the k-th entry in this vector representing the intensity at $m/z = k$. The spectra are then normalized such that its L_1-norm of the relative abundance is 1.0. In addition, a baseline intensity of 10^{-7} is added, so that the minimum intensity is 10^{-7}. This was done due to the particular form of the error measure that was used in the model as we describe below. The spectra represented in this manner were used as the desired output of the GNN model. Because the length of the output vector is 1000, this model can only predict the portion of the mass spectrum with $m/z < 1000$. For molecules whose masses are greater than 1000, the model cannot predict the complete spectra, and therefore not recommended. Furthermore, all molecules whose masses are greater than 850 amu and smaller than 50 amu were excluded when training this model, so it is also not recommended to use this model to predict for molecules with a mass significantly beyond this range.

The model output is a vector of size 1000, making the task of training the NN model much more challenging, with much of the challenge associated with the choice of an appropriate error function for use in training. As a result, the present model differs considerably from those in previous work by the authors [37,38].

There is an intrinsic symmetry in the mass spectrum, that is, if a fragment with mass x exists, a fragment with mass $M–x$ is likely to be present in the spectrum, where M is the molecular mass. This intrinsic symmetry can be exploited in the ML model by using the bidirectional prediction approach by Wei et al. [48]. Specifically, the prediction is

$$v = \sigma(g) \odot v^f + [1 - \sigma(g)] \odot v^r, \tag{1}$$

where v^f and v^r are forward and reverse predictions, respectively, g is an affine transformation of the last hidden layer X (i.e., $g = W \cdot X + b$), σ is the sigmoid function, and $\odot$ denotes component-wise multiplication. (See Wei et al. [48] for full details.).

Predictions from the model are scaled such that the intensity of the highest peak is equal to 999 and values are rounded to the nearest integer. This is done to mimic the format of the NIST20 reference library.

3.3. Loss function

We used mean-absolute-error as the loss function for normal boiling points and Kováts retention indices. In these two models, this loss function is better than the mean-squared error typically used for regressions [37,38].

For mass spectra, as stated above, a large part of the difficulty in training a robust and reliable model to predict mass spectra relies on the error function used in training the model. A function is needed to compute the "distance" between two spectra, giving a small distance (low error) when two spectra are similar and a large distance (high error) when two spectra have very little in common. Evaluating this similarity involves two dimensions, the m/z value and the relative abundance (or intensity). Since the distance function returns a single scalar representing the similarity of two spectra, it is easy for many details to be obscured in the sum over hundreds of m/z values and thus not accessible to optimization. We considered two different loss functions, the earth-mover's distance (EMD, Eq. (2)), a widely used similarity measure for histograms [72] and therefore also applicable to mass spectra [73], and the symmetrized version of the Kullback-Leibler divergence (KL, Eq. (3)), inspired by a message-passing NN model that predicts infrared spectra [74]. We trained two GNN models using either one of these two as the loss function, and the predictive error of the two models are assessed using two additional error metrics (mass-weighted mean-square error, Eq. (4), and mass-weighted cosine similarity, Eq. (5)). The model trained with KL divergence produces slightly lower mass-weighted mean-square error, and higher mass-weighted cosine similarity. In addition, the predicted spectra from the model trained with EMD have more spurious lines. Therefore, we chose symmetrized version of KL divergence as the loss function. The definition of these similarity measures or error metrics are given below:

$$\mathrm{EMD}(\boldsymbol{u}, \boldsymbol{v}) = \sum_k \left| \mathscr{C}(\boldsymbol{u}, \boldsymbol{v}) \right|, \tag{2}$$

$$\mathrm{KL}(\boldsymbol{u}, \boldsymbol{v}) = \sum_k \left[u_k \ln\left(\frac{u_k}{v_k}\right) + v_k \ln\left(\frac{v_k}{u_k}\right) \right], \tag{3}$$

$$\mathrm{WMSE}(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{N} \sum_k \left(\frac{m_k \sqrt{u_k}}{\left\| \sum_k (m_k \sqrt{u_k})^2 \right\|} - \frac{m_k \sqrt{v_k}}{\left\| \sum_k (m_k \sqrt{v_k})^2 \right\|} \right)^2 \tag{4}$$

$$\mathrm{WCS}(\boldsymbol{u}, \boldsymbol{v}) = \frac{\sum_k m_k \sqrt{u_k} m_k \sqrt{v_k}}{\left\| \sum_k (m_k \sqrt{u_k})^2 \right\| \cdot \left\| \sum_k (m_k \sqrt{v_k})^2 \right\|}, \tag{5}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are the reference and predicted spectra, respectively; $\mathscr{C}(u, v)$ is the cumulative sum of the difference of spectra u and v evaluated over all values of m/z; index k runs from 1 to the maximum allowed value of m/z (N, 1000 in the present case); u_k is the relative abundance (intensity) of the kth m/z peak for the known (library) spectrum, and v_k is the corresponding quantity for the predicted spectrum.

3.4. Activation function of the last layer

Scalar quantities were normalized via computing the z-score $y' = \frac{y - \mu_y}{\sigma_y}$, where y' is the value of the normalized quantity, y is the corresponding unnormalized quantity, μ_y is the mean of y, and σ_y is the standard deviation of y. Therefore, the distribution of y' is centered at 0, with both positive and negative values, and a linear activation is appropriate. For mass spectra, on the other hand, since all intensities are non-negative values, and they span three orders of magnitude, an exponential activation function is used.

3.5. Training and validation

To facilitate the eventual model validation, the data set was divided into 10 equally sized "folds" of randomly selected data. During model training, 80 % of the data was used as the training set with an additional 10 % of the data used as a validation set. The remaining 10 % of the data is used as a testing set.

The training was conducted using the "Adam" optimizer with mini batches, and using hyperparameters tuned for each problem. The batch size is 32 for retention index and boiling point, and 64 for mass spectrum. The learning rate is 2×10^{-4}. Early stopping is used for the training to mitigate overfitting. Briefly, we monitor the loss function of the validation set and if it does not improve for a certain number of epochs (150 steps for retention index, 300 for boiling point, and 100

steps for mass spectrum), the training is terminated.

4. Results and discussion

4.1. Mass spectrum

There are a number of approaches that can be taken in evaluating the performance of a model used to predict mass spectra. Among these, two strategies emerge. The first is to use a measure that describes the error in peak location and height on a peak-by-peak or aggregate basis. The second is to evaluate the performance of the model in its ultimate application, in this case matching library spectra. Both approaches are utilized in this article. However, as assessing the performance of the model based on the Kullback-Leibler measure produces values that are not intuitive, additional attention is given to measures of the first type.

The first of these additional measures is the root sum of squared errors (RSSE), defined as

$$\mathrm{RSSE}(u,v) = \sqrt{\sum\nolimits_k (u_k - v_k)^2} \tag{6}$$

where u and v are the reference and predicted spectra, respectively. This function, like the Kullback-Leibler measure, but unlike the MSE, has the advantage that the units may be directly compared to those of the mass spectrum (i.e., intensity). Since the spectra are normalized prior to training, the values of this metric may be compared to unity; an error metric of one means that the differences in the peak heights of the reference and predicted spectra differ by as much as the sum of the peak heights in either spectrum. Note that differences in peak heights are important, but differences in the location of these peaks (corresponding to the value of m/z) are much more important from a physical point of view, in particular because many library matching functions rely on accurately matching the m/z values.

The second additional measure is the sum of absolute errors (SAE), defined as

$$\mathrm{SAE}(u,v) = \sum\nolimits_k \left| u_k - v_k \right| \tag{7}$$

This error measure is expected to give similar information to the RSSE.

Values for the Kullback-Leibler, RSSE, and SAE measures are given in Table 2 for the model presented in this article. The results are presented for the training, validation, and testing data sets given by a 10-fold cross-validation testing protocol. Briefly, we divided the data set into 10 sets (or "folds") and then trained the model 10 times, each time using 8 folds as the training set, 1 fold as the validation set, and the remaining 1 fold as the testing set. The minimum, maximum, median, mean, and standard deviation of the mean errors are presented. It is immediately apparent that the model used in this study is subject to strong overfitting as evidenced by the larger error measures in the testing versus training data.

The performance of the present model may be compared to that of the model of Wei et al. [48]. From the values in Table 3, it can be seen that their model experiences much less overfitting. However, the values of the Kullback-Leibler and SAE measures are considerably larger for that model compared to the results of the present study. The results for the Kullback-Leibler measure can be rationalized by considering that this was the error function used in training the present model. In contrast, the differences in the values of RSSE and SAE are not explained by the choice of the training function. The present model gives RSSE and SAE values that are 4–6 times smaller than the model of Wei et al. with a smaller standard deviation. Finally, note that the analysis of the data set of Wei et al. [48] has been limited to the testing set used in that paper, meaning that the model is tested on approximately 11,500 spectra instead of the full set.

Table 2
Performance of the present model based on 10-fold cross validation.

Set	n	min ϵ	max ϵ	median ϵ	mean ϵ	σ (ϵ)
Kullback-Leibler						
training	2,303,208	0.0000	6.5821	0.3064	0.3411	0.1961
validation	287,901	0.0172	14.8265	0.9039	1.1534	0.9694
testing	287,901	0.0159	20.5033	0.9041	1.1549	0.9727
RSSE						
training	2,303,208	0.0012	0.8768	0.0806	0.0944	0.0518
validation	287,901	0.0129	1.0279	0.1693	0.1882	0.1069
testing	287,901	0.0110	1.0807	0.1695	0.1885	0.1071
SAE						
training	2,303,208	0.0023	1.7404	0.3694	0.3775	0.1269
validation	287,901	0.0458	1.9940	0.6948	0.7130	0.3275
testing	287,901	0.0529	1.9869	0.6952	0.7134	0.3279

Table 3
Performance of the model of Wei et al. [48].

Set	n	min ϵ	max ϵ	median ϵ	mean ϵ	σ (ϵ)
Kullback-Leibler						
training	236,355	0.0000	31.2071	5.2212	5.6492	2.9698
validation	11,505	0.2563	33.2832	5.5240	6.2335	3.3530
testing	11,494	0.4299	31.7778	5.5184	6.2369	3.4011
RSSE						
training	236,355	0.0150	1.3602	0.5666	0.5735	0.1017
validation	11,505	0.2468	1.3497	0.6514	0.6718	0.1603
testing	11,494	0.2263	1.3878	0.6533	0.6747	0.1635
SAE						
training	236,355	0.0151	15.9228	4.2806	4.4963	1.5801
validation	11,505	0.7554	18.7222	4.1441	4.4892	1.7326
testing	11,494	0.7415	15.6727	4.1561	4.4978	1.7301

Next, we compare the results of our model to those of the models developed by Zhu and Jonas. Fig. 2 compares the mass-weighted cosine similarity and the Stein dot product [75] among indicated models. As we can see, the models by Zhu and Jonas, the best model at present, are indeed more precise than ours and the model by Wei et al., especially the accuracy on the 10th percentile (bottom of the bars).

The issue of overfitting is difficult to overcome in the present model. The primary method of reducing or controlling overfitting in the present model is the use of dropout layers. However, this strategy was less effective than desired as were other strategies such as using more data in the validation set during training. The model of Wei et al. [48] employs a specialized approach (deep residual learning [76]) to improve model training and avoid overfitting and the results suggest that this is quite successful. The deep residual learning approach is not part of the

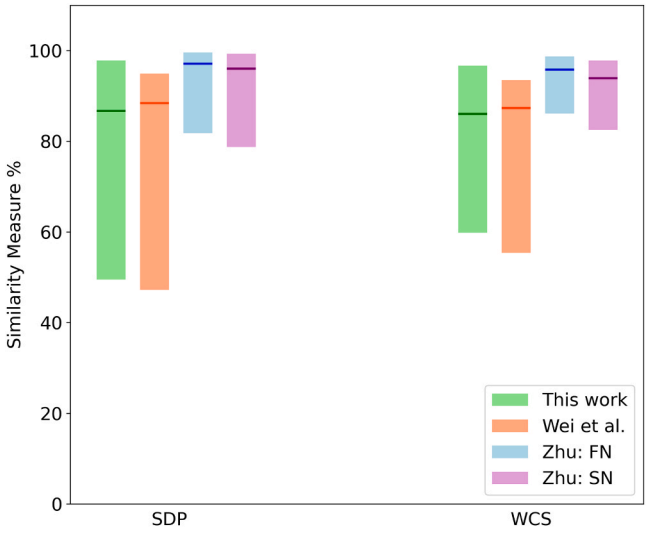

Fig. 2. Mass spectra prediction performance of four models on the NIST library: our model, the model by Wei et al., two models by Zhu and Jonas (SubsetNet, SN, and FormulaNet, FN). The bottom and top of the bars represent the 10th and the 90th percentiles, respectively, with the middle bold tick representing the median.

MEGNet model used in this work, but it will be incorporated into the model in future applications. It is expected that this will lead to a significant improvement in model performance.

Three additional error measures were computed to more fully characterize model performance: mass-weighted mean-square error (as employed by Wei et al. [48]), cosine similarity, and earth mover's distance. All spectra are normalized such that their L_1 norm is 1.0 before computing the similarity measures.

Results for the Kullback-Leibler measure and the three additional similarity measures defined above are given as histogram plots in Fig. 3, where the results of the present model and the model of Wei et al. [48] are both shown. In general, the present model has higher frequency of small errors, but has a longer tail (i.e., more errors with larger values) as compared to the results of Wei et al. [48]. One unexpected result seen in these plots is that the model of Wei et al. peaks at larger error values compared to the present model, whereas it might be expected that the histogram would peak at the smallest error values. This appears to be an artifact of the tuning of that model to maximize the performance on the recall@ 10 metric (a measure of the number of times the correct spectrum was ranked among the 10 best matching scores). When that model is trained in a similar manner to the present work, the errors are larger. This result might have been anticipated by the authors of that article in which they suggested that a GNN approach could outperform their MLP-based model.

Finally, the performance of the model is assessed by comparing the results of the model developed in this work against the reference data, plotted as a mass spectrum (Fig. 4) with the logarithm of the intensity on the *y* axis to better show intensities that differ by orders of magnitude. Each panel in the figure indicates the relative quality of the result by giving the percentile ranking (*k*th percentile ranking means this prediction is better than *k*% of all predictions) based on the value of the KL measure. In the figure, it is clear that spectra that are poorly predicted have significant spurious and missing peaks and that the intensities can be quite different. In contrast, spectra that are well predicted have few

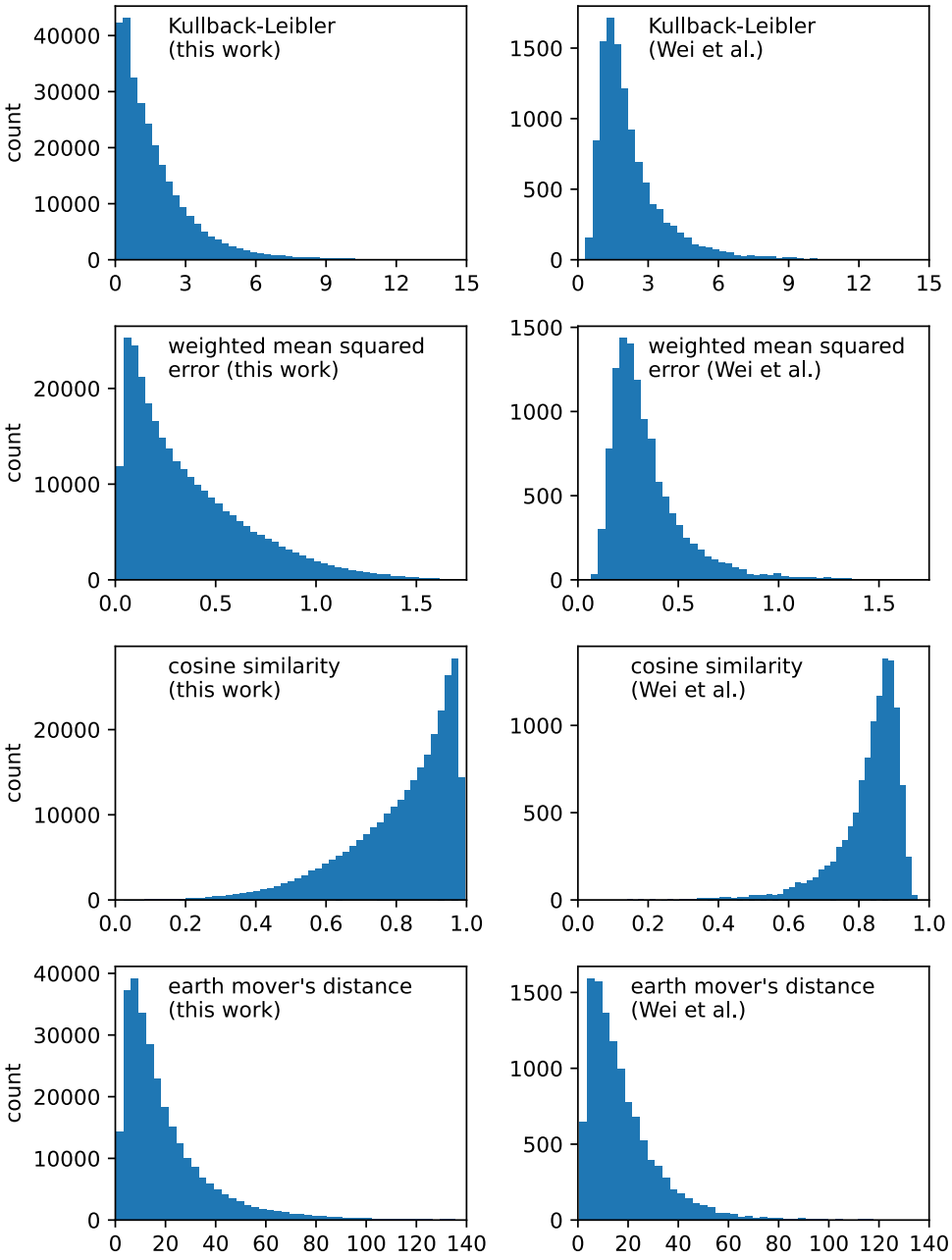

Fig. 3. Distributions of similarity measure values for the predicted spectra.

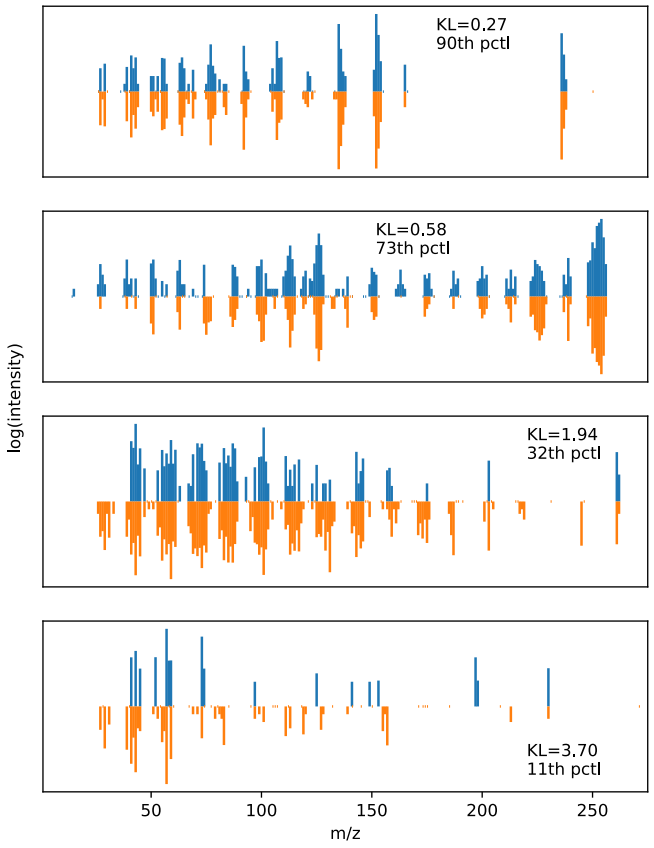

Fig. 4. Selected predictions of the model used in this work for four randomly-chosen chemical compounds representing different levels of prediction accuracy. A higher percentile rank indicates better agreement between the experimental spectrum (above, blue) and the prediction (below, orange).

missing or spurious peaks and the intensities are more consistent.

Scaling and rounding of the model output (as described above) eliminates a number of spurious peaks with very small intensity. This procedure is sufficient to remove peaks above the molecular mass without the need for an additional filtering step. This is a strong indication that the training error function in the present model is working well for this problem.

It is interesting that the present model is trained in a few hundred epochs and gives the results shown herein. In contrast, the work of Wei et al. [48] used 100,000 epochs for model training. This suggests another potential avenue for improving the present model that is being pursued.

In analyzing the predictions made by the present model, it was noted that molecules with large numbers of rings or those with multiple, fused ring systems tended to be less well predicted, whereas molecules with a significant alkane character tended to be better predicted. This suggests that additional features describing ring systems in more detail may lead to model improvement, and this is currently being investigated.

4.2. *Kováts retention indices and boiling points*

The overall statistics of the 10-fold cross-validation for the two models are shown in Table 4. One can immediately notice that the errors of the validation and testing sets are much larger than those of the training sets, which means that overfitting is present in both models. The early-stopping strategy used during the training cannot completely prevent overfitting. We tested other strategies such as dropout and adding regularization terms to the loss function. These strategies indeed decrease the gap between training error and validation/testing error, indicating that they are effective in mitigating overfitting. However, these strategies led to larger overall validation errors than those shown in Table 4. As a result, we did not incorporate dropout and regularization in our final models.

Table 4
Summary statistics of the 10-fold cross validation procedure. The mean value and standard deviation of the mean absolute error (MAE) and the root mean square error (RMSE) over 10 runs is given for each of the 3 sets used.

	Kovats (unitless)		Boiling Point (K)	
Set	MAE	RMSE	MAE	RMSE
Training	9.38 ± 0.86	18.27 ± 1.87	2.30 ± 0.44	2.80 ± 0.51
Validation	27.84 ± 0.67	57.77 ± 1.97	5.56 ± 0.52	7.78 ± 1.66
Testing	28.09 ± 0.72	58.43 ± 1.93	5.77 ± 0.40	7.81 ± 0.94

Next, we compare the performance of our GNN models with other models on Kováts retention indices and normal boiling points. Table 5 shows the comparison between our GNN model on Kováts retention indices and the other two models on the same property. One is a group increment model by Stein et al. [41], whose predictions are part of the NIST reference library of mass spectrum. The other model is a convolutional neural network (CNN) by Matyushin et al. [45], which used SMILES representation of the molecule. The original CNN model was trained with the NIST 08 library, which contains fewer (experimental) determinations of the retention index than are available in the NIST 20, so we retrained the CNN model with NIST 20.

It is clear from Table 5 that both ML models perform significantly better than the group increment model by Stein et al., likely due in part to the limitations in the group increment methodology and in part due to the use of the less powerful linear least squares for model optimization, while the machine-learned GNN model can more systematically explore datasets and discover patterns and relationships embedded in data. Our model also achieved smaller prediction errors than those of the CNN model. We suggest that the superior results are likely due to the better molecular representation used in GNN and to the larger parameter space of our model.

For the normal boiling point model, we characterized its performance by comparing it to the method of Stein and Brown as implemented in the EPI Suite package [77]. The MAE of the Stein and Brown method over all compounds in our data set is 11.84 K with a median error of 7.99 K and a standard deviation of 12.84 K. Our model is performing about a factor of two more accurate.

The results above demonstrate the importance of using a robust, expressive molecular representation and model. Next, we discuss the importance of high-quality data sets.

4.3. *Importance of high-quality data*

We applied the same GNN methodology to predict the retention indices, trained with two NIST mass spectrum data sets: the NIST 20 library described above and the 2017 version of the library. With the same hyperparameters, we found that the predictive error of the model trained on the 2017 version of the library is about twice as large as the error of the model trained on the newer 2020 version. To investigate what causes the larger error when trained with the NIST 17 library, we computed the molecular fingerprints of all the compounds in the NIST 17 library, and found all pairs of molecules whose similarity score is high while the difference in the retention index is large. By excluding those pairs (only about 1000 compounds) from the NIST 17 data set, the MAE of the validation error decreases significantly to about 32, which is close to what we report in Table 4. This suggests that those pairs in the NIST 17 library may be problematic. Indeed we found that most of those pairs were removed in the newer NIST 20 library, and we achieved a better performance with the new library.

During the training of the normal boiling point model, we noticed that predictions for certain molecules consistently exhibited large errors, even when they were in the training set. We took note of these and

Table 5
Evaluation of the performance of the model of Stein et al. [41], the model of Matyushin et al. [45], and the model described in this article for selected chemical functionalities for the (unitless) Kovats retention index. The column labeled *N* indicates the number of data items used in computing the mean (μ) and sample standard deviation (s) of the absolute error, $|\epsilon| = |\mathrm{RI}_{\mathrm{experiment}} - \mathrm{RI}_{\mathrm{predicted}}|$.

Molecule	*N*	Stein		Matyushin		Present Work	
type		μ	s	μ	s	μ	s
ether	55,755	112.60	120.21	34.66	52.16	22.99	43.87
amide	24,033	119.26	112.20	38.75	56.73	25.96	47.88
contains O	92,136	114.36	119.88	38.03	56.90	26.65	49.64
hydrocarbon	2010	61.67	94.10	33.27	49.57	27.77	46.75
aromatic	70,501	121.93	119.82	42.48	59.36	29.68	50.95
has a ring	79,091	125.44	125.90	43.37	60.77	30.80	53.19
contains N	51,536	124.71	119.04	46.78	64.90	33.51	56.98
aldehyde	1176	114.23	121.06	45.89	55.80	38.66	49.55
contains S	9707	127.60	120.98	52.85	71.25	39.87	64.53
N heterocycle	22,382	155.50	138.27	61.12	75.80	46.49	67.99
ketone	5611	168.16	172.95	60.98	80.86	46.94	72.89
O heterocycle	9488	182.43	173.88	63.13	73.99	48.28	66.49
alcohol	7102	141.22	153.86	64.01	80.56	52.26	75.42
carboxylic acid	1444	101.78	123.98	63.44	95.51	56.09	94.95
contains P	35	131.00	172.38	66.18	79.00	66.31	76.85
contains P	994			39.64	79.09	31.70	73.73
all compounds	102,761	113.84	119.44	38.97	57.08	27.74	50.00

Source: Adapted from ref. [37] with permission. Copyright 2021 Elsevier.

checked the source literature, and identified a number of literature and data abstraction errors (on the order of 100 data points) in the input data set. Once the errors were either corrected or removed from the data set, we achieved a significant improvement in the model accuracy, as expected.

In both cases, the (potentially) problematic data are only a few percent of the total data, yet they produce a large impact on the performance of our model. Since errors in the data set are essentially unavoidable, strategies are needed to reduce the impact of "bad" data. As suggested in our boiling point model, machine learning could be an effective approach to detect errors in the data set. Developing models that are less likely to be influenced by occasional errors in the training data could be another solution.

5. Conclusion

In summary, we present the application of GNN to three high-quality experimental data sets curated by NIST, namely the Kováts retention indices, the mass spectrum library, and the normal boiling point data from the NIST Thermodynamics Research Center (TRC) SOURCE Data Archival System. For retention indices, we have shown that the GNN methodology significantly outperforms the classical group increment model in terms of predictive accuracy. It also outperforms the CNN model, another popular deep learning methodology, due to the more expressive representation used in GNN for molecules. We have demonstrated that the quality of data is a key to an accurate predictive ML model, as shown in the retention index and boiling point models. For mass spectrum, by using symmetrized version of the Kullback-Leibler function, we presented a GNN model that achieves a modest improvement over a previous model developed by the Wei et al.[48] The performance of the model developed in this work appears to be limited by the nature of the distance metric employed in training. Nevertheless, the results produced by the present model are of value as demonstrated by their performance in various measures including the recall@10 metric.

The success of these GNN models demonstrates the capability of ML methodologies to predict molecular properties, when an appropriate molecular representation (molecular graph in this work) is chosen, and when a sufficiently large, high-quality data set is available. However, high-quality data are usually scarce in physical sciences, so developing ML techniques that perform well on limited data would clearly be an important subject in the near future. Another direction for future exploration is to use 3D structural information (distances, angles, dihedral angles) and equivariant GNNs.

Disclaimer

Certain equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement of any product or service by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

CRediT authorship contribution statement

Thomas Allison: Methodology, Software, Writing – review & editing. **Walid Keyrouz:** Writing – review & editing. **Anthony Kearsley:** Supervision, Writing – review & editing. **Chen Qu:** Writing – original draft. **Barry Schneider:** Project administration, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors gratefully acknowledge the assistance of Dr. Chris Muzny, Dr. Demian Riccardi, and Dr. Eugene Paulechka, all of the NIST Thermodynamics Research Center for providing access to their data set and for assistance in checking and correcting errors in the data. The authors gratefully acknowledge the NIST Mass Spectral Data Center for providing the data used in this study.

References

[1] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489, https://doi.org/10.1038/nature16961.

[2] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, S. Zhao, Applications of machine learning in

drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477, https://doi.org/10.1038/s41573-019-0024-5.
[3] S. Ekins, A.C. Puhl, K.M. Zorn, T.R. Lane, D.P. Russo, J.J. Klein, A.J. Hickey, A. M. Clark, Exploiting machine learning for end-to-end drug discovery and development, Nat. Mater. 18 (5) (2019) 435–441, https://doi.org/10.1038/s41563-019-0338-z.
[4] B.P. MacLeod, F.G.L. Parlane, T.D. Morrissey, F. Häse, L.M. Roch, K.E. Dettelbach, R. Moreira, L.P.E. Yunker, M.B. Rooney, J.R. Deeth, V. Lai, G.J. Ng, H. Situ, R. H. Zhang, M.S. Elliott, T.H. Haley, D.J. Dvorak, A. Aspuru-Guzik, J.E. Hein, C. P. Berlinguette, Self-driving laboratory for accelerated discovery of thin-film materials, eaaz8867, Sci. Adv. 6 (20) (2020), https://doi.org/10.1126/sciadv.aaz8867.
[5] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, J. Hu, Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials, npj Comput. Mater. 6 (1) (2020) 84, https://doi.org/10.1038/s41524-020-00352-0.
[6] R.-R. Griffiths, J.M. Hernández-Lobato, Constrained Bayesian optimization for automatic chemical design using variational autoencoders, Chem. Sci. 11 (2020) 577–586, https://doi.org/10.1039/C9SC04026A.
[7] D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis, J. Zrimec, S. Poviloniene, A. Laurynenas, S. Viknander, W. Abuajwa, O. Savolainen, R. Meskys, M.K.M. Engqvist, A. Zelezniak, Expanding functional protein sequence spaces using generative adversarial networks, Nat. Mach. Intell. 3 (4) (2021) 324–333, https://doi.org/10.1038/s42256-021-00310-5.
[8] A. Nandy, C. Duan, M.G. Taylor, F. Liu, A.H. Steeves, H.J. Kulik, Computational discovery of transition-metal complexes: from high-throughput screening to machine learning, Chem. Rev. 121 (16) (2021) 9927–10000, https://doi.org/10.1021/acs.chemrev.1c00347.
[9] M. Pandey, M. Fernandez, F. Gentile, O. Isayev, A. Tropsha, A.C. Stern, A. Cherkasov, The transformational role of GPU computing and deep learning in drug discovery, Nat. Mach. Intell. 4 (3) (2022) 211–221, https://doi.org/10.1038/s42256-022-00463-x.
[10] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (7698) (2018) 604–610, https://doi.org/10.1038/nature25978.
[11] C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, K.F. Jensen, Prediction of organic reaction outcomes using machine learning, ACS Cent. Sci. 3 (5) (2017) 434–443, https://doi.org/10.1021/acscentsci.7b00064.
[12] M. Meuwly, Machine learning for chemical reactions, Chem. Rev. 121 (16) (2021) 10218–10239, https://doi.org/10.1021/acs.chemrev.1c00033.
[13] B. Zhang, X. Zhang, W. Du, Z. Song, G. Zhang, G. Zhang, Y. Wang, X. Chen, J. Jiang, Y. Luo, Chemistry-informed molecular graph as reaction descriptor for machine-learned retrosynthesis planning, Proc. Natl. Acad. Sci. 119 (41) (2022) e2212711119, https://doi.org/10.1073/pnas.2212711119.
[14] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A. A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.
[15] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Res. 50 (D1) (2021) D439–D444, https://doi.org/10.1093/nar/gkab1061.
[16] J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems, Chem. Rev. 121 (16) (2021) 9816–9872, https://doi.org/10.1021/acs.chemrev.1c00107.
[17] O.T. Unke, S. Chmiela, H.E. Sauceda, M. Gastegger, I. Poltavsky, K.T. Schütt, A. Tkatchenko, K.-R. Müller, Machine learning force fields, Chem. Rev. 121 (16) (2021) 10142–10186, https://doi.org/10.1021/acs.chemrev.0c01111.
[18] J. Behler, Four generations of high-dimensional neural network potentials, Chem. Rev. 121 (16) (2021) 10037–10072, https://doi.org/10.1021/acs.chemrev.0c00868.
[19] H.J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M.A.L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A.P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K.T. Schütt, J. Westermayr, M. Gastegger, R.J. Maurer, B. Kalita, K. Burke, R. Nagai, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Esters, D. Hicks, C. Toher, P.V. Balachandran, I. Tamblyn, S. Whitelam, C. Bellinger, L.M. Ghiringhelli, Roadmap on machine learning in electronic structure, Electron. Struct. 4 (2) (2022) 023004, https://doi.org/10.1088/2516-1075/ac572f.
[20] J.M. Bowman, C. Qu, R. Conte, A. Nandi, P.L. Houston, Q. Yu, δ -machine learned potential energy surfaces and force fields, J. Chem. Theory Comput. 19 (1) (2023) 1–17, https://doi.org/10.1021/acs.jctc.2c01034.
[21] J.S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A.E. Roitberg, Less is more: sampling chemical space with active learning, J. Chem. Phys. 148 (24) (2018) 241733, https://doi.org/10.1063/1.5023802.
[22] N.S. Eyke, W.H. Green, K.F. Jensen, Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening, React. Chem. Eng. 5 (2020) 1963–1972, https://doi.org/10.1039/D0RE00232A.
[23] M. Kulichenko, K. Barros, N. Lubbers, Y.W. Li, R. Messerly, S. Tretiak, J.S. Smith, B. Nebgen, Uncertainty-driven dynamics for active learning of interatomic potentials, Nat. Comput. Sci. 3 (3) (2023) 230–239, https://doi.org/10.1038/s43588-023-00406-5.
[24] H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, Predicting materials properties with little data using shotgun transfer learning, ACS Cent. Sci. 5 (10) (2019) 1717–1730, https://doi.org/10.1021/acscentsci.9b00804.
[25] D. Zhang, S. Xia, Y. Zhang, Accurate prediction of aqueous free solvation energies using 3D atomic feature-based graph neural network with transfer learning, J. Chem. Inf. Model. 62 (8) (2022) 1840–1848, https://doi.org/10.1021/acs.jcim.2c00260.
[26] M. Iman, H.R. Arabnia, K. Rasheed, A review of deep transfer learning and recent advancements, Technologies 11 (2) (2023), https://doi.org/10.3390/technologies11020040.
[27] F. Musil, A. Grisafi, A.P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, Physics-inspired structural representations for molecules and materials, Chem. Rev. 121 (16) (2021) 9759–9815, https://doi.org/10.1021/acs.chemrev.1c00021.
[28] M. Rupp, A. Tkatchenko, K.-R. Müller, O.A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, Phys. Rev. Lett. 108 (2012) 058301, https://doi.org/10.1103/PhysRevLett.108.058301.
[29] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-a. Mller, A. Tkatchenko, Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. Lett. 6 (12) (2015) 2326–2331, https://doi.org/10.1021/acs.jpclett.5b00831.
[30] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. 134 (7) (2011) 074106, https://doi.org/10.1063/1.3553717.
[31] S. De, A.P. Bartók, G. Csányi, M. Ceriotti, Comparing molecules and solids across structural and alchemical space, Phys. Chem. Chem. Phys. 18 (2016) 13754–13769, https://doi.org/10.1039/C6CP00415F.
[32] D. Weininger, SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36, https://doi.org/10.1021/ci00057a005.
[33] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (5) (2010) 742–754, https://doi.org/10.1021/ci100050t.
[34] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2019), https://doi.org/10.1109/TNNLS.2020.2978386.
[35] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: a review of methods and applications, AI Open 1 (2020) 57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.
[36] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572, https://doi.org/10.1021/acs.chemmater.9b01294.
[37] C. Qu, B.I. Schneider, A.J. Kearsley, W. Keyrouz, T.C. Allison, Predicting Kováts retention indices using graph neural networks, J. Chromatogr. A 2021 (1646) 462100, https://doi.org/10.1016/j.chroma.2021.462100.
[38] C. Qu, A.J. Kearsley, B.I. Schneider, W. Keyrouz, T.C. Allison, Graph convolutional neural network applied to the prediction of normal boiling point, J. Mol. Graph. Model. 112 (2022) 108149, https://doi.org/10.1016/j.jmgm.2022.108149.
[39] E. Kováts, Gas-chromatographische charakterisierung organischer verbindungen. teil 1: Retentionsindices aliphatischer halogenide, alkohole, aldehyde und ketone, Helv. Chim. Acta 41 (7) (1958) 1915–1932, https://doi.org/10.1002/hlca.19580410703.
[40] W.P. Eckel, T. Kind, Use of boiling point-Lee retention index correlation for rapid review of gas chromatography-mass spectrometry data, Anal. Chim. Acta 494 (2003) 235–243, https://doi.org/10.1016/j.aca.2003.08.003.
[41] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, Estimation of Kováts retention indices using group contributions, J. Chem. Inf. Model. 47 (2007) 975–980, https://doi.org/10.1021/ci600548y.
[42] A.R. Katritzky, M. Kuanar, S. Slavov, C.D. Hall, M. Karelson, I. Kahn, D.A. Dobchev, Quantitative correlation of physical and chemical properties with chemical structure: utility for prediction, Chem. Rev. 110 (10) (2010) 5714–5789, https://doi.org/10.1021/cr900238d.
[43] J. Yan, J.-H. Huang, M. He, H.-B. Lu, R. Yang, B. Kong, Q.-S. Xu, Y.-Z. Liang, Prediction of retention indices for frequently reported compounds of plant essential oils using multiple linear regression, partial least squares, and support vector machine, J. Sep. Sci. 36 (15) (2013) 2464–2471, https://doi.org/10.1002/jssc.201300254.
[44] A.K. Zhokhov, A.Y. Loskutov, I.V. Rybalćhenko, Methodological approaches to the calculation and prediction of retention indices in capillary gas chromatography, J. Anal. Chem. 73 (3) (2018) 207–220, https://doi.org/10.1134/S1061934818030127.
[45] D.D. Matyushin, A.Y. Sholokhova, A.K. Buryak, A deep convolutional neural network for the estimation of gas chromatographic retention indices, J. Chromatogr. A 2019 (1607) 460395–460398, https://doi.org/10.1016/j.chroma.2019.460395.
[46] S.E. Stein, Chemical substructure identification by mass spectral library searching, J. Am. Soc. Mass. Spectrom. 6 (1995) 644–655, https://doi.org/10.1016/1044-0305(95)00291-K.

[47] S. Stein, Mass spectral reference libraries: an ever-expanding resource for chemical identification, Anal. Chem. 84 (2012) 7274–7282, https://doi.org/10.1021/ac301205z.
[48] J.N. Wei, D. Belanger, R.P. Adams, D. Sculley, Rapid prediction of electron-ionization mass spectrometry using neural networks, ACS Cent. Sci. 5 (4) (2019) 700–708, https://doi.org/10.1021/acscentsci.9b00085.
[49] B. Zhang, J. Zhang, Y. Xia, P. Chen, B. Wang, Prediction of electron ionization mass spectra based on graph convolutional networks, Int. J. Mass Spectrom. 475 (2022) 116817, https://doi.org/10.1016/j.ijms.2022.116817.
[50] R.L. Zhu, E. Jonas, Rapid approximate subset-based spectra prediction for electron ionization-mass spectrometry, Anal. Chem. 95 (5) (2023) 2653–2663, https://doi.org/10.1021/acs.analchem.2c02093.
[51] H. Chang, The myth of the boiling point, Sci. Prog. 91 (3) (2008) 219–240, https://doi.org/10.3184/003685008×360632.
[52] J.C. Dearden, Quantitative structure-property relationships for prediction of boiling point, vapor pressure, and melting point, Environ. Toxicol. Chem. 22 (8) (2003) 1696–1709, https://doi.org/10.1897/01-363.
[53] F. Gharagheizi, S.A. Mirkhani, P. Ilani-Kashkouli, A.H. Mohammadi, D. Ramjugernath, D. Richon, Determination of the normal boiling point of chemical compounds using a quantitative structure-property relationship strategy: application to a very large dataset, Fluid Phase Equil 354 (2013) 250–258, https://doi.org/10.1016/j.fluid.2013.06.034.
[54] D. Cherqaoui, D. Villemin, Use of a neural network to determine the boiling point of alkanes, J. Chem. Soc. Faraday Trans. 90 (1) (1994) 97–102, https://doi.org/10.1039/FT9949000097.
[55] E.S. Goll, P.C. Jurs, Prediction of the normal boiling points of organic compounds from molecular structures with a computational neural network model, J. Chem. Inf. Comput. Sci. 39 (1999) 974–983, https://doi.org/10.1021/ci990071l.
[56] L. Jin, P. Bai, QSPR study on normal boiling point of acyclic oxygen containing organic compounds by radial basis function artificial neural network, Chemom. Intell. Lab 157 (2016) 127–132, https://doi.org/10.1016/j.chemolab.2016.07.007.
[57] L. Jin, P. Bai, Modelling of normal boiling points of hydroxyl compounds by radial basis networks, Mod. Chem. 4 (2) (2016) 24–29, https://doi.org/10.11648/j.mc.20160402.12.
[58] M.R. Fissa, Y. Lahiouel, L. Khaouane, S. Hanini, QSPR estimation models of normal boiling point and relative liquid density of pure hydrocarbons using MLR and MLP-ANN methods, J. Mol. Graph. Model. 87 (2019) 109–120, https://doi.org/10.1016/j.jmgm.2018.11.013.
[59] NIST standard reference database 1A: NIST/EPA/NIH mass spectral library (NIST 20), accessed: September 19, 2022(2020). 10.18434/T4H594, ⟨https://chemdata.NIST.gov/dokuwiki/doku.php?id=chemdata:NISTlibs⟩.
[60] P. Ausloos, C.L. Clifton, S.G. Lias, A.I. Mikaya, S.E. Stein, D.V. Tchekhovskoi, O. Sparkman, V. Zaikin, D. Zhu, The critical evaluation of a comprehensive mass spectral library, J. Am. Chem. Soc. Mass Spectrom. 10 (1999) 287–299, https://doi.org/10.1016/S1044-0305(98)00159-7.
[61] F.E. Grubbs, Sample criteria for testing outlying observations, Ann. Math. Stat. 21 (1) (1950) 27–58, https://doi.org/10.1214/aoms/1177729885.
[62] MEGNet: MatErials Graph Network, accessed: September 19, 2022(2020). ⟨https://github.com/materialsvirtuallab/megnet⟩.
[63] C. Chen, S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. 2 (2022) 718–728, https://doi.org/10.1038/s43588-022-00349-3.
[64] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Comput. Mater. 7 (2021) 185, https://doi.org/10.1038/s41524-021-00650-1.
[65] T. Hsu, T.A. Pham, N. Keilbart, S. Weitzner, J. Chapman, P. Xiao, S.R. Qiu, X. Chen, B.C. Wood, Efficient and interpretable graph network representation for angle-dependent properties applied to optical spectroscopy, npj Comput. Mater. 8 (2022) 151, https://doi.org/10.1038/s41524-022-00841-4.
[66] J. Gasteiger, S. Giri, J.T. Margraf, S. Günnemann, Fast and uncertainty-aware directional message passing for non-equilibrium molecules (2022). arXiv: 2011.14115.
[67] D. Flam-Shepherd, T.C. Wu, P. Friederich, A. Aspuru-Guzik, Neural message passing on high order paths, Mach. Learn.: Sci. Technol. 2 (4) (2021) 045009, https://doi.org/10.1088/2632-2153/abf5b8.
[68] K. Schütt, O. Unke, M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, Proceedings of the 38th International Conference on Machine Learning, PMLR 139 (2021)9377–9388.
[69] J. Brandstetter, R. Hesselink, E. van der Pol, E.J. Bekkers, M. Welling, Geometric and physical quantities improve e(3) equivariant message passing (2022). arXiv: 2110.02905.
[70] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. 13 (2022) 2453, https://doi.org/10.1038/s41467-022-29939-5.
[71] Rdkit: Open-source cheminformatics (2020). ⟨http://www.rdkit.org⟩.
[72] H. Ling, K. Okada, An efficient earth mover's distance algorithm for robust histogram comparison, IEEE Trans. Pattern Anal. Mach. Intell. 29 (5) (2007) 840–853, https://doi.org/10.1109/TPAMI.2007.1058.
[73] S. Majewski, M.A. Ciach, M. Startek, W. Niemyska, B. Miasojedow, A. Gambin, The Wasserstein Distance as a Dissimilarity Measure for Mass Spectra with Application to Spectral Deconvolution, in: L. Parida, E. Ukkonen (Eds.), 18th International Workshop on Algorithms in Bioinformatics (WABI 2018), Vol. 113 of Leibniz International Proceedings in Informatics (LIPIcs), Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, pp. 25:1–25:21, https://doi.org/10.4230/LIPIcs.WABI.2018.25.
[74] C. McGill, M. Forsuelo, Y. Guan, W.H. Green, Predicting infrared spectra with message passing neural networks, J. Chem. Inf. Model. 61 (6) (2021) 2594–2609, https://doi.org/10.1021/acs.jcim.1c00055.
[75] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification, J. Am. Chem. Soc. Mass Spectrom. 5 (1994) 859–866, https://doi.org/10.1016/1044-0305(94)87009-8.
[76] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, 770–778.10.1109/CVPR.2016.90.
[77] United States Environmental Protection Agency, Washington, DC, USA, Estimation Programs Interface Suite, v 4.11 (2023). ⟨https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface⟩.