Graph-Based Deep Learning Models for Thermodynamic Property Prediction: The Interplay between Target Definition, Data Distribution, Featurization, and Model Architecture

Bowen Deng, Thijs Stuyver

Graph-based deep learning models for thermodynamic property prediction: The interplay between target definition, data distribution, featurization, and model architecture

Bowen Deng and Thijs Stuyver*

Ecole Nationale Supérieure de Chimie de Paris, Université PSL, CNRS, Institute of Chemistry for Life and Health Sciences, 75 005 Paris, France

E-mail: thijs.stuyver@chimieparistech.psl.eu

Abstract

In this contribution, we examine the interplay between target definition, data distribution, featurization approaches, and model architectures on graph-based deep learning models for thermodynamic property prediction. Through consideration of five curated datasets, exhibiting diversity in elemental composition, multiplicity, charge state, and size, we examine the impact of each of these factors on model accuracy. We observe that target definition, i.e., using formation instead of atomization energy/enthalpy, is a decisive factor, and so is a careful selection of the featurization approach. Our attempts at directly modifying model architectures result in more modest, though not negligible, accuracy gains. Remarkably, we observe that molecule-level predictions tend to outperform atom-level increment predictions, in contrast to previous findings. Overall,

this work paves the way toward the development of robust graph-based thermodynamic model architectures with more universal capabilities, i.e., architectures that can reach excellent accuracy across data sets and compound domains.

Introduction

Rapid, accurate, and computationally inexpensive estimates of thermodynamic properties are crucial in many applications across the chemical sciences.[1–3] For example, formation (free) energy predictions facilitate straightforward filtering based on thermodynamic driving forces during reaction mechanism/network exploration.[4] Those driving forces can subsequently also be used to estimate kinetic barriers through the empirically observed coupling between reaction and activation energies, also known as the Bell-Evans-Polanyi principle.[5,6]

Empirical, on-the-fly approaches, to estimate energies – be they electronic, enthalpic, or Gibbs-free in nature – without explicitly running quantum chemistry calculations, have a long history. In 1958, Benson and co-workers proposed the first version of group-increment theory.[7] In this theory, molecules are decomposed into predefined atom groups and each of these individual atom groups gets a "group additivity" value assigned, which is initially derived through fitting to formation enthalpies for (a limited) set of experimentally characterized compounds. Several expansions and correction schemes have been proposed to Benson's original approach, gradually improving the accuracy and applicability of this method over the years.[8–11]

With the advent of machine learning (ML) in chemistry, alternative strategies to traditional additivity schemes have become feasible. Deep learning models operating on molecular graphs, i.e., graphs in which nodes correspond to atoms and edges correspond to bonds, have recently shown great promise in producing reasonable thermodynamic predictions.[12] Since these approaches – in contrast to e.g., equivariant neural networks[13] or kernel ridge regression models[14] – do not require a (reliable) geometry as input, they can generate predictions at a millisecond speed even for large molecules spanning hundreds of atoms.[15] Additionally, they

have been demonstrated to yield good predictions for compound classes that have been notoriously difficult to treat with conventional additivity schemes, e.g., polycyclic compounds, which contain ring-strain that cannot easily be decomposed into independent atomic/group contributions.[3,16]

A recent model architecture that is particularly noteworthy in this regard is the one developed by Chen et al.[16] In their approach – conceptually similar to Benson's original additivity ideas – molecular graphs are constructed with an inexpensive, structure-based initial input featurization, after which the graph is passed through a directed message-passing neural network (D-MPNN). The D-MPNN, of which the Chemprop implementation was used,[17,18] iteratively updates the representations of the individual atoms by exchanging information with the neighbors. After a fixed number of iterations, the representation vector for every atom is passed through a feedforward neural network which results in a custom additivity value. These flexible atom-level values, incorporating information about the local molecular environment, are subsequently sum-pooled to yield the final molecule-level prediction (Fig. 1).

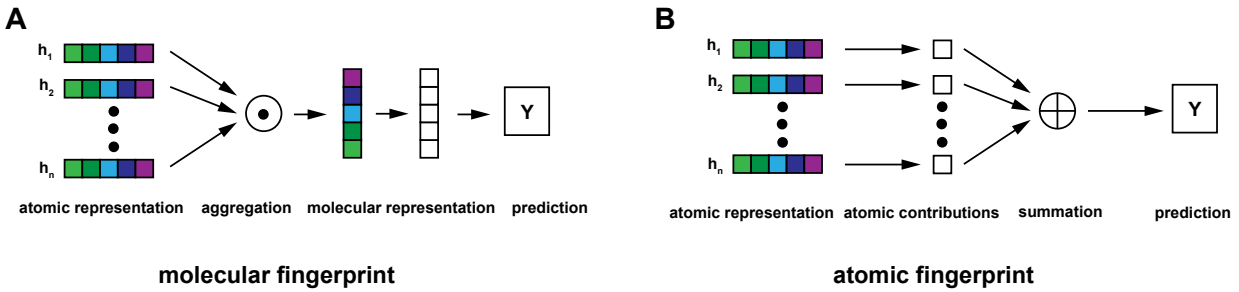


Figure 1: Schematic representations of (a) the molecular fingerprint (FP), (b) the atomic FP architectures. (•) can represent either a mean, sum, or norm operation. (+) indicates sum pooling. In the original Chemprop architecture, learned representations of atomic fingerprints are aggregated to form the molecular FP (cf. panel a), which is then used for the direct prediction of molecular properties. In Chen et al.'s work (cf. panel b),[16] the atomic fingerprints are processed through a fully connected network to determine the atomic additivity values. These values are then summed to derive the overall property of the molecule.

Chen and co-workers demonstrated that this atom-level increment approach extrapolates remarkably well to molecules of varying size, and in a much more robust manner than

conventional graph-based deep learning models, which make their final prediction based on a molecule-level representation (cf. Fig. 1a). In particular, they trained their model on QM9,[19] a dataset consisting of DFT-level properties for 134k neutral, closed-shell molecules with up to 9 heavy atoms (C, O or N), and subsequently refined their model on the small – i.e., only 9,722 molecules with up to 9 heavy atoms – yet high-accuracy, dataset computed at CCSD(T)-F12 level of theory by Green et al.[20] Next, they evaluated their model on the experimental NIST-TRC dataset,[21] consisting of 2524 molecules with up to 42 heavy atoms. With their increment-based model architecture, the experimentally determined formation enthalpies on this last dataset were reproduced within 2 kcal/mol, whereas models making use of molecule-level representations yielded catastrophic errors of tens – and even hundreds – of kcal/mol, as the size of the molecules increased beyond the range encountered in the original training set.

In an independent study, Ward and co-workers observed similar trends when building deep learning-based models for solvation energy prediction: for this property as well, the predictions made by atom-level, increment models turned out to be much more robust to system size variations than those made by models that base their final predictions on molecule-level representations.[22]

At this point, it is important to underscore that the ability to extrapolate towards big molecules is highly desirable from a practical perspective, since accurate electronic structure methods – especially those able to generate reference data for training on par with experimental accuracy – scale unfavorably with system size.[23,24] In practice, calculations for systems exceeding 20-30 atoms tend to quickly become prohibitively expensive, which renders high-throughput dataset generation intractable. Consequently, (accurate) deep-learning training datasets for thermodynamic property prediction will largely remain limited to relatively small molecules for the foreseeable future, while the relevant use cases of thermodynamic models are most certainly not restricted to a specific system size.

Despite the encouraging results by Chen et al.,[16] Ward et al.,[22] as well as others,[3,25] a

universal graph-based predictive model for thermodynamic properties, that can generalize across vast swaths of chemical space remains elusive. In particular, all studies on this topic, up to this point, have focused primarily on QM9-derived datasets, with its peculiar data distribution (e.g., molecules containing small, highly strained rings appear disproportionally represented), its restriction to closed-shell molecules, as well as its limited diversity in terms of elements (only C, N, O, H are present). Consequently, the model architectures, as well as input featurizations developed so far have been tailored and optimized specifically to this type of compounds, and the transferability/expansion to more diverse compound classes has not yet been considered in detail.

In this study, we explore the construction of deep learning models for thermodynamic property prediction that have a more universal applicability. To this end, we curate five distinct open-source thermodynamic datasets, containing closed-shell, radical, and diradical – as well as both charged and uncharged – compounds, in combination with a broad range of elements, i.e., H, C, N, O, F, B, P, S, Cl, Br, and I. We will demonstrate that the models developed so far tend to result in divergent accuracies when retrained on these datasets. Subsequently, several modifications will be explored to make the model performance more accurate and uniform across chemical domains.

Remarkably, we observe that model factors that often receive comparatively little consideration, such as the thermodynamic target definition (e.g., formation vs. atomization enthalpies), and the input featurization, can have outsized effects on the model accuracy, trumping the effect of finetuning the actual neural network architecture in many cases.

Additionally, in contrast to previous findings (*vide supra*), we observe that atom-level increment models generally do not perform on par to molecule-level prediction models; especially when the training and test sets contain a wide range of molecule – and particularly ring – sizes, we observe a decisive advantage for the latter over the former. As such, the conclusions drawn by Chen et al.[16] and Ward et al.[22] do not seem to be universally valid. This highlights an inconvenient truth, namely, that to maximize model accuracy, there ap-

pears to be no way around constructing thermodynamic property datasets that also contain (at least some) diversity in terms of size.

Overall, this work paves the way toward the development of robust graph-based thermodynamic model architectures with more universal capabilities, i.e., architectures that can reach excellent accuracy across data sets and compound domains.

Methodology

Datasets

Five open-source thermodynamic datasets, with molecular graphs in terms of SMILES[26] as input, have been extracted and curated from various sources: QM9, PC9, BDE-db, QMugs and QMugs1.1. Below, a brief description of each of these datasets is provided:

(I) The original version of the QM9 dataset[27] encompasses 134,000 stable, small organic (H, C, N, O) molecules. All properties, including (minimized) energies, enthalpies, and free energies, the corresponding geometries, dipole moments, and polarizabilities were computed using the B3LYP/6-31G(2df,p) level of theory.[28–31]

(II) The PC9 dataset[32] comprises over 99,000 small molecules restricted to containing up to 9 heavy atoms (H, C, N, O, F), and provides calculated total molecular (free) energies/enthalpies at the B3LYP/6-31G(d) level of theory.[28–31] Unlike the QM9 dataset, which includes solely closed-shell neutral compounds, PC9 notably includes species with multiplicities > 1 as well.

(III) The BDE-db dataset[33] contains over 240,000 organic radical species and 40,000 associated closed-shell molecules consisting of H, C, N, and O atoms. Properties such as optimized 3D geometries, enthalpies, Gibbs free energy, vibrational frequencies, Mulliken charges, and spin densities were calculated at the M06-2X/def2-TZVP level of theory.[34,35]

(IV) The QMugs (Quantum-Mechanical Properties of Drug-like Molecules) dataset[36] comprises thermodynamic properties of approximately 660,000 curated molecules, extracted

from the ChEMBL database.[37] Geometry optimizations for these compounds were performed at the xTB level of theory,[38] after which quantum mechanical properties (enthalpies, Gibbs free energies, Mulliken charges, spin densities, etc.) were computed at ωB97X-D/def2-SVP level of theory.[35,39]

(V) The QMugs1.1[40] dataset is an expansion of the original QMugs dataset. It contains almost 72,000 small molecules sourced from a selection of reaction databases, and additional charged and boron-containing compounds were added from ChEMBL.[37] The same level of theory as in the original QMugs dataset was used.[35,39]

Extensive cleaning of these datasets was needed to make them amenable as training data for graph-based deep learning models. An in-depth discussion of the procedure followed can be found in Section S2 of the Supporting Information. After the application of this procedure, final versions of the QM9 (127,007 data points), BDE-db (289,639 data points), PC9 (96,634 data points), QMugs (636,821 data points) and QMugs1.1 (70,546 data points) were obtained, and used throughout this study.

Target definition

Since each of the datasets has originally been computed at different levels of theory, total (free) energy/enthalpy values across the datasets are not directly comparable. Consequently, the raw energetic labels were consistently transformed into thermodynamic quantities, by making use of atom-level reference calculations. As will be discussed below, the choice of the atom reference is not trivial and can affect the accuracy of the resulting model to a non-negligible extent. As such, multiple types of target quantities have been considered.

Atomization enthalpy/energy refers to the amount of energy required to decompose a chemical substance into independent atoms completely, cf. Eq. 1 for the case of the enthalpy,

$$H_a = H_{\text{total}} - \sum_i n_i \times H_i \tag{1}$$

In this equation, H_a represents the atomization enthalpy, H_i is the single atomic enthalpy of the ith atom type (H, C, N, O, etc.), n_i is the number of atoms of the ith atom type that makes up the molecule, and $\mathrm{H_{total}}$ is the total enthalpy computed for the molecule. The reference enthalpies for the atoms are consistently computed at the same level of theory as the total enthalpy; the selected multiplicities for each atom type, and the corresponding enthalpy values, are listed in Table S4 in the Supporting Information.

Formation enthalpy is traditionally defined as the energetic change when one mole of the compound is formed from its constituent elements in their most stable forms under standard conditions.[41] Unfortunately, transforming total enthalpy values into formation enthalpies is a non-trivial task, requiring several parameters for every atom type, which are not readily available for every element in the periodic table, as well as specific energetic quantities at the molecule level (e.g., zero-point energy), which are not accessible for each of the datasets considered in this study.

As such, we adopted a more convenient formation enthalpy/energy definition, where stable, gas-phase molecular allotropes were selected as references, e.g., $\mathrm{H_2}$, $\mathrm{N_2}$, $\mathrm{O_2}$, $\mathrm{F_2}$, $\mathrm{P_4}$ and $\mathrm{C_{60}}$. The advantage of these alternative references is that they can easily be computed with an electronic structure program at any level of theory.[42] The formation enthalpy (or energy) can then be defined as follows,

$$H_f = H_{\mathrm{total}} - \sum_i \frac{n_i}{N_i} \times H(X_N)_i \tag{2}$$

where N_i corresponds to the number of atoms of element i present in the corresponding reference molecule and $H(X_N)_i$ represents the molecular enthalpy of the selected allotrope, with molecular formula X_N, for element i. The reference enthalpy values at every level of theory, used to compute the formation enthalpies in this study, are listed in Table S4 in the Supporting Information.

For QM9, the existing train/validation/test-split was adopted (*vide supra*). For the other datasets, a similar random 80/10/10-split was set. To confirm the robustness of the

conclusions drawn in this study (*vide infra*), alternative data splits were also tested for the smallest (and most challenging) dataset, QMugs1.1, resulting in the recovery of the same trends as those presented below (cf. Supporting Information Section S5). All curated datasets and splits are accessible from `https://doi.org/10.6084/m9.figshare.27262947`.

Hyperparameter selection

For each model tested, only a single set of optimal hyperparameters was determined, and the training set of the QM9 dataset was used to this end. The reason for not finetuning hyperparameters on each individual dataset independently is that one of the goals of this study is to find robust settings that could generalize across compound domains.

Only three hyperparameters were varied in a grid search: activation function (ReLU, PReLU, tanh), number of message-passing steps (n $\in$ {2, 3, 4, 5}) and hidden size (h $\in$ {200, 300, 400}). The results of these hyperparameter searches are summarized in Section S6 of the Supporting Information.

Featurization approaches

We started our work from the original Chemprop model developed by Yang and co-workers[18] in bond message-passing mode,[17] and added the implementation of atom-level fingerprints (FP), in line with the previous work by Chen et al. (cf. Figure 1a).[16] Subsequently, several additional modifications were implemented and tested.

First and foremost, we considered different atom-/bond-featurization approaches. Traditionally, the input featurization for graph-based molecular deep learning models has only been an afterthought, even though they have been demonstrated to influence model performance to a significant extent. In the study by Chen et al.,[16] it was observed that modifying the default input features in the Chemprop model could reduce the mean absolute error on the QM9 dataset by almost 40%. This observation was rationalized by arguing that some of the default features were problematic, e.g., formal charge and bond order, since they are

resonance structure dependent. Furthermore, the authors of this study argued that a more fine-grained encoding of the sizes of the rings in which each atom is involved enables a better description of strain effects.

Here, we extensively evaluate both featurizations, which we denote as "Chemprop" and "Chen et al." respectively. Additionally, we also tested a couple of new featurization approaches. Specifically, compared to the featurization in Ref.,[16] we aimed to refine the ring encoding further. Whereas Chen et al. added a one-hot encoding to the input (atom) representation, indicating whether the considered atom is part of a 3, 4, ..., 8-membered ring, we replaced this part of the representation with a counter, i.e., if an atom is part of two 4-membered rings, the corresponding element in the vector gets assigned a value 2 instead of 1.

Additionally, we introduced the ability to integrate molecule-level input descriptors/features into the model architecture, to facilitate better discrimination between closed-shell, radical, and diradical compounds, as well as between charged (and uncharged) species. To this end, the net molecular charge and multiplicity are determined for the SMILES strings (parallel electron spins are assumed when multiple radical sites are present) and one-hot encoded, after which the resulting vectors are concatenated to the learned atom-level FP coming out of the D-MPNN.

The resulting featurization is denoted "Ours – v1" throughout this manuscript. Additionally, we also considered an expansion of the ring sizes covered. Whereas Chen et al. limited their one-hot encoded ring embedding to ring size 10, we considered in a final featurization – denoted as "Ours – v2" – rings up to size 20. This modification was tested since the QMugs data set,[36] in particular, contains a lot of macrocycles (see Section S2.2 of the Supporting Information). In Table 1, a schematic overview of the different featurizations is provided.

Table 1: Overview of the input features considered in this study.

Type	Feature	Size	Chemprop	Chen et al.	Ours – v1	Ours – v2
Atom	Atom type	100	Yes	Yes	Yes	Yes
	# neighbor bonds	5	Yes	Yes	Yes	Yes
	Formal charge	5	Yes	No	No	No
	Chirality	4	Yes	Yes	Yes	Yes
	# hydrogens	5	Yes	Yes	Yes	Yes
	Hybridization	5	Yes	No	No	No
	Aromaticity	1	Yes	No	No	No
	Atomic mass	1	Yes	Yes	Yes	Yes
	Atom is in ring	1	No	Yes	Yes	Yes
	In $N-$member ring	(3,10)	No	Yes	No	No
	# $N-$member rings	(3,10)	No	No	Yes	No
	# $N-$member rings	(3,20)	No	No	No	Yes
Bond	Bond type	4	Yes	Yes	Yes	Yes
	Is conjugated	1	Yes	Yes	Yes	Yes
	Bond is in ring	1	Yes	Yes	Yes	Yes
	Stereo	6	Yes	Yes	Yes	Yes
	In $N-$member ring	(3,10)	No	Yes	No	No
	# $N-$member rings	(3,10)	No	No	Yes	No
	# $N-$member rings	(3,20)	No	No	No	Yes
Molecule	Charge	5	No	No	Yes	Yes
	Spin multiplicity	3	No	No	Yes	Yes

Model architectures

Finally, we also took a closer look at the neural network architecture itself. In both Chen et al.'s[16] and Ward et al.'s[22] study, a regular (D-)MPNN was used to embed the atoms.[12] Here, we explored in the first instance whether the performance of the embedder can be improved by refining the message-passing mechanism. Significant attention has previously been devoted to the improvement and finetuning of this crucial step in the representation learning process.[17,18,43] Some recent examples include the work by Flam-Shepard and co-workers,[44] as well as by Chen et al.,[45] in which higher-order message passing was considered, with the aim of enabling information diffusion across longer distances within molecular graphs. Here, we introduce a new network module in the D-MPNN, which we call "MLP-Trigonometric" (cf. left-hand side of Figure 1b). This network module is composed of regular multi-layer perceptrons (MLP) and refines the atomic embeddings through advanced skip connections[46–48]

which integrate initial M^0 and subsequent messages M^t. In practice, linear layers F_a and F_b respectively encode (layer-normalized) passed and original messages, after which a sinusoidal interaction term ($\sin\theta$) captures the dynamic relationship between these messages, yielding F_r, followed by a final skip connection to add the subsequent messages M^t,

$$\begin{aligned} F_a &= \text{MLP}(\text{LayerNorm}(m_{vw}^t), W^i) \\ F_b &= \text{MLP}(m_{vw}^0, W^j) \\ F_r &= \text{MLP}(\sigma(F_a \cdot sin(F_b)), W^k) \\ m_{vw}^t &= m_{vw}^t + F_r \end{aligned} \tag{3}$$

where m_{vw}^t is the v, w element of M^t, and m_{vw}^0 is the v, w element of M^0 respectively. We hypothesize that the introduction of these additional skip connections may improve the stability of the message-passing process further and facilitate deeper feature integration.

Additionally, we also considered the integration of a recently developed type of neutral network layer, inspired by the Kolmogorov-Arnold representation theorem,[49] into the message-passing step (cf. right-hand side of Figure 1b). Kolmogorov-Arnold networks (KAN) have recently been proposed as high-performing alternatives to traditional multilayer perceptrons. In the constituent blocks of a KAN, i.e., a "KAN-layer", each node is fully connected to every node in the subsequent layer. For each edge, a separate, learned activation function is applied, rather than simply using weights (cf. the regular MLP-based architecture). Each node then involves a summation operation to all incoming edges.

Learnable activation functions are expressed as weighted combinations of B-splines, where the B-spline basis functions, denoted as B_i, are utilized,

$$\text{KAN-layer}(x) = w_1 \cdot \text{SiLU}(x) + w_2 \cdot \sum_{i=0}^{G+k-1} c_i \cdot B_i(x) \tag{4}$$

The weights w_1, w_2, and the coefficients of the basis function, c_i, are trainable parameters of the spline and x is an individual dimension of an input vector X. The basis function B_i

is selected as a polynomial of degree k, and the grid parameter G controls the complexity of the B-spline construction. Adhering to previous work in this area,[50] the values for k and G were set to 3 and 5 respectively. As activation function, the Sigmoid Linear Unit (SiLU), was selected consistently.

The mathematical expression of the full network module, KAN-Trigonometric, we adopted, can be expressed as follows,

$$\begin{aligned} F_a &= \mathrm{MLP}(\mathrm{LayerNorm}(m_{vw}^t), W^i) \\ K_b &= \mathrm{KAN}(m_{vw}^0) \\ F_r &= \mathrm{MLP}(\sigma(F_a * cos(K_b)), W^j) \\ m_{vw}^t &= m_{vw}^t + F_r \end{aligned} \tag{5}$$

where linear layer F_a and KAN-layer F_b respectively encode (layer-normalized) passed and original messages, F_r involves an interaction between both the outputs from both layers, and A corresponds to a final skip connection.

Results and discussion

Evaluation of the existing model architectures

To set the stage for this study, we start by considering the performance of the original Chemprop model architecture,[18] as well as the analogous architecture with atom-level fingerprints,[16] on each of the curated datasets. Consistent hyperparameters, finetuned on the QM9 dataset, were selected. Both the atomization enthalpy (H_a), and our adopted definition of the formation enthalpy (H_f; *vide supra*) are consistently selected as the target quantities.

Table 2 reveals that the original Chemprop model fails – at times spectacularly – for all selected datasets, except for QM9. Remarkably, a result approaching chemical accuracy is in fact only achieved when the formation enthalpy is used as the target for the latter dataset. This startling preliminary result highlights the limited *out-of-the-box* applicability

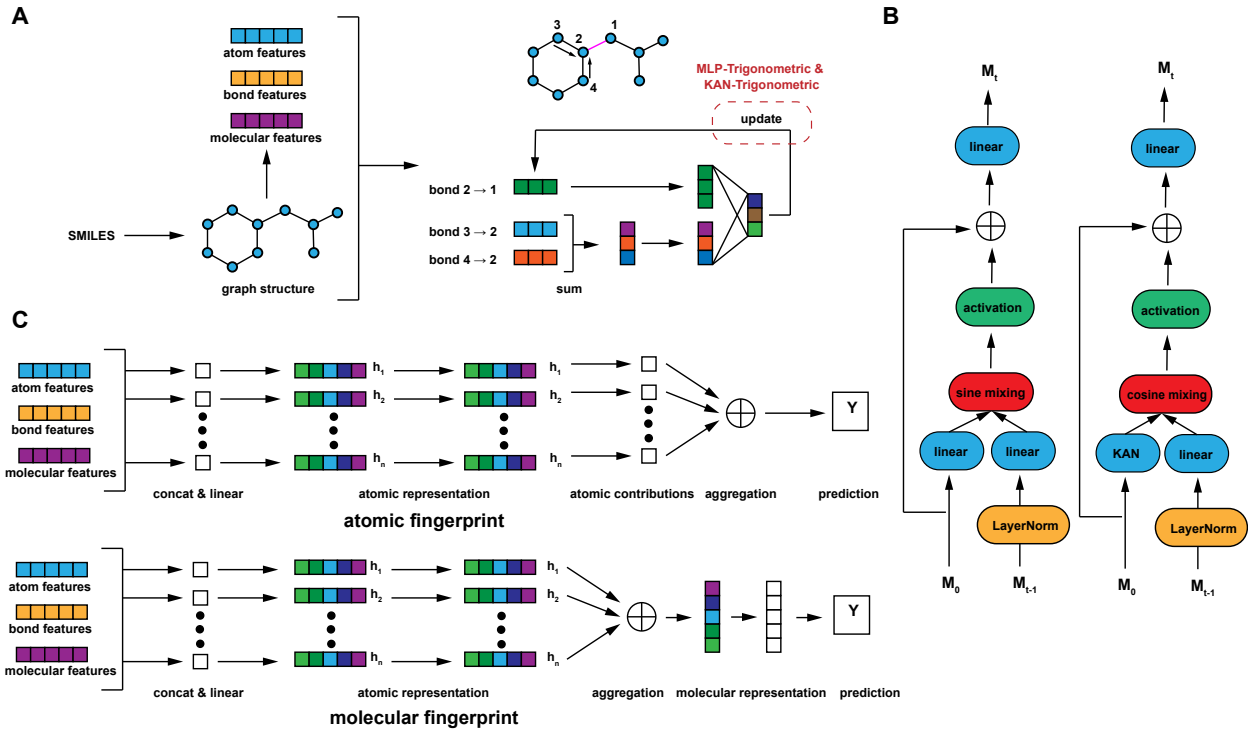


Figure 2: Schematic overview of the architecture used in this work. (a) Construction of the graph and initiation of atom, bond and molecular features (left), followed by bond message passing (right). (b) Modifications introduced to the message passing procedure: MLP-Trigonometric (left) and KAN-Trigonometric (right) block. (c) Schematic overview of the read-out part of the network for the atomic fingerprint (top) and the molecular fingerprint (bottom).

of the currently available graph-based model architectures for thermodynamic property prediction,[17] and it underscores the need for modifications to increase their robustness, which is indeed the focus of the current study.

It should also be noted from Table 2 that switching targets between atomization and formation enthalpies impacts the model accuracy, not only for QM9 but for all datasets. More specifically, one consistently observes significantly higher root mean square errors (RMSE) and mean square errors (MAE) for the atomization enthalpies. This is remarkable since the only difference between both targets is their definition of the atom-level reference. While unexpected at first, a reasonable explanation for this observation can straightforwardly be provided. We will focus on this point in the next subsection, before turning to modifying the actual featurization and model architecture.

Table 2: Performance of the original Chemprop model with optimal hyperparameters determined for the QM9 dataset on the various datasets considered as part of this study. The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Target	Hyperparameters	RMSE (kcal/mol)	MAE (kcal/mol)
QM9	D-MPNN(atomic FP)	Chemprop	H_a	(tanh, 5, 300)	14.93 ± 1.71	5.12 ± 0.39
QM9	D-MPNN(atomic FP)	Chemprop	H_f	(ReLU, 5, 300)	1.93 ± 0.06	1.10 ± 0.04
QM9	D-MPNN(molecular FP)	Chemprop	H_a	(ReLU, 5, 400)	12.98 ± 0.76	2.74 ± 0.15
QM9	D-MPNN(molecular FP)	Chemprop	H_f	(PReLU, 5, 400)	1.91 ± 0.02	1.09 ± 0.06
PC9	D-MPNN(atomic FP)	Chemprop	H_a	(tanh, 5, 300)	36.29 ± 4.04	14.33 ± 2.36
PC9	D-MPNN(atomic FP)	Chemprop	H_f	(ReLU, 5, 300)	5.70 ± 0.39	2.46 ± 0.21
PC9	D-MPNN(molecular FP)	Chemprop	H_a	(ReLU, 5, 400)	30.38 ± 0.94	9.83 ± 2.12
PC9	D-MPNN(molecular FP)	Chemprop	H_f	(PReLU, 5, 400)	5.58 ± 0.05	2.40 ± 0.04
BDE-db	D-MPNN(atomic FP)	Chemprop	H_a	(tanh, 5, 300)	39.76 ± 1.59	13.31 ± 1.26
BDE-db	D-MPNN(atomic FP)	Chemprop	H_f	(ReLU, 5, 300)	3.54 ± 0.06	1.29 ± 0.08
BDE-db	D-MPNN(molecular FP)	Chemprop	H_a	(ReLU, 5, 400)	32.75 ± 0.98	8.53 ± 2.15
BDE-db	D-MPNN(molecular FP)	Chemprop	H_f	(PReLU, 5, 400)	3.53 ± 0.06	1.27 ± 0.03
QMugs	D-MPNN(atomic FP)	Chemprop	H_a	(tanh, 5, 300)	108.84 ± 1.14	62.93 ± 0.30
QMugs	D-MPNN(atomic FP)	Chemprop	H_f	(ReLU, 5, 300)	18.88 ± 0.18	10.37 ± 0.19
QMugs	D-MPNN(molecular FP)	Chemprop	H_a	(ReLU, 5, 400)	31.51 ± 0.82	9.32 ± 0.29
QMugs	D-MPNN(molecular FP)	Chemprop	H_f	(PReLU, 5, 400)	8.70 ± 0.18	3.68 ± 0.04
QMugs1.1	D-MPNN(atomic FP)	Chemprop	H_a	(tanh, 5, 300)	378.38 ± 4.22	211.17 ± 3.98
QMugs1.1	D-MPNN(atomic FP)	Chemprop	H_f	(ReLU, 5, 300)	28.38 ± 0.31	14.08 ± 0.25
QMugs1.1	D-MPNN(molecular FP)	Chemprop	H_a	(ReLU, 5, 400)	233.91 ± 6.89	84.31 ± 2.60
QMugs1.1	D-MPNN(molecular FP)	Chemprop	H_f	(PReLU, 5, 400)	18.74 ± 0.43	7.94 ± 0.10

Why is formation enthalpy an easier target than atomization enthalpy?

To understand why the atom-level energy reference values matter for model training, one needs to appreciate the analogy between the process of reference energy selection and Δ-learning.[51–54] Δ-learning is a powerful technique to improve the accuracy and generalizability of ML models. Instead of training the model directly on the target quantity, in the Δ-learning approach, the model is trained to predict the difference between a – computationally inexpensive – baseline and the full target. Previous studies have highlighted that the effectiveness of the Δ-learning strategy strongly correlates with the accuracy and robustness of the baseline, i.e., accurate baselines typically result in huge gains compared to training a model from scratch, whereas baselines that only correlate mildly with the actual target quantity lead to only negligible results.

Within the context of energy (or enthalpy) predictions, a generic Δ-learning approach

could be defined as follows:

$$E_{\text{total}} = E^0 + \Delta E \tag{6}$$

where E_{total} is the full target quantity, E^0 is the selected baseline, and ΔE is the deviation from the baseline. Comparison of Eq. 6 with Eqs. 1 and 2 readily indicates that the respective energy references can be interpreted as the baseline so that respectively the formation and atomization enthalpy can be identified as the ΔE. Visualizing the distributions of formation and atomization enthalpies in bar plots, one can see that the distribution of the formation enthalpy is much more concentrated, i.e., the former involves a significantly better "baseline" than the latter across the various datasets (cf. Section S2.1 in the Supporting Information).

As such, based on the analogy presented, one can expect a model trained on formation enthalpies to be more accurate and generalizable than one trained on atomization enthalpies. This is perfectly in line with our experimental results. These results underscore that the choice between atomization and formation enthalpy as the target is not a trivial one, but has real consequences for the model performance. In the remainder of this study, we will exclusively focus on formation enthalpies.

Modifying other model factors

Below, modifications to the Chemprop model, other than the target definition, will be introduced one by one, and the corresponding effect on the accuracy will be evaluated. First, we consider the effect of featurization. In Table 3, accuracies for the model architectures both with molecule- and atom-level fingerprints, combined with the alternative featurization from Chen et al., are shown for every curated dataset. Again, consistent hyperparameters, finetuned on the QM9 dataset, were selected.

In line with the previous results by Chen et al.,[16] we observe that this alternative represen-

Table 3: Overview of the regression errors for the D-MPNN models, in combination with the Chen et al. featurization on the various datasets considered as part of this study (formation energy, H_f, has been selected as the target). The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Hyperparameters	RMSE (kcal/mol)	MAE (kcal/mol)
QM9	D-MPNN(atomic FP)	Chen et al.	(PReLU, 4, 400)	1.26 ± 0.01	0.79 ± 0.01
QM9	D-MPNN(molecular FP)	Chen et al.	(ReLU, 4, 400)	1.25 ± 0.01	0.79 ± 0.01
PC9	D-MPNN(atomic FP)	Chen et al.	(PReLU, 4, 400)	4.20 ± 0.08	1.84 ± 0.03
PC9	D-MPNN(molecular FP)	Chen et al.	(ReLU, 4, 400)	4.17 ± 0.09	1.78 ± 0.06
BDE-db	D-MPNN(atomic FP)	Chen et al.	(PReLU, 4, 400)	1.75 ± 0.01	0.89 ± 0.01
BDE-db	D-MPNN(molecular FP)	Chen et al.	(ReLU, 4, 400)	1.68 ± 0.03	0.91 ± 0.02
QMugs	D-MPNN(atomic FP)	Chen et al.	(PReLU, 4, 400)	4.63 ± 0.10	2.67 ± 0.10
QMugs	D-MPNN(molecular FP)	Chen et al.	(ReLU, 4, 400)	3.46 ± 0.74	2.32 ± 0.27
QMugs1.1	D-MPNN(atomic FP)	Chen et al.	(PReLU, 4, 400)	8.79 ± 0.63	4.49 ± 0.54
QMugs1.1	D-MPNN(molecular FP)	Chen et al.	(ReLU, 4, 400)	5.68 ± 0.26	2.59 ± 0.04

tation outperforms the default one in Chemprop by a significant margin across the different datasets. In contrast to their results, however, we observe that for some of the more diverse data sets (QMugs and QMugs1.1), molecular fingerprints significantly outperform atomic ones, indicating that the increment approach may not be universally applicable after all. Furthermore, fairly large errors remain for the PC9, QMugs and QMugs1.1 datasets, particularly in terms of RMSE, which suggests that there are significant outliers in the prediction errors. As such, significant room for improvement remains.

In Table 4, accuracies for the same models, but then with (the first version of) our own featurization are presented (see Section S7 of the Supporting Information for an overview of the effect of only introducing the improved ring embedding, i.e., leaving out the introduction of molecule-level features). Unsurprisingly, for QM9, which contain no (di)radicals nor charged species or macrocycles, the accuracy achieved is almost exactly the same as for the representation by Chen et al.[16] The biggest performance improvement is achieved on the QMugs1.1 dataset: RMSEs are reduced by 1-1.5 kcal/mol. This is a sensible result, as this dataset is arguably the most diverse (it mainly consists of charged species, and it also contains various large, polycyclic compounds). For PC9, which contains only a limited proportion of radical and diradical compounds, and QMugs, which contains macrocycles but no

charged or radical species, the effect of our modified representation is more limited, though we still observe non-negligible gains in terms of RMSE (improvements of 0.2-0.4 kcal/mol).

Somewhat surprisingly, no significant changes in the performance are observed for BDE-db, despite this dataset containing both closed-shell and monoradical species. A plausible explanation for this behavior is that the model easily learns to detect the presence of radicals implicitly when provided with enough training data for both closed-shell and open-shell molecules, compounded with the fact that the compounds in BDE-db tend to be small, which means that 4 or 5 rounds of message-passing ought to be sufficient to disperse this information about the presence of radical centers across (the relevant parts of) the molecules.

Table 4: Overview of the regression errors for the D-MPNN models, in combination with our new featurization, on the various datasets considered as part of this study (formation energy, H_f, has been selected as the target). The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Hyperparameters	RMSE (kcal/mol)	MAE (kcal/mol)
QM9	D-MPNN (atomic FP)	Ours – v1	(ReLU, 4, 300)	1.26 ± 0.02	0.81 ± 0.01
QM9	D-MPNN (molecular FP)	Ours – v1	(PReLU, 4, 400)	1.21 ± 0.01	0.77 ± 0.01
PC9	D-MPNN (atomic FP)	Ours – v1	(ReLU, 4, 300)	4.01 ± 0.08	1.77 ± 0.03
PC9	D-MPNN (molecular FP)	Ours – v1	(PReLU, 4, 400)	4.02 ± 0.06	1.74 ± 0.07
BDE-db	D-MPNN (atomic FP)	Ours – v1	(ReLU, 4, 300)	1.77 ± 0.04	0.97 ± 0.02
BDE-db	D-MPNN (molecular FP)	Ours – v1	(PReLU, 4, 400)	1.63 ± 0.03	0.87 ± 0.02
QMugs	D-MPNN (atomic FP)	Ours – v1	(ReLU, 4, 300)	4.47 ± 0.07	2.61 ± 0.16
QMugs	D-MPNN (molecular FP)	Ours – v1	(PReLU, 4, 400)	3.10 ± 0.01	2.17 ± 0.01
QMugs1.1	D-MPNN (atomic FP)	Ours – v1	(ReLU, 4, 300)	7.22 ± 0.28	3.80 ± 0.12
QMugs1.1	D-MPNN (molecular FP)	Ours – v1	(PReLU, 4, 400)	4.32 ± 0.03	2.39 ± 0.03

Next, we considered the expansion of the ring embedding to ring size 20. Doing so leads – as expected – to a significant improvement in the accuracy of the QMugs and QMugs1.1 datasets when an atomic fingerprint model is used (RMSE reduction of up to 1 kcal/mol), but none of the other models are affected significantly (cf. Table 5).

Finally, we turn to the actual neural network architecture. In Table 6, results are shown for the MLP-Trigonometric version of the D-MPNN (*vide supra*) for every dataset, making use of the first of our own modified featurization (results for the second version follow similar – though slightly subdued – trends and can be found in Section S8 of the Supporting

Table 5: Overview of the regression errors for the D-MPNN models, in combination with the second version of our new featurization, on the various datasets considered as part of this study (formation energy, H_f, has been selected as the target). The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Hyperparameters	RMSE (kcal/mol)	MAE (kcal/mol)
QM9	D-MPNN (atomic FP)	Ours – v2	(ReLU, 4, 300)	1.25 ± 0.01	0.81 ± 0.01
QM9	D-MPNN (molecular FP)	Ours – v2	(PReLU, 4, 400)	1.20 ± 0.01	0.77 ± 0.01
PC9	D-MPNN (atomic FP)	Ours – v2	(ReLU, 4, 300)	3.96 ± 0.04	1.76 ± 0.03
PC9	D-MPNN (molecular FP)	Ours – v2	(PReLU, 4, 400)	3.95 ± 0.14	1.71 ± 0.03
BDE-db	D-MPNN (atomic FP)	Ours – v2	(ReLU, 4, 300)	1.73 ± 0.03	0.95 ± 0.02
BDE-db	D-MPNN (molecular FP)	Ours – v2	(PReLU, 4, 400)	1.63 ± 0.03	0.86 ± 0.01
QMugs	D-MPNN (atomic FP)	Ours – v2	(ReLU, 4, 300)	3.67 ± 0.02	2.54 ± 0.01
QMugs	D-MPNN (molecular FP)	Ours – v2	((PReLU, 4, 400)	3.09 ± 0.01	2.16 ± 0.01
QMugs1.1	D-MPNN (atomic FP)	Ours – v2	(ReLU, 4, 300)	6.73 ± 0.34	3.63 ± 0.08
QMugs1.1	D-MPNN (molecular FP)	Ours – v2	(PReLU, 4, 400)	4.47 ± 0.13	2.47 ± 0.06

Information). Once more, hyperparameters were selected through finetuning on the QM9 dataset.

Table 6: Overview of the regression error on the MLP-Trigonometric model, in combination with the first version of our new featurization, on the various datasets considered as part of this study (formation energy, H_f, has been selected as the target). The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Hyperparameters	RMSE	MAE
QM9	MLP-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	1.18 ± 0.02	0.72 ± 0.01
QM9	MLP-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	1.10 ± 0.01	0.68 ± 0.01
PC9	MLP-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	3.84 ± 0.05	1.64 ± 0.01
PC9	MLP-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	3.83 ± 0.06	1.55 ± 0.03
BDE-db	MLP-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	1.62 ± 0.04	0.84 ± 0.03
BDE-db	MLP-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	1.52 ± 0.02	0.82 ± 0.01
QMugs	MLP-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	4.07 ± 0.02	2.53 ± 0.04
QMugs	MLP-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	3.07 ± 0.02	2.15 ± 0.02
QMugs1.1	MLP-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	6.72 ± 0.25	3.46 ± 0.21
QMugs1.1	MLP-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	4.38 ± 0.15	2.32 ± 0.06

From Table 6, it can be concluded that the introduction of the MLP-Trigonometric module in the D-MPNN tends to improve the performance of every model architecture, though to a varying extent. For QM9, BDE-db and PC9, the improvements are meaningful, yet modest in absolute terms (0.1-0.2 kcal/mol). More significant improvements are observed for the atomic FP model trained on the QMugs and QMugs1.1 datasets (RMSE reduced by

0.4 and 0.5 kcal/mol respectively).

To further assess the effect of introducing the MLP-Trigonometric module in the message-passing network, we also applied the molecular fingerprint version of the resulting model on a (subset of) the benchmarking datasets of MoleculeNet[55] (cf. Section S9 of the Supporting Information). For all datasets considered in this additional benchmarking, our modified model architecture beats the original Chemprop – oftentimes by a significant margin (5-25% reduction in the RMSEs on the regression tasks tested). Furthermore, our finalized model outperforms the vast majority of the model architectures tested in the previous study by Fang et al.[56] – some of which make use of 3D information – on most of the datasets, underscoring the universal validity of the modifications introduced here.

Finally, we also considered the introduction of KAN-layers as a stand-in for conventional feedforward layers. In contrast to some earlier reports, we do not observe any meaningful improvements upon switching between both types of layers. Even worse, most attempts to introduce KAN-layers resulted in a spectacular deterioration of the model accuracy. Only when introducing a KAN-layer in the message-passing module, did we obtain a somewhat comparable result as for MLP-Trigonometric (cf. Table 7). Even then, MLP-Trigonometric retains its advantage for most of the thermodynamic datasets considered.

In Section S8 of the Supporting Information, alternative combinations of featurization and model architectures are presented as well. Similar trends are observed in these cases.

Overall, while the final models considered, cf. Tables 6 and 7, are still not able to uniformly reach chemical accuracy (MAE <1 kcal/mol) across all of the different datasets, we are within striking distance for most of them. What is remarkable is that for the datasets containing a wide range of molecule sizes, i.e., QMugs and QMugs1.1, molecular FP model architectures still outperform atomic FP ones by a significant margin. This suggests that a molecular FP provides a more "holistic" representation of the molecule, in contrast to the inherently local, atom-level alternative. Note that this observation is in direct contradiction to the previous conclusions by Chen et al.[16] and Ward et al.[22] The atomic FPs may still be

Table 7: Comparison of Regression Error on the KAN-Trigonometric model, in combination with our new featurization, on the various datasets considered as part of this study (formation energy, H_f, has been selected as the target). The standard deviations were determined based on the RMSEs/MAEs for five replicates.

Dataset	Model	Featurization	Hyperparameters	RMSE	MAE
QM9	KAN-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	1.14 ± 0.01	0.70 ± 0.01
QM9	KAN-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	1.11 ± 0.01	0.69 ± 0.01
PC9	KAN-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	3.90 ± 0.09	1.64 ± 0.02
PC9	KAN-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	3.88 ± 0.04	1.54 ± 0.03
BDE-db	KAN-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	1.61 ± 0.04	0.83 ± 0.02
BDE-db	KAN-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	1.56 ± 0.02	0.81 ± 0.01
QMugs	KAN-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	4.35 ± 0.10	2.64 ± 0.04
QMugs	KAN-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	3.09 ± 0.02	2.17 ± 0.02
QMugs1.1	KAN-Trigonometric (atomic FP)	Ours – v1	(PReLU, 4, 400)	7.51 ± 0.52	4.03 ± 0.37
QMugs1.1	KAN-Trigonometric (molecular FP)	Ours – v1	(PReLU, 5, 400)	4.33 ± 0.11	2.33 ± 0.02

better at extrapolating to molecule sizes unseen during training, but molecular FPs appear to have a clear edge when it comes to interpolation.

Conclusions

In this study, we have taken a closer look at graph-based deep learning models for thermodynamic property prediction. At the onset, we demonstrated that the Chemprop model with its default settings does not perform well on most thermodynamic datasets other than QM9. Subsequently, we aimed to identify some of the reasons for the limited success of Chemprop. First and foremost, we identified target definition as a major factor impacting the accuracy of the model: formation energies/enthalpies are a much easier target to learn than atomization energies/enthalpies, despite the comparable cost of generating both labels from raw computed energy values. Secondly, we identified the input featurization as being impactful. Adopting the input featurization introduced by Chen et al.[16] markedly improves model accuracy, and additional gains can be made by refining the ring encoding and introducing molecule-level features. Finally, some additional gains can be made by finetuning the model architecture, cf. the developed MLP-/KAN-Trigonometric modules to enhance the message-passing mechanism. These architecture-based gains are however relatively modest

compared to the aforementioned factors, underscoring the necessity to consider the model in its entirety when predicting thermodynamic properties.

Remarkably, we find that for all datasets, the molecule-level fingerprint outperforms the atom-level one. For the datasets that only contain compounds of a uniform length, i.e., QM9, PC9, and BDE-db, the performance improvements of the former compared to the latter are small to negligible, but for QMugs and QMugs1.1, these become significant. This suggests that while atom-level, increment model architectures may be more capable of extrapolating to large molecules, the molecule-level model architectures result in superior performance upon interpolation across molecule sizes.

Overall, our final models (with molecular FPs) reach MAEs on the order of 1-2 kcal/mol and RMSEs of 1-4 kcal/mol across all datasets. Note that this is an accuracy that is in line with the intrinsic accuracy of DFT calculations.[23] As our graph-based models can generate predictions on a millisecond scale (in contrast to geometry-based ones),[13–15] they are particularly appealing as rapid filtering tools in high-throughput (reaction) screening applications.

Acknowledgement

TS acknowledges the French National Agency for Research (ANR) for a CPJ grant (ANR-22-CPJ1-0093-01), BD thanks the National Center for Scientific Research (CNRS) for generous financial support. This work was granted access to the HPC resources of IDRIS under the allocation 2023-100732 granted by GENCI.

Data and Software Availability

All code described in this paper, is available under the open-source MIT License on GitHub, `https://github.com/chimie-paristech-CTM/thermo_GNN`. The (filtered/cleaned) datasets can be downloaded at `https://doi.org/10.6084/m9.figshare.27262947`

Author contributions

BD: software, calculations, formal analysis, writing. TS: conceptualization, methodology, formal analysis, writing, supervision, funding acquisition.

Conflicts of interest

None

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.xxxx.

- Additional information as noted in the text, including message passing neural network, data processing and analysis, reference energies and enthalpies, tests on alternative data splits for selected datasets, hyperparameter settings, ablation study of featurization approaches, and additional analysis of MLP-/KAN-Trigonometric model performance) (PDF).

`supporting_information.pdf` (1.5 MB)

References

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(2) Rupp, M.; Von Lilienfeld, O. A.; Burke, K. Guest editorial: Special topic on data-enabled theoretical chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.

(3) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate thermochemistry with small data

sets: A bond additivity correction and transfer learning approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(4) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(5) Bell, R. P. The theory of reactions involving proton transfers. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **1936**, *154*, 414–429.

(6) Evans, M.; Polanyi, M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans. Faraday Soc.* **1936**, *32*, 1333–1360.

(7) Benson, S. W.; Buss, J. H. Additivity rules for the estimation of molecular properties. Thermodynamic properties. *J. Chem. Phys.* **1958**, *29*, 546–572.

(8) Joback, K. G.; Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **1987**, *57*, 233–243.

(9) Constantinou, L.; Gani, R. New group contribution method for estimating properties of pure compounds. *AIChE J.* **1994**, *40*, 1697–1710.

(10) Zhao, Q.; Savoie, B. M. Self-consistent component increment theory for predicting enthalpy of formation. *J. Chem. Inf. Model.* **2020**, *60*, 2199–2207.

(11) Zhao, Q.; Iovanac, N. C.; Savoie, B. M. Transferable Ring Corrections for Predicting Enthalpy of Formation of Cyclic Compounds. *J. Chem. Inf. Model.* **2021**, *61*, 2798–2805.

(12) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. International conference on machine learning. 2017; pp 1263–1272.

(13) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.

(14) van Gerwen, P.; Fabrizio, A.; Wodrich, M. D.; Corminboeuf, C. Physics-based representations for machine learning properties of chemical reactions. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045005.

(15) Spiekermann, K. A.; Stuyver, T.; Pattanaik, L.; Green, W. H. Comment on 'physics-based representations for machine learning properties of chemical reactions'. *Mach. Learn.: Sci. Technol.* **2023**, *4*, 048001.

(16) Chen, L.-Y.; Hsu, T.-W.; Hsiung, T.-C.; Li, Y.-P. Deep learning-based increment theory for formation enthalpy predictions. *J. Phys. Chem. A* **2022**, *126*, 7548–7556.

(17) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: a machine learning package for chemical property prediction. *J. Chem. Inf. Model.* **2023**, *64*, 9–17.

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

(19) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 1–7.

(20) Grambow, C. A.; Li, Y.-P.; Green, W. H. Accurate thermochemistry with small data sets: A bond additivity correction and transfer learning approach. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(21) NIST Thermodynamics Research Center NIST/TRC Table Database. CD-ROM, 2004.

(22) Ward, L.; Dandu, N.; Blaiszik, B.; Narayanan, B.; Assary, R. S.; Redfern, P. C.; Foster, I.; Curtiss, L. A. Graph-based approaches for predicting solvation energy in multiple solvents: open datasets and machine learning models. *J. Phys. Chem. A* **2021**, *125*, 5990–5998.

(23) Casetti, N.; Alfonso-Ramos, J. E.; Coley, C. W.; Stuyver, T. Combining molecular quantum mechanical modeling and machine learning for accelerated reaction screening and discovery. *Chem. Eur. J.* **2023**, *29*, e202301957.

(24) Folmsbee, D.; Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *Int. J. Quant. Chem.* **2021**, *121*, e26381.

(25) Pinheiro, G. A.; Mucelini, J.; Soares, M. D.; Prati, R. C.; Da Silva, J. L.; Quiles, M. G. Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset. *J. Phys. Chem. A* **2020**, *124*, 9854–9866.

(26) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput.* **1988**, *28*, 31–36.

(27) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 1–7.

(28) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *J. Chem. Phys.* **1993**, *98*, 1372–1377.

(29) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(30) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(31) Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. 6-31G* basis set for atoms K through Zn. *J. Chem. Phys.* **1998**, *109*, 1223–1229.

(32) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminform.* **2019**, *11*, 1–15.

(33) St. John, P. C.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci. Data* **2020**, *7*, 244.

(34) Zhao, Y.; Truhlar, D. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(35) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(36) Isert, C.; Atz, K.; Jiménez-Luna, J.; Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci. Data* **2022**, *9*, 273.

(37) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; others ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(38) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(39) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.

(40) Neeser, R. M.; Isert, C.; Stuyver, T.; Schneider, G.; Coley, C. W. QMugs 1.1: Quantum mechanical properties of organic compounds commonly encountered in reactivity datasets. *Chem. Data Collect.* **2023**, *46*, 101040.

(41) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

(42) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; others Gaussian 16 Software. Revision C. 01, Gaussian, Inc., Wallingford: CT, USA; 2016.

(43) Azabou, M.; Ganesh, V.; Thakoor, S.; Lin, C.-H.; Sathidevi, L.; Liu, R.; Valko, M.; Veličković, P.; Dyer, E. L. Half-Hop: A graph upsampling approach for slowing down message passing. International Conference on Machine Learning. 2023; pp 1341–1360.

(44) Flam-Shepherd, D.; Wu, T. C.; Friederich, P.; Aspuru-Guzik, A. Neural message passing on high order paths. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 045009.

(45) Chen, J.; Schwaller, P. Molecular hypergraph neural networks. *J. Chem. Phys.* **2024**, *160*.

(46) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 770–778.

(47) Xu, K.; Zhang, M.; Jegelka, S.; Kawaguchi, K. Optimization of graph neural networks: Implicit acceleration by skip connections and more depth. International Conference on Machine Learning. 2021; pp 11592–11602.

(48) Li, H.; Xu, Z.; Taylor, G.; Studer, C.; Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems* **2018**, *31*.

(49) Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljačić, M.; Hou, T. Y.; Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756* **2024**,

(50) Blealtan efficient-kan. 2024; `https://github.com/Blealtan/efficient-kan`, GitHub repository.

(51) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(52) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; Von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*.

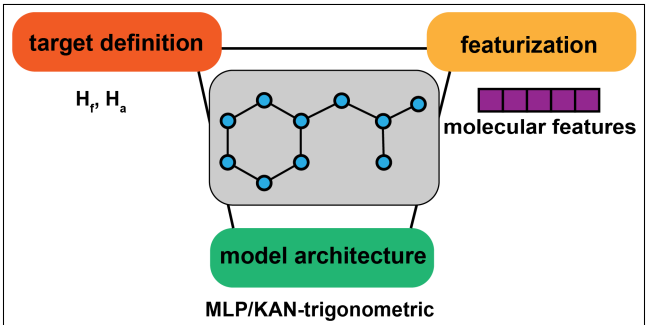
(53) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.

(54) Rakotonirina, V. D.; Bragato, M.; Heinen, S.; von Lilienfeld, O. A. Combining Hammett σ constants for Δ-machine learning and catalyst discovery. *Digit. Discov.* **2024**, –.

(55) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.

(56) Fang, X.; Liu, L.; Lei, J.; He, D.; Zhang, S.; Zhou, J.; Wang, F.; Wu, H.; Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.* **2022**, *4*, 127–134.

TOC Graphic



Graph-based deep learning models for thermodynamic property prediction: The interplay between target definition, data distribution, featurization, and model architecture