Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials

Peter Bjørn Jørgensen
Department of Applied Mathematics and Computer Science
Technical University of Denmark
pbjo@dtu.dk

Karsten Wedel Jacobsen
Department of Physics
Technical University of Denmark
kwj@fysik.dtu.dk

Mikkel N. Schmidt
Department of Applied Mathematics and Computer Science
Technical University of Denmark
mnsc@dtu.dk

Abstract

Neural message passing on molecular graphs is one of the most promising methods for predicting formation energy and other properties of molecules and materials. In this work we extend the neural message passing model with an edge update network which allows the information exchanged between atoms to depend on the hidden state of the receiving atom. We benchmark the proposed model on three publicly available datasets (QM9, The Materials Project and OQMD) and show that the proposed model yields superior prediction of formation energies and other properties on all three datasets in comparison with the best published results. Furthermore we investigate different methods for constructing the graph used to represent crystalline structures and we find that using a graph based on K-nearest neighbors achieves better prediction accuracy than using maximum distance cutoff or the Voronoi tessellation graph.

1   Introduction

The current workhorse for screening of new molecules and materials is Density Functional Theory (DFT), but machine learning methods show a potential for trading DFT accuracy for a several orders of magnitude decrease in computation time. Until recently machine learning methods for predicting properties of molecules and materials have been based on hand crafted feature descriptors, such as Coulomb matrix (Rupp et al., 2012), bag-of-bonds (Hansen et al., 2015), fingerprints (e.g. Rogers and Hahn 2010) and histogram of angles (Faber et al., 2017). With the availability of large databases of molecules and materials we are now seeing a shift towards data-driven representation learning as we have seen in the computer vision field.

The graph neural network model was introduced by Gori et al. (2005); Scarselli et al. (2009) and regained attention when Li et al. (2015) expanded the model for different graph problems. In the last few years a number of graph-based models for molecules have been proposed (Duvenaud et al., 2015;

Kearnes et al., 2016; Schütt et al., 2017a,b; Gilmer et al., 2017). These models can all be cast into the framework of message passing on a molecular graph as shown by Gilmer et al. (2017).

The molecular graphs used as input for these models are either the topological graph defined by the chemical bonds (with or without bond lengths) or a fully connected graph where the pairwise distances between all the atoms are used as edge features. The topological approach does not directly apply to crystalline structures because the chemical bonds are less well-defined and the fully connected approach is impossible because the structure is infinite. The first application of neural message passing for materials is SchNet (Schütt et al., 2017c) where a constant cutoff distance is used. As noted by Schütt et al. (2017c) this may lead to "isolated atoms" if the cutoff distance is too small and the computational burden increases with the size of the cutoff. This motivates us to investigate different ways to define the molecular graphs based on nearest neighbor cutoff. Concurrently with the work presented in this paper Xie and Grossman (2018) have also employed neural message passing on graphs based on Voronoi tessellation and K-nearest neighbors, but they do not include a comparison between the methods.

The edge neural network with set-to-set readout function (enn-s2s) (Gilmer et al., 2017) demonstrated state of the art prediction accuracy on the 13 properties of the QM9 (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012) dataset consisting of 134k molecules. The recently proposed SchNet (Schütt et al., 2017b,c) network improves the accuracy on eight out of the twelve properties. In both of these models the information exchanged between the atoms in the message passing scheme depends on the representation of the sending atom and the edge feature on which the message is passed, but is independent of the representation of the receiving atom. Gilmer et al. (2017) proposed the use of a "pair message" network that includes the state of the receiving atom, but its predicting performance is inferior to the enn-s2s network. In this work we propose to extend the SchNet model with an edge update network such that the edge feature depends on the representation of the atoms that the edge connects. This in turn means the information exchanged between the atoms also depends on the receiving atom. Edge update networks was also utilised in the Weave module proposed by Kearnes et al. (2016), but there the edges are forced to be undirected and the prediction accuracy is below that of enn-s2s (Gilmer et al., 2017).

We benchmark the proposed edge update network on QM9 molecules, Materials Project (Jain et al., 2013) and OQMD (Saal et al., 2013; Kirklin et al., 2015) datasets and it shows an improvement over current state of the art results on all three datasets. For crystalline structures we show that it is beneficial to use a K-nearest neighbor graph rather than a graph defined by a constant cutoff distance as used in previous work (Schütt et al., 2017c).

The paper is organised as follows. We present the proposed model within the neural message passing framework in Section 2 and we show how it is different from other models within the same framework in Section 3. We introduce the three datasets in Section 4, which are used for benchmarking the model in Section 5 and we conclude the paper in Section 6.

2   Message Passing Neural Networks

We describe message passing neural networks similarly to (Gilmer et al., 2017) as a model operating on a graph G with vertex features x_v and edge features ε_{vw}. Each vertex has a hidden state h_v^t and each edge has a hidden state e_{vw}^t which are updated in a number of interaction steps T. Vertices are updated using a message function $M_t(\cdot)$ and a state transition function $S_t(\cdot)$

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}^t), \tag{1}$$

$$h_v^{t+1} = S_t\left(h_v^t, m_v^{t+1}\right), \tag{2}$$

where $N(v)$ denotes the neighbourhood of v, i.e. the vertices that have an edge to v. The edges are updated by an edge update function $E_t(\cdot)$ that depends of the previous edge state and the states of the sending and receiving vertices of the current step

$$e_{vw}^{t+1} = E_t\left(h_v^{t+1}, h_w^{t+1}, e_{vw}^t\right). \tag{3}$$

After the T interaction steps a readout function $R(\cdot)$ is applied which maps the set of vertex states into a single entity

$$\hat{y} = R\left(\{h_v^T \in G\}\right). \tag{4}$$

The readout function must be invariant to permutation of the vertex set, which is often achieved via summation over the vertex features. The functions M_t, S_t, E_t and R are all implemented as neural networks with trainable parameters and can be optimised using gradient descent. Some models, for example Deep Tensor Neural Network (Schütt et al., 2017a), use weight sharing across the interaction steps, thus $M_t \equiv M$, $S_t \equiv S$ and $E_t \equiv E$ for all time steps t. A range of graph convolution models (Duvenaud et al., 2015; Li et al., 2015; Battaglia et al., 2016; Kearnes et al., 2016; Schütt et al., 2017a) including Laplacian based models (Bruna et al., 2013) can be cast into this message passing framework as shown by Gilmer et al. (2017).

2.1 SchNet with Edge Update Network

We now describe how our model fits into the message passing framework described above. As vertex input features, x_v, we use the atomic numbers which are translated into an embedding vector for each atomic number as in (Schütt et al., 2017a,b,c). The initial hidden state $h_v^0 = \ell(x_v)$ is thus the result of a lookup function $\ell : \mathbb{Z} \to \mathbb{R}^C$. The hidden state is a representation of the atom and its chemical environment and the idea behind the message passing interaction steps is to refine this representation based on the surrounding atoms and their chemical environments. Since the interaction between atoms depends on the distance between them (Coulomb's law) we use the interatomic distances as initial edge features. We then expand the initial edge feature using a radial basis function. Denoting the distance between atom v and w as d_{vw}, the edge between node v and w in the graph has initial feature vector

$$(\varepsilon_{vw})_k = \exp\left(-\frac{(d_{vw} - (-\mu_{\min} + k\Delta))^2}{\Delta}\right), k = 0 \ldots k_{\max} \tag{5}$$

where $\mu_{\min}$, Δ, and $k_{\max}$ are chosen such that the centers of the functions covers the range of the input features. In all our experiments we set $\mu_{\min} = 0\,\text{Å}, \Delta = 0.1\,\text{Å}, k_{\max} = 150$. This expansion makes it easier for the neural network model to decorrelate input and output similar to how 1-hot-encoding is preferred over integer coding for categorical features.

The role of the hidden edge representation e_{vw}^t is to control how the two connected atoms interact. The idea of using an edge update network is to let the updated atomic hidden states influence this interaction. We use an edge update function at each interaction step implemented as a two layer feed-forward neural network. The input to the network is the concatenation of the edge representation and the hidden states of the receiving and sending vertices. Thus

$$e_{vw}^0 = E_0(h_v^0, h_w^0, \varepsilon_{vw}) = g(W_{E2}^0 g(W_{E1}^0(h_v^0; h_w^0; \varepsilon_{vw}))), \tag{6}$$

and similarly for the subsequent steps

$$e_{vw}^{t+1} = g(W_{E2}^{t+1} g(W_{E1}^{t+1}(h_v^{t+1}; h_w^{t+1}; e_{vw}^t))), \tag{7}$$

where $g(x) = \ln(e^x + 1) - \ln(2)$ is the shifted soft-plus activation function $(\cdot; \cdot)$ denotes vector concatenation and $\{W_{E1}^0, W_{E2}^0, W_{E1}^{t+1}, W_{E2}^{t+1}\}$ are trainable weight matrices. This update makes the network edges directional.

The message function itself is only a function of the sending node and can be written as

$$M_t(h_v^t, h_w^t, e_{vw}^t) = M_t(h_w^t, e_{vw}^t) = (W_1^t h_w^t) \odot g(W_3^t g(W_2^t e_{vw}^t)), \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\{W_1^t, W_2^t, W_3^t\}$ are trainable weight matrices. This message function enables the interpretation of the function on the right hand side of the element-wise multiplication as a filter-generating function (Schütt et al., 2017b,c) $f_g^t(e_{vw}^t) = g(W_3^t g(W_2^t e_{vw}^t))$, a continuous analogous to the filters applied over a discrete domain in a convolutional neural network for image processing. We visualize this function towards the end of this section. The shifted soft-plus activation function is chosen to follow Schütt et al. (2017b,c) and can be seen as an infinitely differentiable alternative to the rectified linear unit (ReLU) activation function.

The state transition function applies a two layer neural network on the sum of incoming messages and adds this to the current hidden state as in Residual Networks (He et al., 2015):

$$S_t\left(h_v^t, m_v^{t+1}\right) = h_v^t + W_5^t g(W_4^t m_v^{t+1}), \tag{9}$$

where $\{W_4^t, W_5^t\}$ are trainable weight matrices. After a number of interaction steps T all the information about the property we want to predict must be contained in the set of hidden node states.
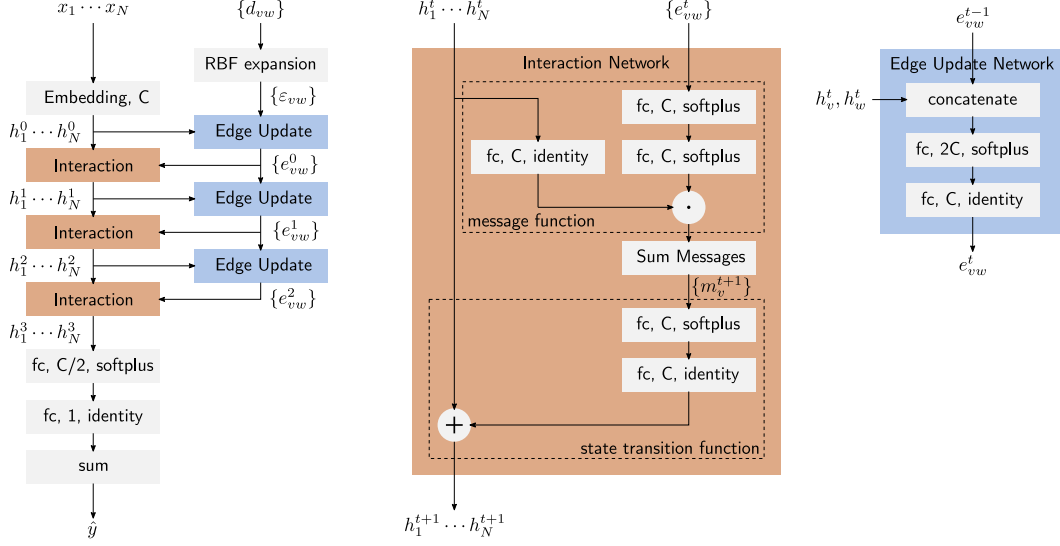
3

Figure 1: Flow of computations of the proposed model. The dimension of the hidden state for each atom state is C and the "fc"-blocks are fully connected layers annotated with the output dimension and the applied activation function.

We apply a readout function for which we use a two layer neural network that maps the hidden representation to a scalar and finally we sum over the contribution from each atom, i.e.

$$R\left(\{h_v^T \in G\}\right) = \sum_{h_v^T \in G} W_7 g(W_6 h_v^T), \tag{10}$$

where $\{W_7, W_6\}$ are also trainable weight matrices. The model architecture and flow of computations is illustrated in Figure 1. The original SchNet architecture is obtained if we "remove" the edge update function, i.e. by setting $E_t(h_v^t, h_w^t, e_{vw}^{t-1}) = e_{vw}^{t-1}$. This property enables direct comparison between the two models.

We want to visualize the filter generating function to qualitatively assess what the model learns, but because of the edge update network, the edge feature of the first layer e_{vw}^0 does not only depend on the distance d_{vw} between the two atoms, but also on the atom embeddings, i.e.

$$f_g^0(d_{vw}) = f_g^0(E_0(h_v^0, h_w^0, \text{RBF}(d_{vw}))), \tag{11}$$

where $\text{RBF}(\cdot)$ is the radial basis function expansion defined in (5). We can now plot the filter response of the first layer as a function of the sending and receiving atoms and the distance between them as shown in Figure 2 and Figure 3. We see that the filter-generating function is almost identical across the sending atom species but the variation across the receiving atoms is more significant. This indicates that the model has learned to shape the messages depending on the receiving atom. The ability to condition the filter on the pair of sending and receiving atoms is the key difference between the proposed model and the SchNet model which is using the same filter for all combinations of sending and receiving atoms.

3 Related Work

Our model is closely related to the other message passing neural network models used for molecular properties prediction. However, the only message passing neural network that has explored the use of edge updates is the Weave module (Kearnes et al., 2016). The message function is $M_t(h_v^t, h_w^t, e_{vw}^t) = \alpha(W_1^t e_{vw}^t)$ and the state transition function is $S_t(h_v^t, m_v^{t+1}) = \alpha(W_3^t(\alpha(W_2^t h_v^t); m_v^{t+1}))$ where $\alpha(\cdot)$ is the rectified linear unit activation function, $(\cdot; \cdot)$ denotes vector concatenation and $\{W_1^t, W_2^t, W_3^t\}$ are weight matrices. Unlike our model where the edge updates are interleaved with the node updates, the edge updates are done in parallel, i.e.

$$\begin{aligned}
e_{vw}^{t+1} &= E_t\left(h_v^t, h_w^t, e_{vw}^t\right) \\
&= \alpha(W_6^t(\alpha(W_4^t(h_v, h_w)) + \alpha(W_4^t(h_w, h_v)), \alpha(W_5^t e_{vw}^t)))
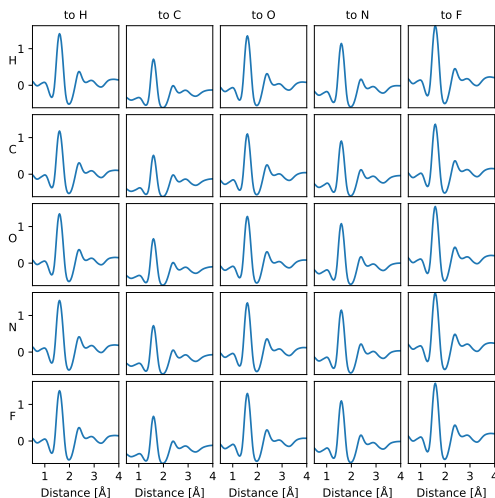\end{aligned} \tag{12}$$

4

Figure 2: An example of one of the learned filters (out of 64) at the first layer of the message passing architecture trained on prediction of formation energy U_0 of the QM9 dataset. The filter depends on the embedding of the sending (the rows) and the receiving (the columns) nodes.
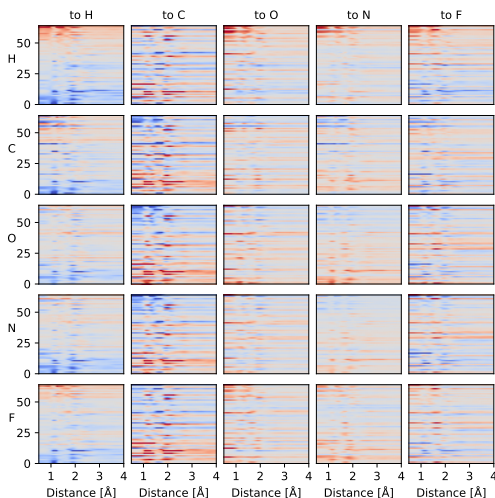

Figure 3: Filter variation for all 64 learned filters at the first layer of the message passing architecture trained on prediction of formation energy U_0 of the QM9 dataset. Each row of pixels corresponds to one of the 64 filters and the color encodes the deviation (-0.4 is blue and 0.4 is red) from the corresponding average filter (averaged over sending and receiving atom). The enumeration of the filters is arbitrary and we have sorted them according to the value of the deviation at 1.0 for the filter H→H. The filter generating function shows a higher dependence on the species of the receiving atom than the species of the sending atom.

Another difference is that the edge updates in the Weave module are invariant to permutation of h_v and h_w and the edges are thus undirected by design. The Weave module does not use the Euclidean distance between the atoms as input features, but uses the graph distance on the chemical graph. The model was tested on a range of different classification tasks as well as drug efficacy, photovoltaic efficiency and solubility regression tasks. The model was reimplemented by Faber et al. (2017) and Gilmer et al. (2017). The version by Faber et al. (2017) includes a few modifications. The edges are no longer forced to be undirected, i.e.

$$
\begin{aligned}
e_{vw}^{t+1} &= E_t\left(h_v^t, h_w^t, e_{vw}^t\right) \\
&= \alpha(W_6^t(\alpha(W_4^t(h_v, h_w)), \alpha(W_5^t e_{vw}^t)))
\end{aligned}
\tag{13}
$$

and the Euclidean distance between the atoms is included in the message function by division of the edge feature with a range of different powers of the distance d_{vw}:

$$
M_t(h_v^t, h_w^t, e_{vw}^t) = \alpha\left(\underset{k\in\{0,1,2,3,6\}}{\text{concatenate}}\left(\frac{W_1^t e_{vw}^t}{d_{vw}^k}\right)\right)
\tag{14}
$$

Both versions of the graph convolution model based on the Weave module shows similar performance on the QM9 regression benchmark (Gilmer et al., 2017; Faber et al., 2017), but the mean absolute prediction error is significantly higher than that of the enn-s2s and SchNet models.

4   Datasets

We benchmark the proposed model on three publicly available datasets.

Quantum Machines 9 (QM9) (Ramakrishnan et al., 2014; Ruddigkeit et al., 2012)   The dataset contains $133\,885$ examples of stable small organic molecules with up to 7 heavy atoms (CONF) and

Table 1: Target properties of the QM9 dataset.

Target	Description
$\varepsilon_{\text{HOMO}}$	Energy of highest occupied molecular orbital (HOMO)
$\varepsilon_{\text{LUMO}}$	Energy of lowest occupied molecular orbital (LUMO)
$\Delta\varepsilon$	Difference between LUMO and HOMO
ZPVE	Zero point vibrational energy
μ	Dipole moment
α	Isotropic polarizability
$\langle R^2 \rangle$	Electronic spatial extent
U_0	Internal energy at $0\,\text{K}$
U	Internal energy at $298.15\,\text{K}$
H	Enthalpy at $298.15\,\text{K}$
G	Free energy at $298.15\,\text{K}$
C_v	Heat capacity at $298.15\,\text{K}$

up to 29 atoms in total including H. The 12 target properties for each example are shown in Table 1. All properties are calculated at the B3LYP/6-31G(2df,p) level of quantum chemistry. Following Gilmer et al. (2017); Schütt et al. (2017c) we randomly select a training set of $110\,000$ molecules, a validation set of $10\,000$ molecules and the remaining $13\,885$ examples are used for testing. The validation set is used for early stopping and model selection.

The Materials Project (Jain et al., 2013) This dataset contains geometries and formation energies of $69\,640$ inorganic compounds with input structures taken from the ICSD database (Bergerhoff et al., 1983). We use the latest version of the database (version 2.0.0). The target property is heat of formation. The number of examples is reduced to $69\,539$ after we exclude all materials with noble gases (He, Ne, Ar, Kr, Xe) because they occur so infrequently in the dataset. This brings the number of different elements in the dataset down to 84. Following (Schütt et al., 2017c) we randomly sample $60\,000$ of the examples to use as the training set. Of the remaining examples we use a set of 5000 for validation and the remaining 4539 examples are used for testing.

Open Quantum Materials Database (OQMD) (Saal et al., 2013; Kirklin et al., 2015) Is also a database of inorganic structures. We have extracted the database from the supplemental material of (Ward et al., 2017). Again we consider materials with noble gases as outliers and we also exclude (highly unstable) materials with a heat of formation of more than $5\,\text{eV/atom}$. We thus exclude 210 out of $435\,792$ examples. As with the materials project we use 20% ($87\,116$) of the examples as a test set, 5000 for validation and the remaining $343\,466$ for training.

5 Results

We train the models with ADAM (Kingma and Ba, 2014) with initial learning rate 5×10^{-4} or 1×10^{-4} (when the higher learning rate leads to instability) and decrease the learning rate by multiplying with 0.96 every $100\,000$ gradient steps. We use a minibatch size of 32, run the optimization for up to 1×10^7 gradient steps, compute the validation error every $50\,000$ steps and terminate if there was no improvement within the last 1×10^6 steps. When applying the model on the two materials datasets (Materials Project and OQMD) we use an average over messages rather than a sum in (1) because Schütt et al. (2017c) found that it increases the stability of the optimization. For these datasets (as well as for some of the QM9 properties, specifically $\varepsilon_{\text{HOMO}}$, $\varepsilon_{\text{LUMO}}$, $\Delta\varepsilon$, ZPVE, μ, α and $\langle R^2 \rangle$) we also take the average rather than the sum in the readout function (10) because the target property is formation energy *per atom*. When the readout function uses a sum we first estimate the target mean and standard deviation atom-wise, i.e. $\hat{\mu} = \sum_i \frac{t_i}{n_i}, \hat{\sigma} = \sqrt{\sum_i (\frac{t_i}{n_i} - \hat{\mu})^2}$ where t_i and n_i are the target and number of atoms for the ith training sample. The targets are then normalised using $\tilde{t}_i = \frac{t_i - n_i \hat{\mu}}{\hat{\sigma}}$ such that each scalar of the sum (10) is expected to have zero mean and unit variance for a given sample with known n_i. When the readout function uses the average rather than a sum the targets are normalised to zero mean and unit variance in the "usual" way, irrespectively of the number of atoms. For the QM9 experiments we use a hidden state representation of dimension $C = 64$ and

Table 2: Mean absolute error of formation energy predictions for V-RF, SchNet and the proposed model. For QM9 the error is in meV and for the Materials Project and OQMD the numbers are in $\mathrm{meV/atom}$. The lowest error is highlighted in bold. We have obtained the V-RF results by running the implementation provided by the authors (Ward et al., 2017), while SchNet results have been obtained by running our own SchNet implementation. The numbers in parenthesis are the estimated 95th percentile, which have been obtained by sampling the test set (with replacement) 1×10^6 times.

Model	QM9 U_0	Mat.Proj.	OQMD
V-RF	-	76.8 (79.8)	74.5 (75.1)
SchNet	13.6 (14.2)	31.8 (33.3)	27.5 (27.9)
Proposed Model	**10.5** (11.1)	**22.7** (24.0)	**14.9** (15.2)

for the two materials datasets we increase the dimensionality to $C = 256$ because the number of different elements in these datasets is 84 rather than 5 in QM9. In agreement with the results shown in (Schütt et al., 2017c) and also noticed in (Gilmer et al., 2017) we did not see a gain in prediction accuracy when using more than 3 interaction steps, so the results presented here are are all with $T = 3$.

5.1 Edge Update Network

To assess the effect of the edge update network we train the proposed model with and without the edge update network to predict formation energies of the three datasets. Without the edge update network the model reduces to the SchNet model. For the two materials datasets (Materials Project and OQMD) we use a cutoff distance of $5\,\mathring{A}$ as used in (Schütt et al., 2017c) when constructing the graphs. In the benchmark we also include a state of the art non-deep learning, graph-based method, which creates a graph based on the Voronoi diagram, extracts a number of features from the graph and uses a random forest regression model for making predictions and we refer to this method as V-RF (Ward et al., 2017). The model also uses the spatial information, but the interatomic distances are normalised such that the model's predictions are independent of the scaling. We benchmark all three models on the same training and test set, but V-RF also uses the validation set for training because early stopping is not used for this model. The mean absolute error for the test set predictions are shown in Table 2. In all three benchmarks we observe a big improvement in prediction accuracy when using the edge update network. This is not only due to the increase in number of parameters, because merely increasing the number of interaction steps does not improve the performance of the model.

We also train the proposed model to predict the 12 properties of the QM9 dataset. The results of SchNet (Schütt et al., 2017c) and enn-s2s (Gilmer et al., 2017) are included as references. The results are shown in Table 3. The proposed model demonstrates a significant improvement in the prediction of 9 out of 12 properties and matches the existing results for the remaining 3.

5.2 Choosing The Cutoff

We want to investigate the importance of choosing a cutoff when constructing the graph used as input to the neural message passing models. The choice of cutoff is important because the computational complexity of the algorithm increases linearly with the number of edges in the graph. On the other hand if the number of edges is too small the interaction between the nodes of the graph may be too limited. In (Schütt et al., 2017c) and in our experiments above a constant cutoff distance of $5\,\mathring{A}$ is used. We use the OQMD dataset for this experiment and use the formation energy as the target for predictions. With this cutoff distance some of the atoms of the dataset becomes isolated and further increasing the cutoff distance comes with prohibitive computational cost. Alternatively we can use a K-nearest cutoff method such that each atom receives messages from the K nearest neighboring atoms no matter how far away they are. This makes the connections between atoms asymmetrical, but that is not necessarily a problem. Finally we can use the neighbors as obtained through Voronoi tessellation, i.e. two atoms are connected if they are neighboring cells in a Voronoi diagram. The connections are symmetrical and we also avoid isolated atoms. We use the software package Voro++ (Rycroft, 2009) to compute the tessellation.

Table 3: Mean absolute error of predictions for different target properties of the QM9 dataset using 110k training examples. The lowest error is highlighted in bold. SchNet and enn-s2s results are from (Schütt et al., 2017c) and (Gilmer et al., 2017) respectively. The numbers in parenthesis are the estimated 95th percentile, which have been obtained by sampling the test set (with replacement) 1×10^6 times.

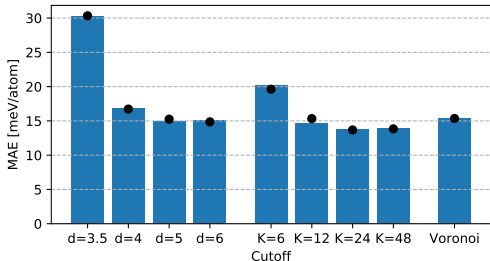| Target	Unit	SchNet	enn-s2s	Proposed	(95th)
$\varepsilon_{\text{HOMO}}$	meV	41	43	**36.7**	(37.3)
$\varepsilon_{\text{LUMO}}$	meV	34	37	**30.8**	(31.3)
$\Delta\varepsilon$	meV	63	69	**58.0**	(58.9)
ZPVE	meV	1.7	**1.5**	1.49	(1.52)
μ	Debye	0.033	0.030	**0.029**	(0.029)
α	Bohr^3	0.235	0.092	**0.077**	(0.082)
$\langle R^2 \rangle$	Bohr^2	**0.073**	0.180	**0.072**	(0.075)
U_0	meV	14	19	**10.5**	(11.1)
U	meV	19	19	**10.6**	(11.2)
H	meV	14	17	**11.3**	(11.9)
G	meV	14	19	**12.2**	(12.7)
C_v	cal/molK	**0.033**	0.040	**0.032**	(0.033)

Figure 4: MAE of formation energy predictions on OQMD using different cutoffs (cutoff distance in Ångström, K-nearest and Voronoi) for constructing the graphs. The black dots show the mean absolute error for the validation set used for early stopping. The lowest error is 13.7 meV/atom using K-nearest cutoff with K=24.
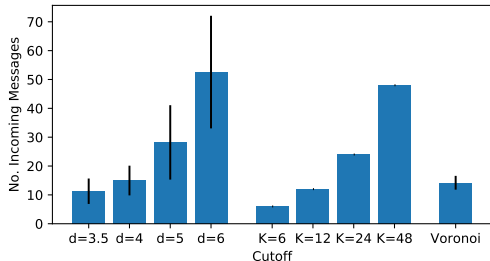

Figure 5: Average number of incoming messages per atom for all atoms in OQMD. The error bars show the standard deviation.

The prediction accuracy of the model when using different cutoff methods is shown in Figure 4. We also show the average number of incoming edges across all atoms in the OQMD dataset in Figure 5, which serves as a proxy for the computational complexity. The results shows that the K-nearest cutoff is more efficient in terms of achieving a low MAE for a given average number of edges per atom and we achieve the lowest error ($13.7\,\text{meV}$) with $K = 24$. This is not only caused by eliminating the "isolated atoms problem", because the Voronoi tessellation method is also without this problem. One reason could be that training the model is more stable when the number of incoming messages is constant.

6   Conclusion

We have proposed a novel neural message passing model for molecules and materials by extending the SchNet model with an edge update network that allows the information exchanged between atoms to be dependent on the sending and receiving atom. This simple extension leads to a remarkable improvement in prediction accuracy. By inspecting the learned filters for the molecular formation energy prediction task we found that the edge updates in the first layer has a higher dependence on the receiving atom than the sending atom. Through numerical simulations we demonstrate improvements in accuracy on formation energy prediction of molecules and materials across all three benchmark

datasets. We also highlight the problem of choosing the cutoff when constructing the graphs used as input for the model. We found that using K-nearest neighbors cutoff yields lower error than using a constant cutoff distance. We believe these results are important for applications of message passing neural networks for predicting properties of molecules and materials and hope to see more applications and further improvements of the model architecture in the future.

References

Battaglia, P. W., Pascanu, R., Lai, M., Rezende, D., and Kavukcuoglu, K. (2016). Interaction networks for learning about objects, relations and physics. 1612.00222.

Bergerhoff, G., Hundt, R., Sievers, R., and Brown, I. D. (1983). The inorganic crystal structure data base. *J. Chem. Inf. Comput. Sci.*, 23(2):66–69.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2013). Spectral networks and locally connected networks on graphs. 1312.6203.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. 1509.09292.

Faber, F. A., Hutchison, L., Huang, B., Gilmer, J., Schoenholz, S. S., Dahl, G. E., Vinyals, O., Kearnes, S., Riley, P. F., and von Lilienfeld, O. A. (2017). Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.*, 13(11):5255–5264.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272.

Gori, M., Monfardini, G., and Scarselli, F. (2005). A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 2, pages 729–734 vol. 2.

Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K.-R., and Tkatchenko, A. (2015). Machine learning predictions of molecular properties: Accurate Many-Body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.*, 6(12):2326–2331.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. 1512.03385.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. a. (2013). The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.*, 30(8):595–608.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. 1412.6980.

Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., and Wolverton, C. (2015). The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1:15010.

Li, Y., Tarlow, D., Brockschmidt, M., and Zemel, R. (2015). Gated graph sequence neural networks. 1511.05493.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data*, 1:140022.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754.

Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.*, 52(11):2864–2875.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108(5):058301.

Rycroft, C. H. (2009). VORO++: a three-dimensional voronoi cell library in c++. *Chaos*, 19(4):041111.

Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials design and discovery with High-Throughput density functional theory: The open quantum materials database (OQMD). *JOM*, 65(11):1501–1509.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. *IEEE Trans. Neural Netw.*, 20(1):61–80.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. (2017a). Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.*, 8:13890.

Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. (2017b). SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. 1706.08566.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. (2017c). SchNet - a deep learning architecture for molecules and materials. 1712.06113.

Ward, L., Liu, R., Krishna, A., Hegde, V. I., Agrawal, A., Choudhary, A., and Wolverton, C. (2017). Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. *Phys. Rev. B Condens. Matter*, 96(2):024104.

Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120(14):145301.