VIMPNN: A PHYSICS INFORMED NEURAL NETWORK FOR ESTIMATING POTENTIAL ENERGIES OF OUT-OF-EQUILIBRIUM SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Simulation of molecular and crystal systems enables insight into interesting chemical properties that benefit processes ranging from drug discovery to material synthesis. However these simulations can be computationally expensive and time consuming despite the approximations through Density Functional Theory (DFT). We propose the Valence Interaction Message Passing Neural Network (VIMPNN) to approximate DFT's ground-state energy calculations. VIMPNN integrates physics prior knowledge such as the existence of different interatomic bounds to estimate more accurate energies. Furthermore, while many previous machine learning methods consider only stable systems, our proposed method is demonstrated on unstable systems at different atomic distances. VIMPNN predictions can be used to determine the stable configurations of systems, i.e. stable distance for atoms – a necessary step for the future simulation of crystal growth for example. Our method is extensively evaluated on a augmented version of the QM9 dataset that includes unstable molecules, as well as a new dataset of infinite- and finite-size crystals, and is compared with the Message Passing Neural Network (MPNN). VIMPNN has comparable accuracy with DFT, while allowing for 5 orders of magnitude in computational speed up compared to DFT simulations, and produces more accurate and informative potential energy curves than MPNN for estimating stable configurations.

1 INTRODUCTION

Chemical simulations have many useful industrial applications ranging from drug discovery to the production of materials for daily use (Schütt et al., 2017). Simulating crystal systems in particular provides useful properties such as surface absorption, chemical reactions, and surface magnetism (Bilek & Skála, 1978). Quantum Mechanical (QM) simulations can be used for the calculation of ground-state energies based upon the interaction of atoms. By simulating a chemical system at different interatomic distances, it is possible to determine a stable configuration where the atomic interaction is at an equilibrium (minimum potential energy). While QM simulations can be less time consuming than physical experimentation, they typically require large amounts of computing resources and do not scale well to larger system sizes (Jiang et al., 2003; Baima et al., 2017).

To address this difficultly in determining the ground-state energy, the Kohn-Sham Density Functional Theory (DFT) may be used to simplify the calculations by considering the electronic density in place of individual electrons. DFT has proved useful in QM due to its good trade-off between speed of computation and chemical accuracy (Cohen et al., 2012). However, as DFT calculations are proportional to the number of interacting electrons (Lanyon et al., 2010) computation for large systems remains intractable. As many interesting and realistic systems are formed from a large number of atoms, a computationally efficient and accurate method for chemical property estimation that scales well to these large systems would prove significantly useful in practical applications (Gomes et al., 2008). Moreover, several simulations are often necessary to identify stable configurations of systems away from local minima. Thus, fast approximation methods are desirable in such scenarios.

Machine learning (ML) has been proposed as an efficient method for classification of (stable) molecules and estimation of chemical properties by learning to reproduce the results of DFT, with recent advancements through a Message Passing Neural Network (MPNN) (Gilmer et al., 2017). However, unstable molecular and crystal structures are often not considered, although this would allow discovering new stable configurations.

In this article we demonstrate the applicability of ML for the energy prediction of unstable molecular and crystal systems. We propose a new DNN framework – the VIMPNN (Valence Interaction Message Passing Neural Network) that considers the different interatomic interactions driven by different valences. It produces comparable accuracy to that of DFT while also improving the computation time by 5 orders of magnitude. We demonstrate that our method also produces more accurate energy estimations than that of MPNN.

ML in quantum chemistry is made possible in part by the availability of large datasets of chemical structures and their (measured or computed) energies in addition to other chemical properties. Online repositories such as Quantum Machine[1] store a collection of datasets resulting from various QM simulation methods for the intention of training predictive models. QM9 (Blum & Reymond, 2009; Montavon et al., 2013) is one such dataset containing 134K molecules using Carbon, Hydrogen, Oxygen, Nitrogen, and Fluorine atoms. However, as all molecules in this dataset are at their stable configuration, it is not suitable for designing and testing accurate energy prediction methods for unstable configurations. We address this shortcoming by augmenting the QM9 dataset with out-of-equilibrium configurations for 10K of its molecules at a [90%, 150%] range of atomic distances. In addition, we create two datasets composed of infinite- (i.e. periodic) and finite-size crystals of Aluminium and Copper atoms.

In summary, the contributions of this work are:

- A new VIMPNN model that accounts for the physics of atomic bonds within a molecule or crystal to improve the accuracy of the chemical properties estimations. This model also introduces auxiliary chemical property estimations to help learn descriptors that are closer to the physics of the problem.
- A new use of ML for accurate ground-state energy estimation of out-of-equilibrium molecule and crystal systems.
- The public release of new infinite- and finite-size crystal datasets, as well as an augmented QM9 dataset.

The remainder of this article is organised as follows: Section 2 discusses the related work concerning various ML approaches for molecular properties estimation. We describe the creation of 3 new datasets through DFT simulations in Section 3. In Section 4 we introduce our new physics informed DNN for accurate energy estimation. We evaluate the VIMPNN and describe the results in Section 5. In Section 6 we give our concluding remarks on the method and results and available future work.

2 Related Work

Several ML methods have been used for estimating a variety of chemical properties from different representations of chemical systems. Li et al. (2015) construct a covariance matrix from Euclidean distances of different atomic configurations, then use Bayesian regression to estimate interatomic forces. Wang et al. (2013) use kernel ridge regression on the numbers of seven coarse *'building blocks'* (such as CH_2) to estimate many chemical properties such as electron affinity and atomization energy. Recent research by Shi et al. (2019) use strain information and a Fourier transform-based representation of crystal lattice for a feed-forward neural network (NN) to estimate an electronic bandgap structure in silicon. The bandgap refers to the energy difference between the valence and conduction bands in insulators and semiconductors, and therefore the energy required to transition between stable excitation states. Silicon crystal in a known equilibrium state is strained in the range of -10% to 10% in each strain component (such as thermal properties) to search the bandgap space for another stable state, for example transforming silicon from a semiconductor to a metal. Therefore, this work uses the NN to describe the bandgap as a function of the strain tensors. In contrast, our method infers the effect of spatial deformations of systems through a different energy estimate.

[1]`http://quantum-machine.org/`

Some approaches focus on the prediction of stable geometries for chemical systems. Timoshenko et al. (2018) use experimental spectra data as input for an NN to estimate the probability distribution of the stable bonding distance between pairs of atoms. In their work, they consider different combinations of Pd and Au atoms (i.e. an NN for Au-Au bonding, another for Pd-Au bonds). When considering crystals made of two different atom types, Takahashi & Takahashi (2019) use 8 quantities associated with each atom type (such as atomic radii, electronegativity, and number of atoms) to perform a random forest classification on how the resulting stable crystal structure would form among a set of 492 different possible structures. The predicted lattice type then requires further optimisation using DFT to find its exact size (i.e. scaling factor). However, their method may also predict metastable configurations (i.e. local minima) rather than stable ones (i.e. global minimum). Whereas the aforementioned methods attempt to directly obtain the stable distance between atoms, we adopt the approach of estimating energies for various geometries with the aim to minimise these energies into stable configurations. While this approach may require more training data and more complex modelling to handle arbitrary configurations, we argue that it has a stronger generalisation potential. It also has the added benefit of allowing the (future) observation of progressive crystal growth.

Previous works on estimating the potential energy of chemical systems include that of Rupp et al. (2012) who used a a feed-forward NN to predict the energies of molecules from the Coulomb matrix (a pairwise matrix that describes the low-level electrostatic interaction between atoms). However, predictions on this matrix are sensitive to permutations of atoms which can result in different molecular property values being estimated for a same crystal. Montavon et al. (2012) provides an invariant solution by training the NN on a set of randomly permuted matrices. To show the benefit of their method, they train different ML models on the QM7 dataset and improve on the accuracy of previous approaches by a factor of 3. A recent advancement by Gilmer et al. (2017) provides a more flexible and accurate representation for molecules as a chemical graph where nodes describe atoms and edges encode the distances and bond type between them. Their Message Passing Neural Network (MPNN) simulates the atomic interactions through the passing of messages between the nodes, and synthesises them within additional hidden nodes. So far, this work has only been evaluated on stable configurations of molecules from the QM9 dataset. We further evaluate it on unstable molecules and crystals. Furthermore, we extend on their approach to allow for more accurate predictions better accounting for the physics of interatomic interactions.

3 Datasets

Many previous ML approaches and associated datasets only consider the case when molecules and crystals are at their stable configuration, therefore learning the interaction of atoms only at an equilibrium state. In performing experimentation on materials, the stable configuration may not be known prior to simulation. Therefore predicting accurately for out-of-equilibrium configurations is necessary, and has to be supported by a dataset of unstable systems. We create 3 datasets to train and benchmark VIMPNN. The first is an augmented QM9 dataset (Section 3.1) to investigate the complexity of atomic interactions from a variety of atom types in small but diverse molecular systems. In the second dataset, we create infinite-size crystals (Section 3.2) for learning regular bonding patterns that arise in periodic structures. Thirdly, a dataset containing finite crystals of an increasing size and complexity to test performance in large scale interactions (Section 3.3).

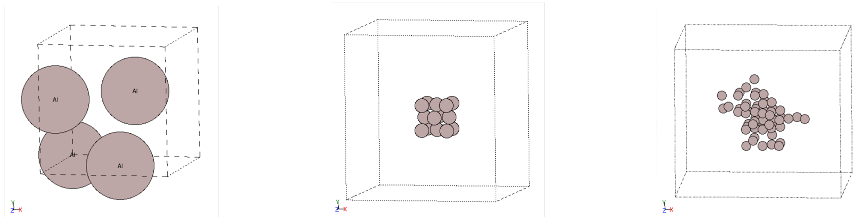
3.1 Augmented QM9 Dataset

An augmented QM9 dataset is created by taking the first 10,000 molecules of QM9 and modifying the interatomic distances at 10 regular intervals between 90 and 150% of the original stable configuration. This dataset therefore contains 100K different systems. At each interval, the ground-state energy is calculated using DFT[2] to serve as the target energy for our supervised learning task

3.2 Infinite Crystal

Learning to estimate the potential energy for an infinitely sized crystal might benefit from the regular pattern in the lattice structure, possibly reducing the complexity of the internal data model the ML algorithm must build. A dataset of infinite-sized crystals, although not necessarily relevant from a

[2]We use CP2K (`https://www.cp2k.org/`) to calculate the ground-state energies with DFT.



(a) Unit-cell of Aluminium atoms for an infinite fcc crystal. (b) Basic fcc pattern in a finite system. (c) Configuration for random atom positions.

Figure 1: Crystal structures for infinite and finite crystal of varying sizes.

physical point of view, is therefore an interesting case study for the design and training of ML based methods. To create the infinite crystal we use the simple Face-Centred Cubic (fcc) Bravais lattice illustrated in Fig. 1a. This lattice structure is created for both Aluminium and Copper atom types. We iteratively compress/dilate the cell size from 90 to 150% of the stable configuration and compute the corresponding ground-state energies using DFT. The change of atomic distances are performed isomorphically (the same change in all spatial axes).

3.3 FINITE CRYSTAL

In contrast to the regular lattice pattern created by an infinite crystal, finite crystals allow for learning over more complex atomic interactions by placing atoms at a random location within the lattice pattern. Such a dataset would enable evaluation of an ML-based method's ability to learn how each atom contributes to the final ground-state energy, rather than learning the general fcc lattice pattern with respect to the interatomic distances.

To generate our dataset, we start with a basic configuration of 14 atoms following the fcc crystal pattern (Fig. 1b). We then iteratively introduce a new atom at a random lattice position into the system (Fig. 1c). Every time a new atom has been added, the distance between all atoms in the system are compressed/dilated at 10 regular intervals from 90 to 150% of the stable distance. Atoms are randomly added until the system reaches a size of 114 atoms, before the process is repeated to create more different crystal structures. We use 20 pseudo-random seeds for the placements of new atoms, therefore creating 20 different variations crystals at each different size (i.e. number of atoms), and 2280 total number of crystal structures. The ground-state energy is again computed at every interval using DFT, which took up to 12 hours per structure in the 114 atom cases.

4 METHODOLOGY

VIMPNN estimates the ground-state energies for molecular or crystal systems through learning the energy as a function of the geometry of bonded atoms. By querying the VIMPNN at various atomic distances, we may find a stable configuration for bonding (see Fig. 2). VIMPNN extends MPNN (Gilmer et al., 2017) that simulates atomic interactions through the passing of messages between pairs of atoms (Section 4.1). However, VIMPNN further accounts for the physics of the system by learning how the different forms of bonding between atoms incur different changes in the nodes' states (Section 4.2), and also reinforces the physical relevance of the analysis and learnt features through estimating auxiliary physical properties of the system (Section 4.3). A representation of the VIMPNN model is depicted in Fig. 2.

4.1 PREREQUISITE: MPNN

MPNN learns to predict ground-state energy from an undirected graph $\mathcal{G}$, where atoms are nodes and edges e_{vw} are feature descriptors such as atomic distance between two nodes v and w. MPNN learns by iterating through three functions for a fixed number of iterations (recommended between 3 and 8): 1) the Message function (M_t in Eq. 1) creates a *message packet* symbolising the action of a neighbouring atom on node v. A final *message* vector m_v^t for node v is created in Eq. 1 to combine the actions of all neighbours. 2) An update function (U_t in Eq. 2) updates the hidden state h_v^t of node

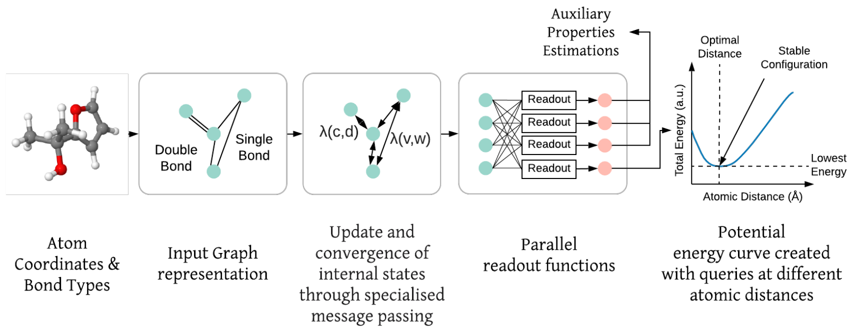


Figure 2: The proposed VIMPNN model to estimate the ground-state energy as a function of the interatomic distance and valence through message passing. Messages exchanged by atoms account for their bond types, and parallel readout functions further pushes the NN towards more physically relevant reasoning through the estimation of low-level physical properties.

v using its previous hidden state h_v^{t-1} and calculated message m_v^t. It is implemented by a Gated Recurrent Unit (GRU). 3) The Readout function (R in Eq. 3) takes the set of hidden state of all nodes at all timesteps and estimates ground-state energy.

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

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

$$y_{energy} = R\left(\left\{h_v^T | v \in \mathcal{G}\right\}\right) \tag{3}$$

Full details on the implementation of these three functions can be found in the original paper (see Gilmer et al. (2017)). Next, we detail how VIMPNN adapts these functions to incorporate physics properties in order to better handle out-of-equilibrium systems.

4.2 Accounting for Bond Types

The type of bonds between different atoms are an important factor when considering their interaction and its contribution to the energy of the system. Chemical bonds between atoms can be characterised by the number of valence electrons exchanged in the process of bonding. For example, ionic bonding requires one donor electron and one acceptor, while in situations where the bonding atoms have similar electronegativity, some electrons would be shared between the bonding atoms resulting in a covalent bond. Moreover, most atomic systems have minimum potential energy (stable bond) at an optimal atom separation (Feinberg & Ruedenberg, 1971).

Therefore, the potential energy of a system may be better estimated by accounting for the contribution of different bond types. Gilmer et al. (2017) acknowledged this fact and introduced bond type information in the edges e_{vw} of the undirected graph representing the chemical system. This information was used together with interatomic distance by the message function M_t in Eq. 1 to estimate the influence of neighbouring atoms on a node. Thus, the NN nodes may exchange different messages depending on their bond type. Gilmer et al. (2017) demonstrated that this approach results in better energy estimates than when using interatomic distance alone.

These encouraging results inspired us to take this physics integration principle further and to explore complementary ways of introducing bond type information into the simulation of atomic interactions. In the present work, we propose to implement different messaging channels based on bond type, with a direct effect on the update of nodes (Eq. 2). Bond type is quickly predetermined based on the atoms' valency and electronegativity using the RDKit software[3]. In practice, the idea of having different messaging channels based on bond type may be expressed in different ways. We experimented with a) having separate messages $m_v|_{BT}$ for each bond type, computed as in Eq. 1 to combine all message packets of a same bond type. The messages $m_v|_{BT}$ would then be simply combined as a weighted sum, where the weights are learnt by the NN to control the influence of each $m_v|_{BT}$. This final

[3]https://www.rdkit.org/

combined message is provided to the GRU of Eq. 2 to compute a node update where different bond types contribute differently. Another possibility is b) concatenating the messages $m_v|_{BT}$ before providing them to the GRU of Eq. 2. Finally, we also experimented with c) having one GRU for each bond type, handling its corresponding message $m_v|_{BT}$, with the updates of all GRUs being summed to compute the final nodes' update. In addition, in the case a), we tried weighting the summed $m_v|_{BT}$ by: i) a vector of coefficients of same size as the message (i.e. element-wise weighting), and ii) a simple scalar. Similarly, in case b), we also experimented with i) letting the GRU handle the concatenated $m_v|_{BT}$ freely, and with ii) imposing a similar handling of the different $m_v|_{BT}$ through using the same GRU parameters, duplicated for each $m_v|_{BT}$ with a simple (learnt) scaling to allow different weightings of the $m_v|_{BT}$ messages in the obtained node update. We found that the best results were obtained in the case a.ii) with combining all messages $m_v|_{BT}$ in a simple way while letting the NN weight the influence of each message $m_v|_{BT}$. It provides a new formulation for the message vector of MPNN:

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \cdot \lambda(v, w) \tag{4}$$

where $\lambda(v, w)$ is the appropriate bond type weight for the pair of nodes v, w.

4.3 Estimating Auxiliary Physical Properties

We hypothesise that, by encouraging the VIMPNN's hidden states to relate more to basic physical properties, we may obtain a more accurate energy estimator that generalises better to new systems. In addition, as the bond type is characterised by the valence property of atoms involved in the bonding process, a more physically relevant hidden state may better support the differentiation of messages and node updates per bond type introduced in Section 4.2. In practice, we encourage this greater relationship with physical parameters through the estimation of auxiliary properties through parallel readout functions. We experimented with: 1) the number of atoms of each type present in the system (Eq. 5), 2) the number of orbitals associated with each atom type (Eq. 6) – a property that is directly relevant to the determination of bond type –, 3) a probability distribution for the scaling to the stable interatomic distance for each bond type, estimated as a Gaussian function (Eq. 7).

$$y_{atoms} = R(\{h_v^T | v \in \mathcal{G}\}) \tag{5}$$

$$y_{orbitals} = R(\{h_v^T | v \in \mathcal{G}\}) \tag{6}$$

$$y_{pdf} = R(\{h_v^T | v \in \mathcal{G}\}) \tag{7}$$

For each of these auxiliary estimations, a mean-squared error loss term (weighted with an $\alpha = 0.3$ hyper-parameter) is minimised during training.

5 Results and Discussion

In this section, the performance impact for each of the physics integration strategies introduced in Section 4 are first measured in turn on the augmented QM9 dataset introduced in Section 3.1 (Section 5.1). The best performing VIMPNN model is then further evaluated on all three datasets (see Section 3), using MPNN as a baseline for comparison (Section 5.2). In particular, we include a test on the generalisation ability of VIMPNN by training it on crystals of up to 25 atoms to predict the energies of up to 75 atom crystals (Section 5.3). We finally explore the physics relevance and interpretability of the learnt model by visualising its internal states (Section 5.4).

5.1 Physics Integration Performances

For this experiment, all models are trained on the augmented QM9 dataset for 360 epochs, while using validation data to track any signs of overfitting, and to save the model with the best generalisation loss during training. After training is complete, the best performing model is re-loaded and used to create predictions for the test dataset. We report the mean-absolute error (MAE) on energy estimation, the mean-squared error (MSE), and the relative error (RE $= \frac{|\hat{y}-y|}{|y|}$) between the true y and predicted $\hat{y}$ energy values. The average and standard deviation of these metrics are provided in Table 1.

Table 1: Evaluation of the impact of different physics integration strategies on the accuracy of energy estimation. Results are presented in the format: mean (std). 'BT' denotes 'bond type'.

	STRATEGY	MAE	MSE	RE
	No BT information	0.2195 (1.169)	1.4155 (17.913)	0.0027 (0.013)
	MPNN (BT specialised messages)	0.0909 (0.476)	0.2348 (6.679)	0.0012 (0.005)
	BT specialised node updates (case a.ii)	**0.0670** (0.141)	**0.0243** (0.406)	**0.0009** (0.002)
	# atoms	0.1713 (0.986)	1.0023 (13.940)	0.0021 (0.011)
Auxiliary estimates of	# orbitals	0.1946 (0.545)	0.3352 (7.223)	0.0025 (0.006)
	BT distance scaling	0.1194 (0.698)	0.5012 (9.548)	0.0015 (0.008)

Table 2: Evaluation of the MPNN and VIMPNN on the augmented QM9, infinite-, and finite-size crystal datasets.

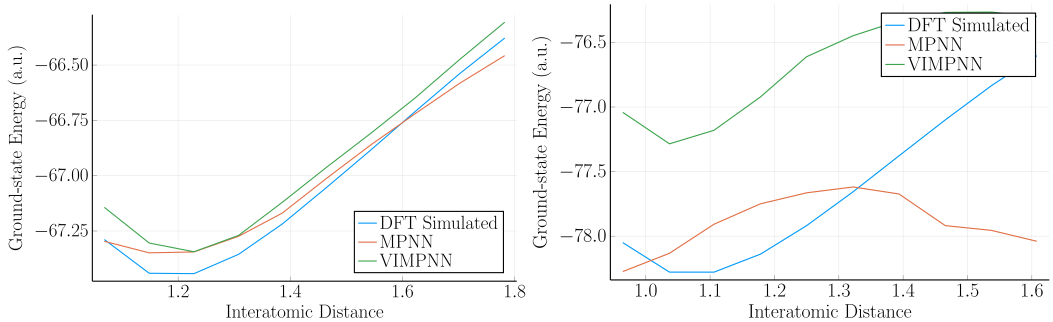
MODEL	DATASET	MAE	MSE	RE
MPNN	Augmented QM9	0.0909 (0.476)	0.2348 (6.679)	0.0012 (0.005)
	Infinite Crystals	**0.0335** (0.032)	**0.0022** (0.005)	**0.0012** (0.002)
	Stable Finite Crystals	2.9060 (4.200)	26.0691 (71.750	0.0047 (0.005)
	Finite Crystals	3.4466 (4.401)	31.2310 (71.743)	0.0058 (0.006)
VIMPNN	Augmented QM9	**0.0646** (0.197)	**0.0430** (1.093)	**0.0008** (0.002)
	Infinite Crystals	0.0335 (0.044)	0.0030 (0.009)	0.0015 (0.002)
	Stable Finite Crystals	**0.5131** (0.537)	**0.5519** (1.162)	**0.0016** (0.002)
	Finite Crystals	**2.3868** (3.557)	**18.3361** (56.050)	**0.0042** (0.004)

We compare all physics integration strategies listed in Section 4 against the baseline of MPNN with no bond type information used. We can see in Table 1 that all proposed methods have a positive effect on the energy estimation, with the introduction of bond type information having a particularly positive impact. This confirms Gilmer et al. (2017)'s observation that the bond type is a good feature descriptor for the estimation of energy. The use of bond type information in the update of nodes (case a.ii) in Section 4.2) has the strongest positive impact, with MAE decreasing from 0.2195 (baseline) to 0.067. In comparison with MPNN, this suggests that having specialised messaging channels and node updates based on bond type is more effective than having specialised messages in capturing the physics of atomic interactions. However, MPNN does improve significantly on the baseline, therefore in future works it may be interesting to combine these two approaches.

The auxiliary estimation of physical properties are also improved on the baseline models, but less so than the integration of bond type information. We interpret this as a result of the model having to implicitly discover and encode the useful physics representations required for accurate predictions. This more modest effect on performance may still be beneficial in combination to the bond type information strategy, thus we continue to use these auxiliary estimations in the rest of the experiments.

5.2 Evaluation of VIMPNN on out-of-equilibrium molecular and crystal data

In this experiment, we combine our proposed physics integration strategies, namely bond type-specialised node updates (case a.ii) of Section 4.2) and auxiliary estimations of physical properties, into the VIMPNN model. The energy estimation performance is measured on the 3 datasets using the same metrics as in Section 5.1. On the finite crystal dataset, two tests are performed on different subsets of the data. The first seeks to demonstrates a the model's ability to handle larger chemical systems by considering system sizes ranging from 15-75 atoms (see Section 3.3), but without the added complexity of compressing and dilating the interatomic distances. These changes of distances are included into the second test that considers all 6,710 crystal structures of up to 75 atoms due to memory constraints. The performance of MPNN and VIMPNN on the datasets are shown in Table 2.



(a) Example of accurate energy estimations made by both MPNN and VIMPNN.

(b) Example case of non-informative potential energy curve created with MPNN where no stable configuration at minimum energy may be found.

Figure 3: Comparison of Curve Predictions for Both MPNN and VIMPNN.

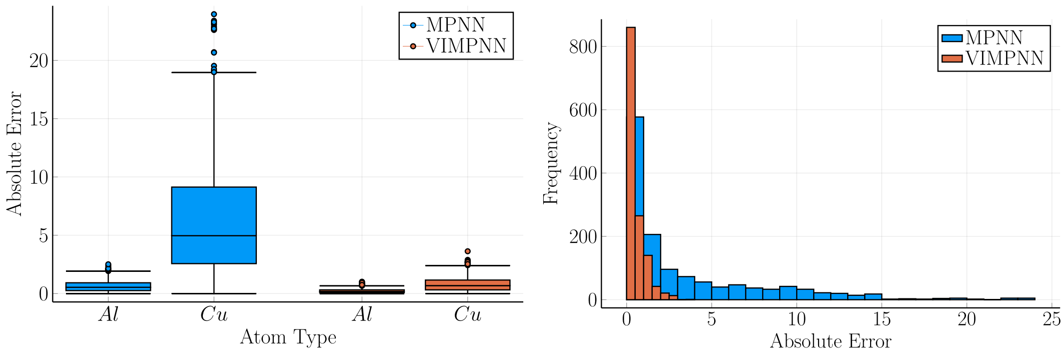
VIMPNN generally obtains better scores than MPNN on all datasets. Improvements are particularly strong on the Stable Finite Crystals dataset. In addition, the potential energy curves produced by several runs of VIMPNN at different interatomic distances prove to be more accurate and informative than those of MPNN for finding the minimum energy configuration (see an example in Fig. 3 for a molecular system). Indeed, we find that MPNN's estimations are more likely to produce curves that are close to the actual energy values but do not contain a minimum at the stable configuration (Fig. 3b). In these cases, MPNN estimates that the energy should always decrease with the reduction in interatomic distance, and is therefore probably not learning the repulsive effect happening when atoms get too close. For these difficult cases, VIMPNN is still able to create informative curves with a reasonable location for the stable configuration, although slightly worsened energy estimation. With exception of the infinite crystals case where there is no significant improvement with both MPNN's and VIMPNN's MAE being 0.335. As there are two types of configurations of infinite crystals, there perhaps is not enough varied information for the proposed methods (such as different bond types) to be of much benefit. By expanding this dataset to include a higher number of structures outside of the fcc lattice and a wider variety of atom types, we may see the usual improvement over MPNN.

When trained on the Finite Crystals dataset, we find MPNN is unable to learn both Copper and Aluminium crystals at the same time. There is a noticeable divergence in the absolute error on the different crystal types that contributes to the lower overall test scores made by MPNN (see Fig. 4a). The MPNN's absolute error for Copper crystals reaches a maximum of 23.96, and mean of 6.25, where as for Aluminium crystals the maximum and mean absolute error is 2.51 and 0.64, respectively. VIMPNN's 3.63 maximum and 0.8 mean absolute error for Copper demonstrates that it has a greater ability to handle both types of crystals during training. While VIMPNN's absolute error is still higher for Copper crystals than Aluminium, the difference is less pronounced and has lower absolute error values overall (Fig. 4b).

As discussed, DFT scales with respect to the number of orbitals in the system, meaning, with heavier and an increasing number of atoms, simulations become costly in terms of compute power and time. Therefore it would be advantageous for VIMPNN to estimate well for large chemical systems. While the VIMPNN has shown a performance improvement over the MPNN for both molecules and crystals, larger systems must still be considered as complex problems. This is clear when we visualise the potential energy curves for both small and large systems. For a system of 15 atoms, the model estimates the stable distance correctly (Fig. 5a) even if the minimum estimated energy is exaggerated. But as the number of atoms increase, the estimations begin to fluctuate, such as with the energy curves for a system containing 75 atoms, and in extreme cases the predicted energies no longer represent a usual curve (Fig. 5b). This indicates a need for accommodating message passing in large scale graph structures where many atomic interactions must be considered simultaneously.

5.3 Generalisation to Larger Crystal Systems

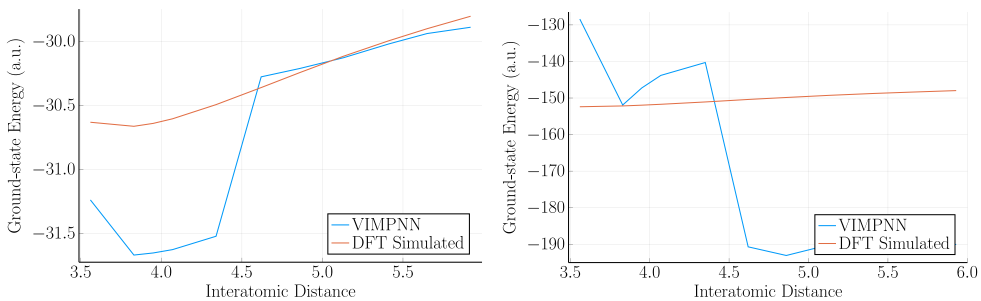
We investigate the VIMPNN's the ability to learn basic principle properties of atomic interactions in small chemical systems that are also transferable to arbitrary system sizes – in particular systems



(a) Error in estimating energy values of different atom types.

(b) Histogram of Absolute Errors for the Finite Crystals dataset.

Figure 4: Absolute Errors for MPNN and VMPNN trained on the Finite Crystals dataset.



(a) Finite Crystal with 15 aluminium atoms.

(b) Finite Crystal with 75 aluminium atoms.

Figure 5: Estimated potential energy curves for both small and large finite crystals.

larger than ones that were used for training. Ability to learn such properties would allow for quicker training times while generalising well irrespective of the graph size. To test whether the VIMPNN is capable of such task, we train the model on a filtered version of the Finite Crystals dataset, including only systems containing 25 or less atoms. The validation and testing remains unfiltered and thus the model estimates energies for systems up to 75 atoms of Copper and Aluminium crystals. Fig. 6 shows that the absolute error increases with the system size, thus the learnt atomic interaction is not being transferred directly. At a system of 15 atoms the MAE is 3.29 (perhaps due to the smaller amount of data used during training), while the MAE for systems of 75 atoms is 13.24.

5.4 Interpretation of the Node's Hidden State for Energy Estimation

When using the VIMPNN in-place of DFT, we are placing trust in the model for computational speed with minimal impact in accuracy. While DFT includes approximations to make computation more tractable, it is grounded in fundamental principles of atomic interaction. The VIMPNN however makes inferences about how atoms interact by searching for a function that generalises well between the input and output data. Indeed, we'll want ensure that it's learning an appropriate representation of atomic interaction, as well as being accurate in it's ground-state energy prediction. To investigate the learnt representations of atomic interactions, we visualise the hidden states during the readout phase.

With a VIMPNN model trained on the infinite crystals dataset, we take a basic 14 Aluminium atom fcc crystal and add an additional atom at 90% of the original distance. We then change the distance of the atom to the rest of the system. At 100 regular intervals between 90% and 120% of the original distance, the estimated energy and the hidden state of the new atom during the readout layer is returned. We then visualise this hidden state as an interpretation of the atoms contribution of the *'change'* in ground-state energy. At the optimal distance (Fig. 7a), the atom's hidden state results in a 0.0 contribution to the increase in energy, and is therefore lowest possible ground-state energy for this crystal (Fig. 7b). As the atom is moved closer, the ground-state energy increases the contribution

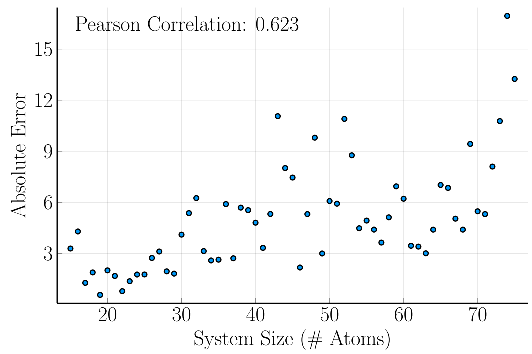
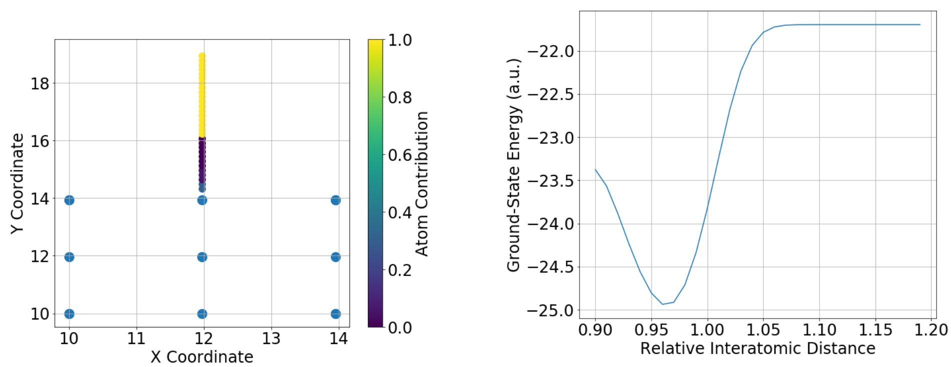


Figure 6: Absolute error when limiting training data to a maximum system size of 25 atoms.



(a) Moving an additional atom an increasing distance to the rest of the crystal.

(b) Estimated potential energy curve when moving added atom away from the crystal.

Figure 7: Visualisation of the VIMPNN nodes hidden state during the readout function for the energy estimation.

value begins to increase to 0.2. Finally, when the atom is pulled away from the rest of the crystal, its contribution reaches 1.0 until it has minimal interaction with the rest of the crystal until a point where the energy no longer changes due to the atom no longer interacting with the rest of the crystal.

6 Conclusion

We have shown how physics informed integration strategies can be used to learn better representations of bond types through an implementation of message channels. Our methods are incorporated into a Valence Interaction Message Passing Neural Network (VIMPNN) that uses specialised node updates where the messages sent between atoms are updated by bond-type weights, in addition to auxiliary system property estimations encouraging more useful representations that generalises better to new systems. The performance of VIMPNN is shown using an augmented QM9 dataset to include unstable configurations, as well as infinite and finite crystals at varying sizes up to 75 atoms. While our method has a performance advantage over an MPNN design, for large chemical systems the estimated energy curves are less representative of the simulated data. Future work should focus on adapting these methods to better accommodate larger and more complex structural patterns that will benefit more complex interactions, in addition to enabling estimations for graph sizes outside what has been included in the training data. Moreover, using basic low-level physical quantities such as valence electrons in different shells as apposed to providing predetermined bond types would make VIMPNN more flexible, enabling methods for generating new crystal systems.

REFERENCES

J. Baima, I. Bush, R. Orlando, R. Dovesi, and A. Erba. Large-Scale Condensed Matter DFT Simulations: Performance and Capabilities of the CRYSTAL Code. Journal of Chemical Theory and Computation, 13(10):5019–5027, 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00687.

O Bilek and L Skála. From Finite to Infinite Crystals: Analytic Solution of Simple Tight Binding Model of Finite SC, FCC and BCC Crystals of Arbitrary Size. Czechoslovak Journal of Physics, 28(9):1003–1019, 1978.

L C Blum and J.-L. Reymond. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database {GDB-13}. J. Am. Chem. Soc., 131:8732, 2009.

Aron J. Cohen, Paula Mori-Sánchez, and Weitao Yang. Challenges for density functional theory. Chemical Reviews, 112(1):289–320, 2012. ISSN 00092665. doi: 10.1021/cr200107z.

M J Feinberg and Klaus Ruedenberg. Paradoxical Role of the Kinetic-Energy Operator in the Formation of the Covalent Bond. The Journal of Chemical Physics, 54(4):1495–1511, 1971.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1263–1272, 4 2017.

André Severo Pereira Gomes, Christoph R. Jacob, and Lucas Visscher. Calculation of local excitations in large systems by embedding wave-function theory in density-functional theory. Physical Chemistry Chemical Physics, 10(35):5353–5362, 2008. ISSN 14639076. doi: 10.1039/b805739g.

Hong Jiang, Harold Baranger, and Weitao Yang. Density-functional theory simulation of large quantum dots. Physical Review B - Condensed Matter and Materials Physics, 68(16):1–9, 2003. ISSN 1550235X. doi: 10.1103/PhysRevB.68.165337.

B. P. Lanyon, J. D. Whitfield, G. G. Gillett, M. E. Goggin, M. P. Almeida, I. Kassal, J. D. Biamonte, M. Mohseni, B. J. Powell, M. Barbieri, A. Aspuru-Guzik, and A. G. White. Towards quantum chemistry on a quantum computer. Nature Chemistry, 2(2):106–111, 2010. ISSN 17554330. doi: 10.1038/nchem.483.

Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. Physical Review Letters, 114(9):1–5, 2015. ISSN 10797114. doi: 10.1103/PhysRevLett.114.096405.

Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, O. Anatole Von Lilienfeld, and Klaus-Robert Müller. Learning Invariant Representations of Molecules for Atomization Energy Prediction. In Advances in Neural Information Processing Systems 25, pp. 440–448, 2012. ISBN 9781627480031. doi: 10.1021/acs.jpclett.5b00831.

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. New Journal of Physics, 15, 2013. ISSN 13672630. doi: 10.1088/1367-2630/15/9/095003.

Matthias Rupp, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters, 108(5):1–5, 2012. ISSN 00319007. doi: 10.1103/PhysRevLett.108.058301.

Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. Nature Communications, 8:6–13, 2017. ISSN 20411723. doi: 10.1038/ncomms13890.

Zhe Shi, Evgenii Tsymbalov, Ming Dao, Subra Suresh, Alexander Shapeev, and Ju Li. Deep elastic strain engineering of bandgap through machine learning. Proceedings of the National Academy of Sciences, pp. 201818555, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1818555116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1818555116.

Keisuke Takahashi and Lauren Takahashi. Creating Machine Learning-Driven Material Recipes Based on Crystal Structure. The journal of physical chemistry letters, 10:283–288, 2019. ISSN 1948-7185. doi: 10.1021/acs.jpclett.8b03527.

Janis Timoshenko, Cody J. Wrasman, Mathilde Luneau, Tanya Shirman, Matteo Cargnello, Simon Russell Bare, Joanna Aizenberg, Cynthia M. Friend, and Anatoly I Frenkel. Probing atomic distributions in mono- and bimetallic nanoparticles by supervised machine learning. Nano Letters, pp. acs.nanolett.8b04461, 2018. ISSN 1530-6984. doi: 10.1021/acs.nanolett.8b04461. URL http://pubs.acs.org/doi/10.1021/acs.nanolett.8b04461.

Chenchen Wang, Ramamurthy Ramprasad, Sanguthevar Rajasekaran, Ghanshyam Pilania, and Xun Jiang. Accelerating materials property predictions using machine learning. Scientific Reports, 3 (1):1–6, 2013. doi: 10.1038/srep02810.