Training atomic neural networks using fragment-based data generated in virtual reality

Silvia Amabilino[*]
silvia.amabilino@bristol.ac.uk

Lars A. Bratholm[*][†]
lars.bratholm@bristol.ac.uk

Simon J. Bennie[*]
simon.bennie@bristol.ac.uk

David R. Glowacki[*][‡][§]
glowacki@bristol.ac.uk

Abstract

The ability to understand and engineer molecular structures relies on having accurate descriptions of the energy as a function of atomic coordinates. Here we outline a new paradigm for deriving energy functions of hyperdimensional molecular systems, which involves generating data for low-dimensional systems in virtual reality (VR) to then efficiently train atomic neural networks (ANNs). This generates high quality data for specific areas of interest within the hyperdimensional space that characterizes a molecule's potential energy surface (PES). We demonstrate the utility of this approach by gathering data within VR to train ANNs on chemical reactions involving fewer than 8 heavy atoms. This strategy enables us to predict the energies of much higher-dimensional systems, e.g. containing nearly 100 atoms. Training on datasets containing only 15k geometries, this approach generates mean absolute errors around $2\,\mathrm{kcal\,mol^{-1}}$. This represents one of the first times that an ANN-PES for a large reactive radical has been generated using such a small dataset. Our results suggest VR enables the intelligent curation of high-quality data, which accelerates the learning process.

1 Introduction

In the recent past, computations were mostly limited by the available processing power. With the machine learning revolution, the issues related to generating and curating data have become equally as important as the algorithms used to process and learn the data (1).

The molecular sciences have seen a surge in popularity of machine learning methods for a variety of applications, from designing new drug molecules (2) to planning synthetic chemistry strategies (3). Multiple research groups have been applying machine learning to the prediction of molecular energies and forces (4; 5), with the goal of accelerating molecular dynamics (MD) simulations. For small systems, ab initio calculations can be used to evaluate accurate energies and forces at each step of an MD simulation. However, this becomes too computationally expensive for larger systems and more approximate methods, such as force fields, are generally used. While much faster, these incur a trade-off in accuracy.

An alternative to force fields is to use accurate data, e.g. from electronic structure calculations, to fit potential energy surfaces (PES). Evaluating fitted PES should be faster than performing electronic

[*] School of Chemistry, University of Bristol, Bristol, BS8 1TS, UK

[†] School of Mathematics, University of Bristol, Bristol, BS8 1TW, UK

[‡] Department of Computer Science, University of Bristol, BS8 1UB, UK

[§] Intangible Realities Laboratory, University of Bristol, BS8 1UB, UK

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

structure calculations, but should provide similar accuracy to the underlying data. A variety of machine learning techniques have been used for fitting PES, for example permutationally invariant fitting (6), cubic splines (7), modified Shepard interpolation (8), interpolating moving least squares (9), Multi-State Empirical Valence Bond theory (MS-EVB) (10; 11), reproducing Kernel Hilbert space interpolation (12). Recently, Kernel Ridge Regression (KRR) and Neural Networks (NNs) have have attracted considerable attention (13; 14). However, KRR suffers from poor scaling with data set size, while NNs memory scaling does not depend on the number of data points. This contributed to NNs popularity for applications with large datasets.

In 2017, NNs were used to fit quantum mechanical DFT calculations in order to learn accurate and transferable potentials for organic molecules with up to 8 heavy atoms (4). These NNs were able to predict the energies of larger molecules with up to 53 atoms. However, they were trained on a dataset containing around 17.2 million compounds. Similarly, another group showed that by training on 15k different molecules and 3 million geometries, they obtained low errors when predicting the energy of molecules outside of the dataset (5).

Generating these large datasets can be extremely computationally expensive if accurate energies are required. Furthermore, sampling data to study the dynamics of reactive systems is especially challenging. A previous study (1) investigated how real-time interactive quantum molecular dynamics in virtual reality (iMD-VR) can be used to generate data for training NNs. iMD-VR relies on human intuition to efficiently sample hyperdimensional PESs (15), as users can interact ‘on-the-fly’ with real-time quantum mechanical molecular dynamics simulations and explore regions of interest on the PES (1). The NN trained on iMD-VR data was found to have similar performance to one trained on constrained molecular dynamics data, but the former had a lower mean absolute error (MAE) when predicting the energies of geometries close to the minimum energy path of the reaction. (1)

In this article, we investigate how real-time iMD-VR can be combined with fragment-based training of NNs to accurately predict the energy of large open-shell reactive systems.

We focus on the reaction of a large hydrocarbon chain ($C_{30}H_{62}$, called ‘squalane’) and a cyano radical (CN). We use a training dataset that does not contain the system we want to predict, but only smaller hydrocarbons with up to 6 carbon atoms. To date, most studies of chemical reaction surfaces with NNs have focused on systems with up to 19 atoms (1). Therefore, with 94 atoms, this system represents (to the best of our knowledge) one of the largest radical systems for which an NN-fitted PES has been developed.

2 Method

2.1 Dataset generation

In this study we focused on smaller datasets compared to recent previous works (4; 5). The dataset generation included multiple steps. We first sampled trajectories of CN performing primary, secondary, and tertiary H-abstraction from different small hydrocarbons: methane, ethane, isobutane, isopentane and isohexane. This was done using the open-source iMD-VR framework ‘Narupa’ (16; 1). We used the SCINE implementation of PM6 as the force engine (17; 18). Fig. 1 shows a representative user-guided abstraction pathway, obtained by bringing the CN into proximity with a primary hydrogen on isopentane (1).

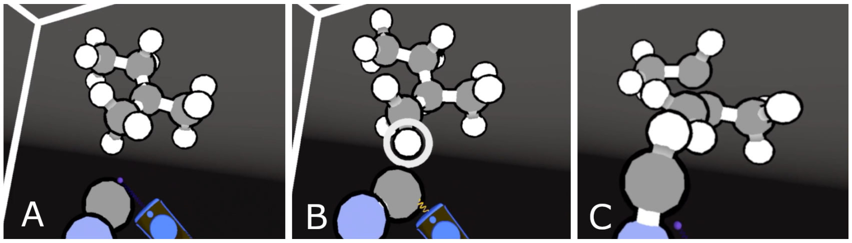

Figure 1: Sampling of a H-abstraction trajectory using interactive molecular dynamics in VR. The reactants (A) are brough in close proximity (B) to form the products (C).

After sampling the abstraction trajectories for each system, the energies were re-calculated at DFT level (PBE functional with TZVP basis set). The energies of different hydrocarbons were scaled so that they had similar magnitudes. We then constructed six mixed datasets all containing 15k data points, but with different ratios of each hydrocarbons:

- Training set 1: 15k methane
- Training set 2: 10k methane, 5k ethane
- Training set 3: 8k methane, 4k ethane, 3k isobutane
- Training set 4: 7.5k methane, 3.5k ethane, 2.5k isobutane, 1.5k isopentane
- Training set 5: 7.5k methane, 7.5k isopentane
- Training set 6: 8k methane, 4k isopentane, 3k isohexane

Following the same procedure outlined above, we also generated a secondary H-abstraction trajectory for squalane. This was used as the test set, to assess the accuracy of the NN-fitted PES.

2.2 Fitting the potential energy surface

We used the Atomic Neural Networks (ANNs) (19) architecture and the Smith formulation of the Atom Centred Symmetry Functions (ACSFs) (4; 20) as the molecular representation. We used the implementation of ANNs and ACSFs present in the QML Python package (21).

After generating the datasets, we optimised the ANNs hyper-parameters for each dataset using a random search (22). After training, the performance of each ANN was evaluated by predicting the energy of $C_{30}H_{62}$ + CN and comparing it to the reference DFT data. A constant offset was removed from all ANNs predictions in order to get a better comparison of the relative predictions quality.

3 Results

Here we attempt to establish whether the ANNs trained on the small systems are transferable to the larger reactive system. In addition, we want to understand how large the ‘fragments’ of the target system need to be before reaching acceptable accuracy.

Table 1: Mean Absolute Errors and R^2 values of the ANN predictions for the trajectory of squalane reacting with CN compared to the DFT reference.

	MAE ($\mathrm{kcal\,mol^{-1}}$)	$\mathbf{R^2}$
Training set 1	9.9 ± 7.3	0.01
Training set 2	10.3 ± 8.6	-0.18
Training set 3	4.9 ± 4.1	0.73
Training set 4	2.4 ± 2.2	0.93
Training set 5	2.6 ± 2.0	0.92
Training set 6	2.0 ± 1.6	0.96

The trends observed are explained below, while the MAEs and the R^2 scores are shown in Table 1. The R^2 scores were calculated with Scikit-learn, (23) and can be negative (24). The ANN trained only on methane can reproduce neither the change in energy between the reactants and the products, nor the oscillations in energy due to the bonds stretching motions. This is expected, as the ANN has not learnt about C-C bonds or secondary H-abstractions. Adding ethane does not improve the results, while adding isobutane halves the MAE and the reactants energies begin to be predicted better. However, the products are not predicted as well, as isobutane contains only primary and tertiary Hs and the squalane trajectory involves a secondary H-abstraction. To get considerable improvement, isopentane has to be included in the dataset. Even adding as few as 1.5k isopentane data points lowers the MAE for the squalane trajectory to around $2.5\,\mathrm{kcal\,mol^{-1}}$. This is because isopentane contains most of the important features required to describe squalane: multiple C-C bonds and primary, secondary and tertiary Hs. In fact, increasing the number of isopentane geometries in the dataset or adding longer hydrocarbons (i.e. training set 5 and 6) does not considerably change the results.

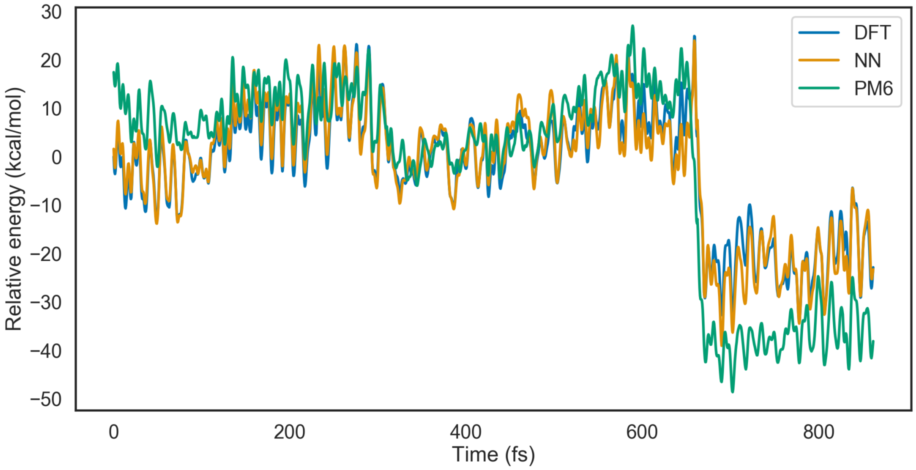

Figure 2: Comparison of the ANN (trained on training set 6) predictions with the PM6 and DFT energies for the trajectory of squalane reacting with CN.

We then compared the energies of the squalane H-abstraction trajectory calculated with DFT, PM6 and the ANN trained on training set 6. A constant offset was also removed for the PM6 energies. Fig. 2 shows that the ANN predictions are much closer to the DFT reference compared to the PM6 energies. More quantitatively, the ANN predictions have a MAE of $2.0\,\mathrm{kcal\,mol^{-1}}$, while the PM6 energies have a MAE of $8.3\,\mathrm{kcal\,mol^{-1}}$. In addition, the ANN energies are about an order of magnitude faster to evaluate than for PM6.

4 Conclusion

This study investigated the transferability of reactive potential energy surfaces (PES) fitted with atomic neural networks (ANNs), with training data generated using interactive molecular dynamics in virtual reality.

ANNs were trained on six different training sets containing different proportions of small hydrocarbons (with at most 6 carbon atoms) reacting with CN. After training, the ANNs were used to predict the energy of squalane ($C_{30}H_{62}$) reacting with CN.

The results showed that only including methane and ethane in the training set gives poor results. From isobutane onwards, the mean absolute errors are under $5\,\mathrm{kcal\,mol^{-1}}$. The results improve considerably (MAE around $2.5\,\mathrm{kcal\,mol^{-1}}$) once isopentane is added to the training set, but then there is only around $0.5\,\mathrm{kcal\,mol^{-1}}$ improvement if longer hydrocarbons are added.

In conclusion, this shows that reactive potential energy surfaces fitted with ANNs can be transferable. However, this is the case only if enough molecules in the training set capture the key chemical interactions of the large system. Furthermore, we showed that good accuracy can be achieved with training sets containing only 15k data point. Following this approach, the cost of generating the dataset and training the neural network can be drastically reduced.

The next step in this project will be to study the prediction of the forces in addition to the energies. If the ANNs predictions for the forces could maintain similar accuracy and speed up as the energies, this would effectively enable to perform accurate molecular dynamics simulations for a new range of systems.

References

[1] S. Amabilino, L. A. Bratholm, S. J. Bennie, A. C. Vaucher, M. Reiher, and D. R. Glowacki, "Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality," *The Journal of Physical Chemistry A*, vol. 123, no. 20, pp. 4486–4499, 2019.

[2] A. Gupta, A. T. Müller, B. J. Huisman, J. A. Fuchs, P. Schneider, and G. Schneider, "Generative recurrent networks for de novo drug design," *Molecular informatics*, vol. 37, no. 1-2, p. 1700111, 2018.

[3] M. H. Segler and M. P. Waller, "Neural-symbolic machine learning for retrosynthesis and reaction prediction," *Chemistry–A European Journal*, vol. 23, no. 25, pp. 5966–5971, 2017.

[4] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," *Chem. Sci.*, vol. 8, pp. 3192–3203, 2017.

[5] K. Yao, J. E. Herr, D. Toth, R. Mckintyre, and J. Parkhill, "The tensormol-0.1 model chemistry: a neural network augmented with long-range physics," *Chem. Sci.*, vol. 9, pp. 2261–2269, 2018.

[6] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," *International Reviews in Physical Chemistry*, vol. 28, no. 4, pp. 577–606, 2009.

[7] S. McKinley and M. Levine, "Cubic spline interpolation," *College of the Redwoods*, vol. 45, no. 1, pp. 1049–1060, 1998.

[8] M. A. Collins, "Molecular potential-energy surfaces for chemical reaction dynamics," *Theoretical Chemistry Accounts*, vol. 108, no. 6, pp. 313–324, 2002.

[9] G. G. Maisuradze and D. L. Thompson, "Interpolating moving least-squares methods for fitting potential energy surfaces: Illustrative approaches and applications," *The Journal of Physical Chemistry A*, vol. 107, no. 37, pp. 7118–7124, 2003.

[10] A. Warshel and R. M. Weiss, "An empirical valence bond approach for comparing reactions in solutions and in enzymes," *Journal of the American Chemical Society*, vol. 102, no. 20, pp. 6218–6226, 1980.

[11] D. R. Glowacki, A. J. Orr-Ewing, and J. N. Harvey, "Non-equilibrium reaction and relaxation dynamics in a strongly interacting explicit solvent: F + cd3cn treated with a parallel multi-state evb model," *The Journal of Chemical Physics*, vol. 143, no. 4, p. 044120, 2015.

[12] O. T. Unke and M. Meuwly, "Toolkit for the construction of reproducing kernel-based representations of data: Application to multidimensional potential energy surfaces," *Journal of Chemical Information and Modeling*, vol. 57, no. 8, pp. 1923–1931, 2017.

[13] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," *Science Advances*, vol. 3, no. 5, 2017.

[14] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, "Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels," *The Journal of Chemical Physics*, vol. 146, no. 24, p. 244108, 2017.

[15] M. O'Connor, H. M. Deeks, E. Dawn, O. Metatla, A. Roudaut, M. Sutton, L. M. Thomas, B. R. Glowacki, R. Sage, P. Tew, M. Wonnacott, P. Bates, A. J. Mulholland, and D. R. Glowacki, "Sampling molecular conformations and dynamics in a multiuser virtual reality framework," *Science Advances*, vol. 4, no. 6, 2018.

[16] "Narupa." `https://gitlab.com/intangiblerealities`, 2019.

[17] M. P. Haag and M. Reiher, "Real-time quantum chemistry," *International Journal of Quantum Chemistry*, vol. 113, no. 1, pp. 8–20, 2013.

[18] A. C. Vaucher, M. P. Haag, and M. Reiher, "Real-time feedback from iterative electronic structure calculations," *Journal of computational chemistry*, vol. 37, no. 9, pp. 805–812, 2016.

[19] J. Behler, "Neural network potential-energy surfaces for atomistic simulations," in *Chemical Modelling: Applications and Theory Volume 7*, vol. 7, pp. 1–41, The Royal Society of Chemistry, 2010.

[20] J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," *The Journal of Chemical Physics*, vol. 134, no. 7, p. 074106, 2011.

[21] A. Christensen, F. Faber, B. Huang, L. Bratholm, A. Tkatchenko, K. Muller, and A. von Lilienfeld, "Qml: A python toolkit for quantum machine learning." `https://github.com/qmlcode/qml`, 2017.

[22] R. T. McGibbon, C. X. Hernández, M. P. Harrigan, S. Kearnes, M. M. Sultan, S. Jastrzebski, B. E. Husic, and V. S. Pande, "Osprey: Hyperparameter optimization for machine learning," Sept. 2016.

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

[24] "Scikit-learn." `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html`.