Leveraging Multitask Learning to Improve the Transferability of Machine Learned Force Fields

Leif Jacobson,[*,†] James Stevenson,[‡] Farhad Ramezanghorbani,[¶] Steven Dajnowicz,[‡] and Karl Leswing[‡]

†*Schrödinger Inc., 101 SW Main Street, Suite 1300, Portland, OR 97204*

‡*Schroödinger Ince., 1540 Broadway, 24th floor, New York, NY 10036*

¶*Nvidia, 41 E 11st, New York, NY 10003*

E-mail: leif.jacobson@schrodinger.com

Abstract

Transferable neural network potentials have shown great promise as an avenue to increase the accuracy and applicability of existing atomistic force fields for organic molecules and inorganic materials. Training sets used to develop transferable potentials are very large, typically millions of examples, and as such, are restricted to relatively inexpensive levels of *ab initio* theory, such as density functional theory in a double- or triple-zeta quality basis set, which are subject to significant errors. It has been previously demonstrated using transfer learning that a model trained on a large dataset of such inexpensive calculations can be re-trained to reproduce energies of a higher level of theory using a much smaller dataset. Here, we show that more generally, one can use hard parameter sharing to successfully train to multiple levels of theory simultaneously. We demonstrate that simultaneously training to two levels of theory is an alternative to freezing layers in a neural network and re-training. Further, we show that training multiple levels of theory can improve the overall performance of

all predictions and that one can transfer knowledge about a chemical domain present in only one of the datasets to all predicted levels of theory. This methodology is one way in which multiple, incompatible datasets can be combined to train a transferable model, increasing the accuracy and domain of applicability of machine learning force fields.

1 Introduction

Machine learning force fields (MLFFs) such as the high dimensional neural network potentials (HDNNPs)[1,2] message passing[3,4] and equivariant[5–8] neural network potentials are currently gaining traction as an alternative to traditional atomistic force fields in computational chemistry and materials science. Such methods promise to increase the accuracy and domain of applicability of model potentials used to describe molecules and materials. These methods use regression techniques, borrowed from the field of machine learning, to train a model which takes cartesian coordinates (and sometimes net charge[9–12]) as input and outputs energies. These energies are trained to reproduce *ab initio* energies to high precision, typically inside of chemical accuracy, usually defined to be 1 kcal/mol. Recently, several such MLFF models have been reported which are meant to be transferable to molecules[11–16] or inorganic materials[17–19] for which they have not been explicitly trained. In constructing a foundational model for organic molecules one trains to a very large dataset of example organic molecules which is representative of the chemical space one anticipates applying such a model. In a pioneering work Smith *et al*[13] trained a model to approximately 18 million conformations of organic molecules, labeled at the ωB97-X/6-31G(d) level. It would be computationally intractable to label a dataset of this size with a more expensive *ab initio* level of theory or even utilizing an atom centered basis set which is more complete. Recognizing this issue, the same authors demonstrated that it is possible to "transfer learn" to a much smaller dataset of higher level labels by freezing several of the layers of the atomic neural networks which make up the pre-trained HDNNP (in this case ANI-1x).[20] This idea of transfer learning from

a large dataset to a small, more expensive one is a very powerful idea which has gained some popularity in the field of MLFF.[21,22]

Multiple large, and mutually incompatible datasets of druglike molecules have already appeared in the literature. The very popular QM9 dataset consists of various properties of 134,000 organic molecules computed at the B3LYP/6-31G(d, p) level,[23] the ANI dataset consists of 20 million examples computed at the ωB97X/6-31G(d) level,[13] the ANI-1x dataset[24] consists of roughly 5 million examples labeled with the ωB97X functional with both the 6-31G(d) and def2-TZVPP basis sets whereas the ANI1-CCX dataset is 500,000 examples labeled with DLPNO-CCSD(T)/CBS.[20] The OrbNet Denali dataset consists of roughly 2 million examples labeled at the ωB97X-D3/def2-TZVP level,[25] and finally the SPICE dataset was reported with 1.1 million examples at the ωB97M-D3(BJ)/def2-TZVPPD level.[26] None of these datasets can be naively combined since they represent different model chemistries. Many settings used in a DFT calculation can affect the absolute energy, and as more datasets appear they will likely have small (or large) differences which would give one pause when combining the data naively. Examples include seemingly minor modifications of a functional (e.g. ωB97X, ωB97X-D, ωB97X-D3), similar but incompatible basis sets (e.g. def2-TZVP, def2-TZVPD), the use of integral approximations such as density fitting or pseudospectral approximations, use or lack of a pseudo or effective core potential and even various integral, self-consistent field convergence, and linear dependence thresholds. Nonetheless, each new dataset entering the literature brings some new value. It is desirable to use methods which are able to extract valuable information from all datasets appearing in the public domain without the need to relabel them, in order to build a more general, transferable model.

One way to benefit from incompatible data is hinted at by the work of Smith *et al.*[20] but we believe can be generalized such that many levels of theory can be simultaneously trained to. In their work, an HDNNP (ANI-1x) was pre-trained to the very large ANI-1x dataset (see above). Several of the layers of each atomic network were then frozen and the remaining unfrozen parameters were re-optimized by training to a much smaller dataset with labels from

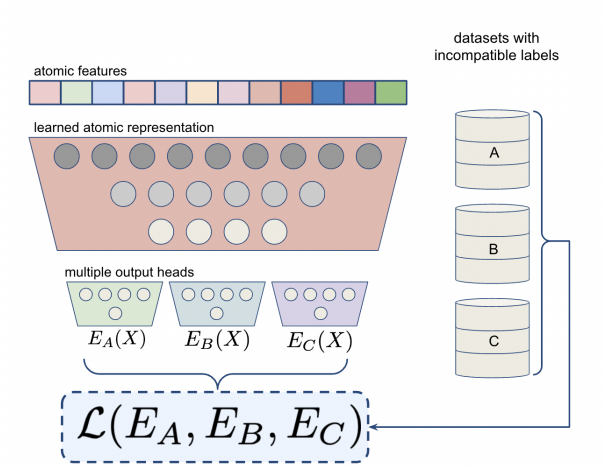

Figure 1: A multitask model shown for one element. The model transforms some input feature vector to a shared latent vector. Each output head transforms this shared vector into a different prediction. All of the parameters of the model are trained simultaneously on a collection of datasets with incompatible labels. There is a one-to-one mapping between output heads and datasets.

a higher level of theory (DLPNO-CCSD(T)/CBS). We can re-interpret the work of Smith, not as a single transfer-learned model, but as a *pair* of models which utilize hard parameter sharing over the frozen layers in the transfer learning protocol. If one were to then merge the frozen layers of the two models (they are identical after all) then these two models could be interpreted as one single *multitask* model with two outputs. This type of architecture is depicted in Fig. 1. Here, for an atomic neural network with a total of N layers, the first M layers are shared and used to transform some input feature vector into an intermediate, or latent, representation which is finally transformed into two or more output energies. In an HDNNP these output atomic energies are summed independently to give molecular energies. This interpretation of transfer learning as multitask learning has several useful advantages: It can be generalized to many outputs, does not require pre-training to one output, both outputs can improve as information about different chemical spaces is incorporated into the latent space that is shared by the various output heads and no empirical selection of frozen layers is required. Finally, we point out that this architecture is not novel and in fact many MLFF methodologies already use multiple output heads to train to different properties such

as energy, dipole moment, polarizability, HOMO-LUMO gap, etc.[3–5] For such models there is no change in architecture required, one must simply update the loss function to include contributions from more than one energy. As far as we are aware multitask learning to multiple levels of theory has not been previously reported. A similar strategy has, however, been recently used by Takamoto *et al.*[19] where the model is conditioned on the task to be predicted (i.e. used as a feature). While no detailed comparison on the effectiveness of that approach was reported it seems such an approach would have similar benefits as the one described in detail here.

In this article we will train single and multitask models on four publicly available datasets (ANI-1x,[20] ANI-ccx,[20] OrbNet Denali[25] and SPICE[26]) and then evaluate the performance of all of the output heads on single point energy tasks covering a range of interaction types. We will train to energies and dipole moments. Some of the datasets also have labeled forces but we do not use this data for consistency. Some of the datasets also lack dipole information but we view this data as secondary in importance. First we will reproduce the results of Smith et al.[20] and show that we can perform multitask training as an alternative to pre-training a model on ANI-1x and then freezing layers and re-training to the ANI-ccx dataset. We will then show that adding the OrbNet Denali dataset can improve the performance of the ANI-1x and ANI-ccx predictions on torsion profiles of both neutral and ionic molecules even though the ANI-1x and ANI-ccx datasets contain no ionic examples. Finally we will show that adding the SPICE dataset improves the performance of intermolecular interactions of the ANI-1x, ANI-ccx and OrbNet Denali outputs, as measured by errors on the popular S66x8 dataset.[27,28]

2 Details

The single task QRNN model was described recently[12] and is only briefly summarized here; the only modification was to add additional output heads to each atomic energy neural

network (NN) such that for a model trained to N tasks, each atomic NN produces N outputs. These atomic energies are summed to produce N total energies. QRNN is an example of an HDNNP and utilizes two separate NN for each atom type that is supported. The first such NN takes modified Behler-Parinello type symmetry functions[1,13] as input and predicts atomic electronegativities. These then enter into a simplified charge equilibration equation that predicts atomic charges. These charges are determined recursively by using them as features into the same NN. We only perform two recursive steps and then use the final charges as features, in addition to geometric symmetry functions and charge-weighted radial symmetry functions, as input to a NN which predicts atomic energies. Finally we sum the atomic energies, together with a damped coulomb interaction based on the predicted atomic charges and Grimme's D2 dispersion correction[29] to predict a total energy.

As in previous work we use a multitask loss function[30]

$$\mathcal{L}_{mtl} = \sum_I \frac{\mathcal{L}_{E_I}}{2\sigma_{E_I}^2} + \frac{\mathcal{L}_q}{2\sigma_q^2} + log(\sigma_q \prod_I \sigma_{E_I}) \ . \tag{1}$$

$\mathcal{L}_{E_I}$ and $\mathcal{L}_q$ are the loss functions for the I^{th} energy task and the charge task, respectively. The inverse weights, σ_{E_I} and σ_q are trainable parameters. This approach obviates the need to hand tune the weights of the individual tasks in the overall loss function. The energy loss function is taken to be the squared error between the predicted energies and energy labels. We also train to the squared error of predicted dipole moments to those predicted by DFT. Unlike the treatment of energy we do not differentiate atomic charge by level of theory, instead assuming that dipole moments from all levels of theory are equivalent. The charges could be treated separately, as we do energies, but leave that to future work. We suspect the differences in atomic charges is much less important to model than the difference in energies, as the electron density tends to be similar between different density functionals,[31] and the partial charges are more dependent on the partitioning scheme. Since the energy predicted by our models is dependent on the predicted atomic partial charges,

it is possible to backpropagate through the whole model and train both of the atomic NNs to energies alone. However, unless the training set contains many examples of atomic pairs with distances greater than the NN cutoff the predicted charges may not produce accurate long range interaction energies. Training to dipole moments aids in giving correct long range electrostatic interactions and helps to refine the charge predictions.

To minimize the loss function we use the AdamaxW optimizer (Adamax with decoupled weight decay) with a weight decay of 1.0e-4. We utilize early stopping with a patience of zero and a maximum of 100 cycles, where each cycle consists of training to 5 million examples in batches of 128. We do not optimize batches to balance the frequency of seeing the various labels. We simply select a random batch and update the loss function for the labels that we encounter. We train ensembles with 7 members and each ensemble member is trained to a 90/10 train/validation random split of the training data which is performed independently at run time (as such there is some degree of overlap between the training sets of the ensemble members). We use exponential moving averaged weights and biases to evaluate validation and test errors, which are updated every 10 batches with a smoothing factor of 0.001. As is standard practice we reduce the range of the energy labels and center them towards zero by subtracting an offset energy for each atom, defined by fitting a linear model to each level of theory. For purposes of multi-label training, this procedure has the further benefit of making the different labels more similar to each other by effectively removing the differences between levels of theory. All models are implemented and trained in a locally modified copy of the TorchANI open source software package,[32] which is an implementation of HDNNP models in PyTorch. Each neural network is trained on a single GPU with single-precision floating point, and all test results are computed with double-precision inference. Hyperparameters defining the network dimensions and AEVs are taken from previous work.[12] The output heads in this study are defined as a single layer, and as such, the multitask model has the same layer dimensions as the single task models described previously, and trained here.[12]

3 Results

Table 1: Summary of properties of datasets used in this work: ANI-1x,[20] ANI-ccx,[33] OrbNet Denali[25] and SPICE.[26] Abbreviated names of the datasets we use throughout are given in parenthesis.

Dataset	Number of examples (millions)	Level of theory	Has dipole moments	Targeted chemical space
ANI-1x (1X)	5	ωB97X/ 6-31G(d)	Yes	neutral molecules
ANI-ccx (CCX)	0.5	DLPNO-CCSD(T)/ CBS	No	neutral molecules
OrbNet Denali (D)	2	ωB97X-D3/ def2-TZVP	No	neutral and ionic molecules
SPICE (S)	1	ωB97M-D3(BJ)/ def2-TZVPPD	Yes	neutral and ionic molecules and clusters

The four datasets we use here (ANI-1x, ANI-ccx, OrbNet Denali and SPICE) will be abbreviated as 1X, CCX, D and S, respectively. We will denote a QRNN multitask model trained to datasets X, Y and Z as QRNN(X, Y, Z). (The order X, Y, Z does not matter as all are trained simultaneously.) Further, we will refer to predictions made by such a model using the output head corresponding to dataset Y as QRNN(X, Y, Z)[Y]. For singletask models we will often omit the square bracket notation since there is no ambiguity in this case. The properties of the four datasets relevant to this work and the abbreviations just described are summarized in Table 1.

The ANI-1x dataset was constructed utilizing active learning to broadly cover the space of conformations of neutral druglike molecules and utilizes the ωB97X functional in conjunction with a rather modest double zeta quality basis set, 6-31G(d).[24] A subset of this data was subsequently labeled with a much higher coupled cluster level of theory combined with a near complete basis set (DLPNO-CCSD(T)/CBS) and was utilized to demonstrate a transfer learning protocol.[20] The authors of that work show that this procedure gives significantly improved results when testing on challenging reaction energies of isomers and

reproduction of torsional energy profiles in reference to high levels of theory. In Sec. 3.1 we show that multitask training on these two datasets is a viable alternative with results nearly indistinguishable from those of the original transfer learning protocol.

As we have pointed out previously[15] the impressive results for rotamer scans of ANI-1x and ANI-ccx trained models degrade when one investigates a broader sampling of chemical space. In Sec.3.2 we train to all four datasets and show that the ANI-1x and ANI-ccx output heads improve in this regard. We also show that predictions at these levels of theory greatly improve for chemical spaces that are lacking in the corresponding datasets: ionic molecules and intermolecular interactions. By multitask learning to datasets that cover these new chemical spaces, knowledge can be transferred amongst the output heads. We offer some concluding remarks in Sec. 5

3.1 Multitask learning on ANI-1x and ANI-ccx datasets

As a first test we explore whether our method can qualitatively reproduce the results of Smith *et al.* Figure 2 shows the error, relative to high level references on the HC7/11[34] and ISOL[35] test sets. In this test each molecule is re-optimized using the potential from each model. These test sets probe hydrocarbon reaction and isomerization energies and organic molecule isomerization energy, respectively. The first interesting feature to note is that the combination of ωB97X with a double-zeta quality basis shows fairly poor agreement with the high level reaction energies, displaying errors as high as 40 kcal/mol. The predictions from ANI-1x closely track the reference it was trained to, as does our multitask model, when using the ANI-1x output head. In contrast, the transfer learned ANI-ccx shows significant improvement over ANI-1x and qualitatively tracks the reference data, with the largest error reduced to 5 kcal/mol. Our multitask model reproduces this behavior and displays a maximum error of 4 kcal/mol. Overall, our multitask model modestly outperforms ANI-1x and ANI-ccx on this test. Compared to the references the models were trained to, the ANI-1x model yields MUEs of 4.4 and 1.5 kcal/mol for the HC7/11 and ISOL tests whereas our

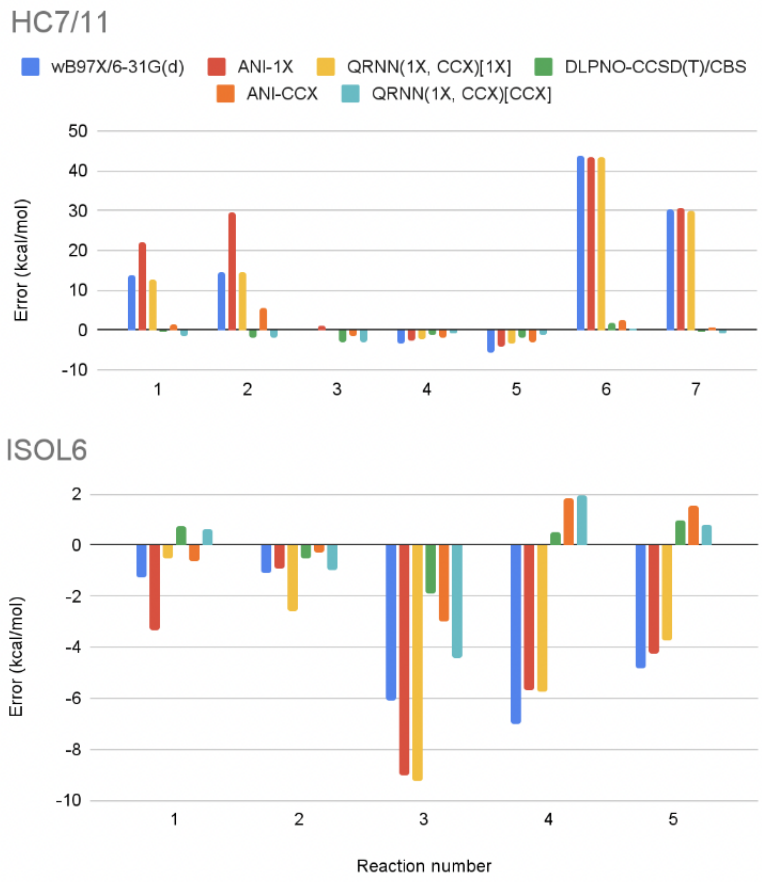

Figure 2: Errors of various methods relative to reference results. The HC7/11 dataset reference is MP2/6-311+G(2df, 2p)[34] whereas the reference for the ISOL test set is CCSD(T)-F12/aug-cc-pVDZ[35] ANI-1x, ANI-ccx, ωB97X-D/6-31G(d) and DLPNO-CCSD(T)/CBS values are taken from Smith *et al.*[20] The QRNN results simultaneously trained to both ANI-1x and ANI-ccx but evaluated on either the ANI-1x [QRNN(1X, CCX)[1X]] or the ANI-ccx [QRNN(1X, CCX)[CCX]] output head reproduce the results of Smith.

multitask model gives MUE of 0.9 and 1.6 kcal/mol. The ANI-ccx model affords MUEs of 1.9 and 0.9 kcal/mol for these benchmarks while our multitask model evaluated on the ANI-ccx output head gives MUEs of 0.7 and 0.9 kcal/mol.

Figure 2 highlights the fact that the two output heads are able to display very different behaviors. For example, the reaction energy in reaction 6 of the HC7/11 dataset differs by over 40 kcal/mol between the ωB97X-D/6-31G* and DLPNO-CCSD(T)/CBS levels of theory. This difference is accurately captured by the multitask model where both output heads predict the reaction energy of their reference data to an accuracy of about 1 kcal/mol.

This is surprising in light of the fact that the output heads are only differentiated by a single, linear layer, this will be discussed further in Sec. 4.

Next, we turn our attention to the performance of the multitask models on reproducing relative energies over torsion scans on a set of small organic molecules. Here we compare single-task QRNN models trained solely to either the ANI-1x or ANI-ccx datasets to a multitask model trained to both and evaluated on the two output heads. Mean Absolute Deviations (MADs) of the relative energy over each torsion scan are computed for a subset of the dataset of Sellers *et al.*[36] We restrict our attention only to molecules composed of the elements Hydrogen, Carbon, Nitrogen and Oxygen (HCNO), of which there are 45 examples in this dataset. Here, the reference energy is CCSD(T)/CBS and we utilize the reference geometries from Sellers. Figure 3 shows that the behavior of the models trained to two datasets is slightly improved over single-task models trained to only a single dataset. The median of the MAD of the models predicting on the ANI-1x output are the same, 0.47 kcal/mol. The median MAD of the single task model on the ANI-ccx output head is 0.31 kcal/mol whereas the multitask model gives a median MAD of only 0.26 kcal/mol. It is also clear from Figure 3 that the spread of errors is much smaller for the multitask model on the ANI-ccx head and the upper quartile is greatly diminished. Training to the combination of the two datasets appears to modestly improve the results on this dataset as compared to training to only the ANI-ccx dataset. Our results are in good agreement with those of Smith *et al.*[20] who report a median MAD of 0.23 kcal/mol for the ANI-ccx model.

3.2 Multitask learning on four datasets

In Sec. 3.1 we demonstrated that our multitask procedure is a simplified, but viable alternative to the transfer learning procedure of Smith *et al.* Next we turn to study the behavior of the procedure when we train to more than two datasets. We now include the OrbNet Denali[25] and SPICE[26] datasets. Here we will use the HCNO subset of a test set of torsion scans of neutral druglike molecules reported previously to probe torsion profiles over a broader re-

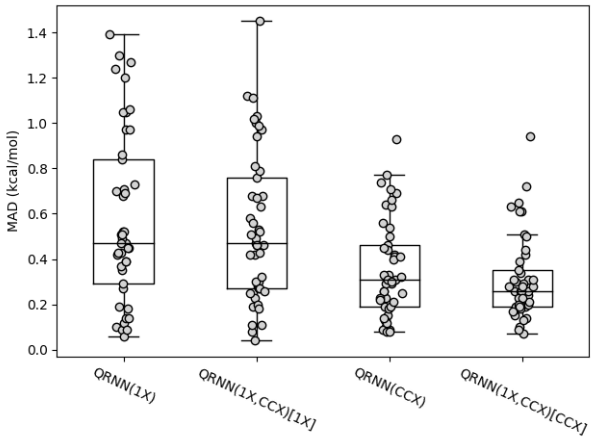

Figure 3: Mean average deviation (MAD) of a relative conformational energies over torsion scans of small organic molecules in kcal/mol relative to CCSD(T)/CBS estimates. Errors of QRNN models trained to the single datasets ANI-1x and ANI-ccx are shown alongside those trained to both datasets in conjunction. The box shows the first and third quartiles whereas a line is given at the median. The whiskers show the maximum and minimum points outside of the box but inside 1.5 times the interquartile range.

gion of chemical space than that probed by the Sellers dataset studied in Section 3.1.[15] We have relabeled this test set with the ωB97X-D3/def2-TZVP level of theory and use that as the reference. We believe this reference is sufficient for our purposes and the error between this level of theory and CCSD(T)/CBS for these relative conformational energies are likely to be much smaller than the errors between the models and CCSD(T)/CBS results.[37] This dataset contains 369 torsion scans and covers a much more significant sample of chemical space, representing a stronger test of transferability.

For each torsion scan we compute an RMSD, choosing the structure with lowest reference energy to define the energy zero. Figure 4 shows the distribution of these errors. Here, we compare results for models in several training experiments. Four models were trained to single datasets (QRNN(1X), QRNN(CCX), QRNN(D) and QRNN(S)), a model trained to two datasets QRNN(1X, CCX) and a model trained to all four datasets QRNN(1X, CCX, D, S). Multitask models predicting on the 1X, CCX and S output heads generally improve over singletask models whereas multitask results slightly degrade when using the D output head. Concretely, the median RMSD increases from 0.49 kcal/mol when training to only OrbNet

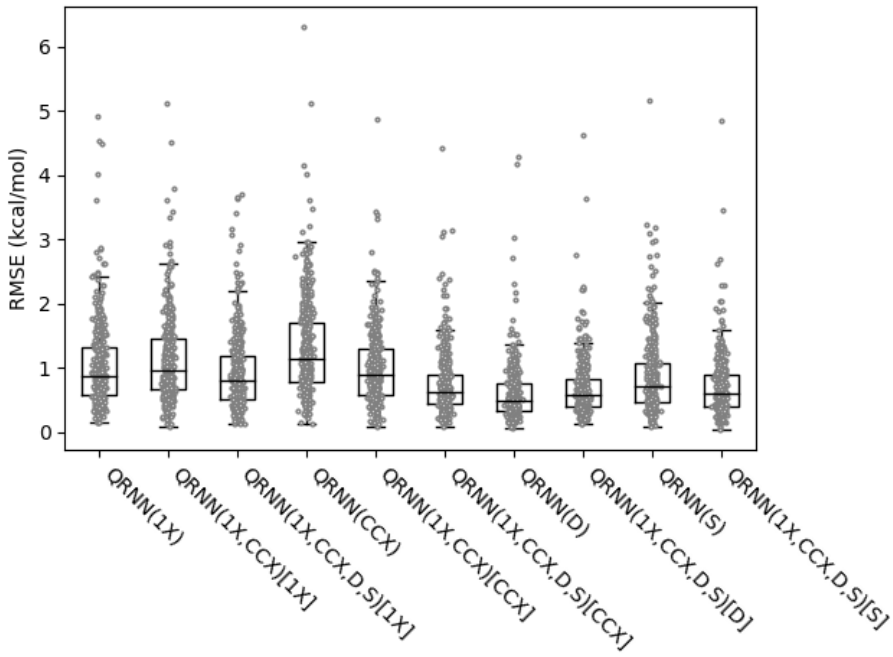

Figure 4: Root mean square deviation (RMSD) of relative conformational energies over torsion scans of small organic molecules in kcal/mol relative to ωB97X-D3/def2-TZVP estimates. The box shows the first and third quartiles whereas a line is given at the median. The whiskers show the maximum and minimum points outside of the box but inside 1.5 times the interquartile range.

Denali data to 0.58 kcal/mol when training to all four datasets. The effect of adding more data is most pronounced in the CCX results where the median RMSD for QRNN(CCX), QRNN(1X, CCX)[CCX] and QRNN(1X, CCX, D, X)[CCX] declines from 1.14 kcal/mol to 0.90 kcal/mol and finally to 0.63 kcal/mol. These results show that additional data at other levels of theory significantly improve the transferability of the model.

Now we turn to investigate the effect of adding data from qualitatively different parts of chemical space. The 1X and CCX datasets focus entirely on neutral organic molecules. As such, performance for torsion scans on ionic molecules are expected to be quite poor. To validate this expectation we evaluate RMSD over the HCNO subset of torsion scans of ionic molecules that we have previously published,[12] but relabeled the reference energies at the ωB97X-D3/def2-TZVP level. The median RMSD of QRNN(1X) and QRNN(CCX) over this set of 247 torsion scans is 1.51 and 2.36 kcal/mol respectively, justifying our expectation.

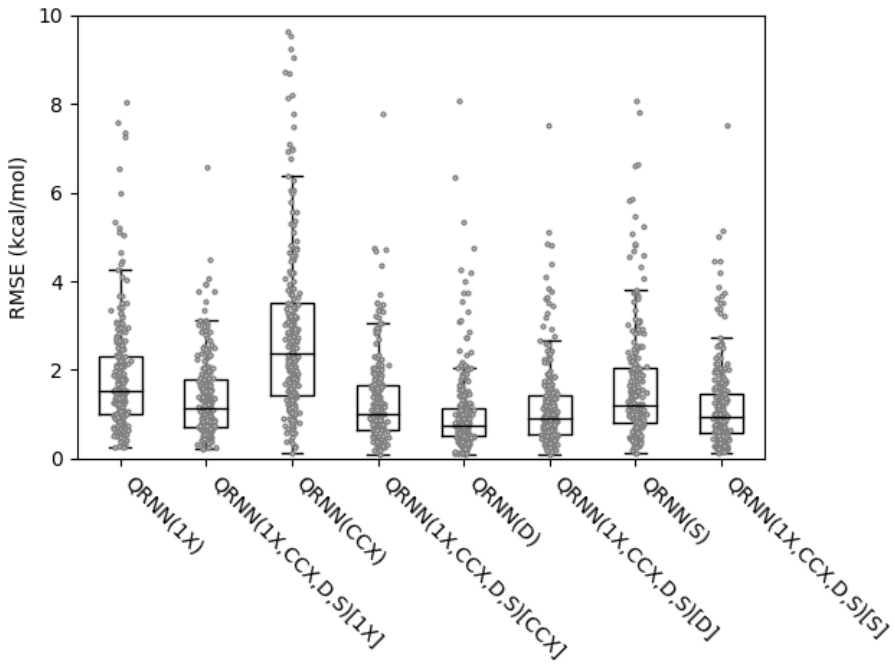

Figure 5: Root mean square deviation (RMSD) of relative conformational energies over torsion scans of ions of small organic molecules in kcal/mol relative to ωB97X-D3/def2-TZVP estimates. The box shows the first and third quartiles whereas a line is given at the median. The whiskers show the maximum and minimum points outside of the box but inside 1.5 times the interquartile range. There are seven large error cases for the QRNN(CCX) model that are not show in order to restrict the range of the plot.

Figure 5 shows that very large errors are also present; In fact the QRNN(CCX) dataset displays a maximum error of 23 kcal/mol (above the y-axis in this plot). As in the neutral torsion results we once again see that the 1X, CCX and S predictions improve when adding additional data whereas the D predictions slightly degrade. The most impressive improvement is for the CCX results whose median error improves from 2.36 for the QRNN(CCX) model to 0.99 kcal/mol for the QRNN(1X, CCX, D, S) model. Again the QRNN(D) models yields the lowest errors with a median error of 0.74 kcal/mol growing to 0.91 kcal/mol when training to all four datasets.

Finally, we investigate errors on intermolecular interactions, as probed by the popular S66x8 benchmark.[27,28] This dataset provides high level estimates (CCSD(T)-F12/CBS) of the binding energy of 66 molecular dimers, each at 8 intermolecular separations. For each of

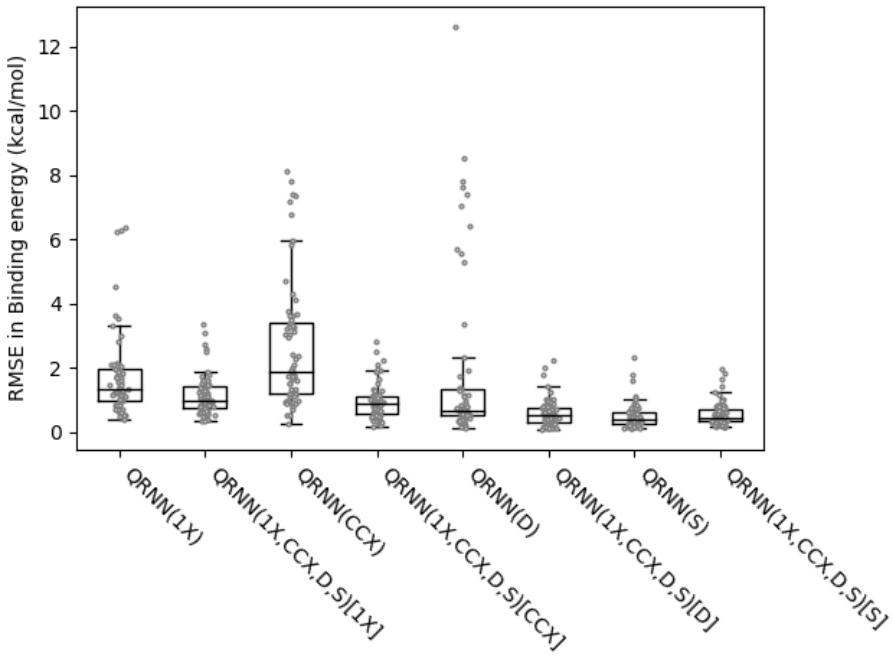

Figure 6: Root mean square deviations (RMSD) of binding energies of small dimers in the S66x8 dataset relative to reference energies at the CCSD(T)-F12/CBS level. The box shows the first and third quartiles whereas a line is given at the median. The whiskers show the maximum and minimum points outside of the box but inside 1.5 times the interquartile range.

the 66 dimers we compute RMSD in binding energy along the scan. The results are shown in Figure 6 and follow a similar trend as seen for the other tests. The predictions on 1X, CCX and D all systematically improve with addition of different datasets whereas the results on predictions of the S dataset slightly worsens. More specifically, the median error for CCX predictions using the QRNN(CCX) model is 1.89 kcal/mol and decreases significantly to 0.89 kcal/mol in the QRNN(1X, CCX, D, S) model. Predictions on the S dataset give the best results and training only to that dataset yields a median error of only 0.42 kcal/mol whereas training to all datasets and evaluating on the S output head yields a median error only negligibly worse, at 0.43 kcal/mol.

Table 2: Summary statistics for the performance of models on all test sets investigated. All errors are mean RMSD given in kcal/mol. The best model for each row is given in bold.

output	1X		CCX		D		S	
training sets	1	4	1	4	1	4	1	4
ZINC rotamers	1.03	1.00	1.32	1.00	**0.60**	0.69	0.86	0.70
ionic rotamers	1.85	1.38	3.08	1.28	**1.04**	1.18	1.64	1.21
S66x8	1.71	1.15	2.66	0.96	1.78	0.61	**0.51**	0.58
overall	1.53	1.17	2.35	1.08	1.14	**0.82**	1.01	0.83

4 Discussion

Table 2 summarizes the multitask training results presented in Sec. 3.2 for training to one or four datasets. In every case the performance of predictions on the 1X and CCX output heads improves when training to four datasets as opposed to one. For tasks that are not represented in the training set of a particular output head (ionic and intermolecular interactions for 1X and CCX) the improvement is significant. In each of the three datasets a model trained to a single level of theory performs best (D or S) and worsens when performing multitask training. However, if we average the errors over the three tasks with equal weight we find that the best model overall is a multitask model. The 1X and CCX datasets primarily focus on single, neutral molecule conformational energies and these results highlight the fact that information can be transferred amongst different output heads by utilizing datasets covering different regions of chemical space and different interactions. This may sacrifice some accuracy on an output head for a particular task that is already well represented in a dataset. Certainly, providing all data at the targeted level of theory would be superior, but incorporating freely available datasets also provides significant value when building a foundational model.

There are a few mechanisms by which information can be transferred between the output heads during training. The first is the most direct, which is that the latent representation of the input molecule is improved and this representation is shared amongst all of the output heads. This improved description can improve predictions on all output heads. Another

mechanism is the improvement of the description of the point charges, which are shared by all output heads. Two of the datasets we train to (CCX and D) do not contain dipole moments to train to and thus the charges are only trained via minimization of energy errors. Datasets that add dipole moments may give improved point charges and thus improved performance. The fact that the predictions on the D output head are often the best and this dataset does not have any dipole moments to train to suggests that this is a secondary effect, though it may be quite important for prediction of long range energies probed by the S66x8 dataset, for which the single-task D model is poor.

The output heads used in this work only perform a linear transformation of the latent space to the final output. This implies that the various outputs are also linearly related. At first glance it seems quite surprising to suggest that two levels of theory as seemingly different as CCSD(T) and ωB97X, based on dramatically different physical assumptions, could be linearly related. A few factors help to explain this. First, one can formally relate two levels of theory using parameters from the multitask model, but doing so requires that the weight matrices of the final layer are invertible, which may not be the case; any parameters from the latent space which are given zero weight by one output head and nonzero weight by another will lead to a non-invertible mapping. Further, the transformation operates on atomic energies, which are then summed, so in order to transform between levels of theory by this means one must possess a reasonable decomposition of the total energy into atomic contributions. These atomic energies are nonlinear functions of the atomic positions and may be difficult to determine. It therefore seems unlikely to us that it would be generally possible to find a transformation between levels of theory given only pairs of molecular energies.

We have seen some evidence that the linear transformation just discussed is not sufficiently flexible. In three cases we see that training to a single dataset can outperform the multitask model on a given task. This indicates that the multitask model does not have sufficient capacity when training to datasets that yield models that perform worse on that particular task. On the other hand, as one increases the expressiveness of the output trans-

formation the amount of information shared between heads will likely decrease, resulting in less information transfer. One must therefore balance this trade off.

There are three limitations worth mentioning. Our models utilize point charges to account for long range interactions and to distinguish molecules of differing net charge[12] and in this article we do not differentiate charges by level of theory; our model predicts only one set of point charges and multiple energies, essentially assuming that all DFT functionals yield approximately equivalent densities. Future work may distinguish charges between different levels of theory. Second, we do not define a mechanism besides the latent space for sharing information between output heads. This may be of interest if training to two levels of theory that are highly correlated, for example one training set which utilizes some approximation to four index integrals and one that does not. In such a case harmonically tethering the parameters of the outputs may be a successful strategy. The final limitation is that the chemical elements found in any dataset should ideally be present in all datasets. In an HDNNP trained with separate neural networks for each chemical element this seems to be a requirement for successful prediction on molecules containing a specific element. It is possible that models utilizing an atom embedding featurization scheme could infer the behavior of elements on a level of theory for which there is no training data.

5 Conclusions

In this article we have presented a conceptually simple procedure to incorporate datasets computed at different levels of theory into the training of a transferable MLFF. Due to the different interests and preferences of independent researchers, each new dataset that enters the public domain will contain complementary and incomplete data relative to what is needed to train a universal MLFF. These datasets will likely continue to be published utilizing different density functionals, basis sets and approximations. We believe that a method similar to what we have discussed in this article will be widely adopted when training foundational

MLFFs in the future. We have demonstrated that one can transfer knowledge of different tasks, such as the description of the conformational energies of ions and intermolecular interactions, when adding datasets of different levels of theory. That is, the different datasets do not need to share the same chemical space.

Acknowledgement

We would like to thank Adrian Roitberg, Alex Urban and Nong Artrith for many useful discussions. This work was supported by Gates Ventures.

Supporting Information Available

All geometries used to analyze errors statistics in the main text, with energy labels for the trained models and reference energies, where appropriate, are given in supplementary_data.tar.gz. Each test set is given as a separate json file and a text file titled README.txt describes the contents. Original references for each test set are given.

References

(1) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(2) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037–10072.

(3) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(4) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(5) Schütt, K. T.; Unke, O. T.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *ArXiv* **2021**,

(6) Thölke, P.; Fabritiis, G. D. Equivariant Transformers for Neural Network based Molecular Potentials. International Conference on Learning Representations. 2022.

(7) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, E., Tess; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Comm.* **2022**, *13*, 1453.

(8) Batatia, I.; Kovacs, D. P.; Simm, G. N. C.; Ortner, C.; Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. Advances in Neural Information Processing Systems. 2022.

(9) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. General-Purpose Machine Learning Potentials Capturing Nonlocal Charge Transfer. *Acc. Chem. Res.* **2021**, *54*, 808–817.

(10) Xie, X.; Persson, K. A.; Small, D. W. Incorporating Electronic Information into Machine Learning Potential Energy Surfaces via Approaching the Ground-State Electronic Energy as a Function of Atom-Based Electronic Populations. *Journal of Chemical Theory and Computation* **2020**, *16*, 4256–4270, PMID: 32502350.

(11) Zubatyuk, R.; Smith, J.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a Neural Network to Attach and Detach Electrons from Molecules. *Nat. Commun.* **2021**, *12*, 4870.

(12) Jacobson, L. D.; Stevenson, J. M.; Ramezanghorbani, F.; Ghoreishi, D.; Leswing, K.; Harder, E. D.; Abel, R. Transferable Neural Network Potential Energy Surfaces for

Closed-Shell Organic Molecules: Extension to Ions. *J. Chem. Theory Comput.* **2022**, *18*, 2354.

(13) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(14) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(15) Stevenson, J.; Jacobson, L. D.; Zhao, Y.; Wu, C.; Maple, J.; Leswing, K.; Harder, E.; Abel, R. Schrodinger-ANI: An Eight-Element Neural Network Interaction Potential with Greatly Expanded Coverage of Druglike Chemical Space. *ChemRxiv* **2019**,

(16) Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., 3rd OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020**, *153*, 124111.

(17) Deng, B.; Peichen, Z.; Jun, K.; Han, K.; Bartel, C. J.; Ceder, G. CHGNet: Pretrained universal neural network potential for charge-informed atomistic modeling. *ArXiv* **2023**,

(18) Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comp. Sci.* **2022**, *2*, 718 – 728.

(19) Takamoto, S.; Shinagawa, C.; Motoki, D.; Nakago, K.; Li, W.; Kurata, I.; Watanabe, T.; Yayama, Y.; Iriguchi, H.; Asano, Y.; Onodera, T.; Ishii, T.; Kudo, T.; Ono, H.; Sawada, R.; Ishitani, R.; Ong, M.; Yamaguchi, T.; Kataoka, T.; Hayashi, A.; Charoenphakdee, N.; Ibuka, T. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nature Commun.* **2022**, *13*, 2991.

(20) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

(21) Daru, J.; Forbert, H.; Behler, J.; Marx, D. Coupled Cluster Molecular Dynamics of Condensed Phase Systems Enabled by Machine Learning Potentials: Liquid Water Benchmark. *Phys. Rev. Lett.* **2022**, *129*, 226001.

(22) Chen, M. S.; Lee, J.; Ye, H.-Z.; Berkelbach, T. C.; Reichman, D. R.; Markland, T. E. Data-Efficient Machine Learning Potentials from Transfer Learning of Periodic Correlated Electronic Structure Methods: Liquid Water at AFQMC, CCSD, and CCSD(T) Accuracy. *J. Chem. Theor. Comput.* **2023**, *0*, 0, PMID: 36730728.

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

(24) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.

(25) Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G. A.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R.; Miller, T. F., III OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **2021**, *155*, 204103.

(26) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Wang, Y.; De Fabritiis, G.; Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *ArXiv* **2022**,

(27) Brauer, B.; Kesharwani, M. K.; Kozuch, S.; Martin, M. L., Jan The S66x8 bench-

mark for noncovalent interactions revisited: explicitly correlated *ab initio* methods and density functional theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 20905.

(28) Řezáč, J.; Riley, K.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theor. Comput.* **2011**, *7*, 2427–2438.

(29) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.

(30) Cipolla, R.; Gal, Y.; Kendall, A. Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* **2018**, 7482–7491.

(31) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289–320.

(32) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J. Chem. Inf. Model.* **2020**, *60*, 3408–3415.

(33) Smith, J. S.; Zubatyuk, R.; Nebgen, B. T.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.

(34) Pevarati, R.; Zhao, Y.; Truhlar, D. G. Generalized gradient approximation that recovers the second-order density-gradient expansion with optimized across-the-board performance. *J. Phys. Chem. Lett.* **2011**, *2*, 1991–1997.

(35) Luo, S.; Zhao, Y.; Truhlar, D. G. Validation of electronic structure methods for isomerization reactions of large organic molecules. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13683.

(36) Sellers, B. D.; James, N. C.; Gobbi, A. Generalized gradient approximation that recovers the second-order density-gradient expansion with optimized across-the-board performance. *J. Chem. Inf. Model.* **2017**, *57*, 1265–1275.

(37) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics* **2017**, *115*, 2315–2372.