

# SchNetPack: A Deep Learning Toolbox For Atomistic Systems

K.T. Schütt,<sup>1,\*</sup> P. Kessel,<sup>1</sup> M. Gastegger,<sup>1</sup> K. Nicoli,<sup>1</sup> A. Tkatchenko,<sup>2,†</sup> and K.-R. Müller<sup>1,3,4,‡</sup>

<sup>1</sup>*Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

<sup>2</sup>*Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg*

<sup>3</sup>*Department of Brain and Cognitive Engineering, Korea University,*

*Anam-dong, Seongbuk-gu, Seoul 136-713, South Korea*

<sup>4</sup>*Max-Planck-Institut für Informatik, Saarbrücken, Germany*

SchNetPack is a toolbox for the development and application of deep neural networks to the prediction of potential energy surfaces and other quantum-chemical properties of molecules and materials. It contains basic building blocks of atomistic neural networks, manages their training and provides simple access to common benchmark datasets. This allows for an easy implementation and evaluation of new models. For now, SchNetPack includes implementations of (weighted) atom-centered symmetry functions and the deep tensor neural network SchNet as well as ready-to-use scripts that allow to train these models on molecule and material datasets. Based upon the PyTorch deep learning framework, SchNetPack allows to efficiently apply the neural networks to large datasets with millions of reference calculations as well as parallelize the model across multiple GPUs. Finally, SchNetPack provides an interface to the *Atomic Simulation Environment* in order to make trained models easily accessible to researchers that are not yet familiar with neural networks.

## I. INTRODUCTION

One of the fundamental aims of modern quantum chemistry, condensed matter physics and materials science is to numerically determine the properties of molecules and materials. Unfortunately, the computational cost of accurate calculations prove prohibitive when it comes to large-scale molecular dynamics simulations or the exhaustive exploration of the vast chemical space. Over the last years however, it has become clear that machine learning is able to provide accurate predictions of chemical properties at significantly reduced computational costs. Conceptually, this is achieved by training a machine learning model to reproduce the results of reference calculations given the configuration of an atomistic system. Once trained, predicting properties of other atomistic systems is generically cheap and has been shown to be sufficiently accurate for a range of applications [1–18].

A common subclass of machine learning models for quantum-chemistry are *atomistic neural networks*. There exist various architectures of these models, which can be broadly split into two categories: descriptor-based models that take a predefined representation of the atomistic system as input [19–24] and end-to-end architectures that learn a representation directly from atom types and positions [25–28].

SchNetPack provides a unified framework for both categories of neural networks. While we plan to support more architectures in the future, SchNetPack currently includes implementations for SchNet [15, 27], an end-to-end continuous convolution architecture, as well as

Behler–Parrinello networks which are based on atom-centered symmetry functions (ACSF) [19, 29] and an extension thereof which uses weighted atom-centered symmetry functions (wACSF) [24].

SchNetPack furthermore contains functionality for accessing popular benchmark datasets, training neural networks on (multiple) GPUs to predict a variety of chemical properties. It is built in an extensible manner and is implemented using the PyTorch deep learning framework.

The remainder of the paper is structured as follows. In Section II, we present how models in SchNetPack are structured and briefly review (w)ACSF and SchNet representations. Section III outlines how SchNetPack manages the training process for atomistic neural networks and gives an overview of the integrated datasets. Section IV summarizes details about the implementation, while Sections V and VI provide code examples for training an atomistic neural network and calculating a power spectrum using the interface to the Atomic Simulation Environment (ASE). Section VII presents results of SchNetPack on standard benchmarks, before we conclude and give an outlook on future extensions.

## II. MODELS

Models in SchNetPack have two principle components: *representation* and *prediction* blocks (see Figure 1). The former takes the configuration of the atomistic system as an input and generates feature vectors describing each atom in its chemical environment. The latter uses these atom-wise representations to predict the desired properties of the atomistic system. The only difference between descriptor-based and end-to-end architectures is whether the representation block is fixed or learned from data. In the following two sections, we will explain the possible

\* kristof.schuett@tu-berlin.de

† alexandre.tkatchenko@uni.lu

‡ klaus-robert.mueller@tu-berlin.de

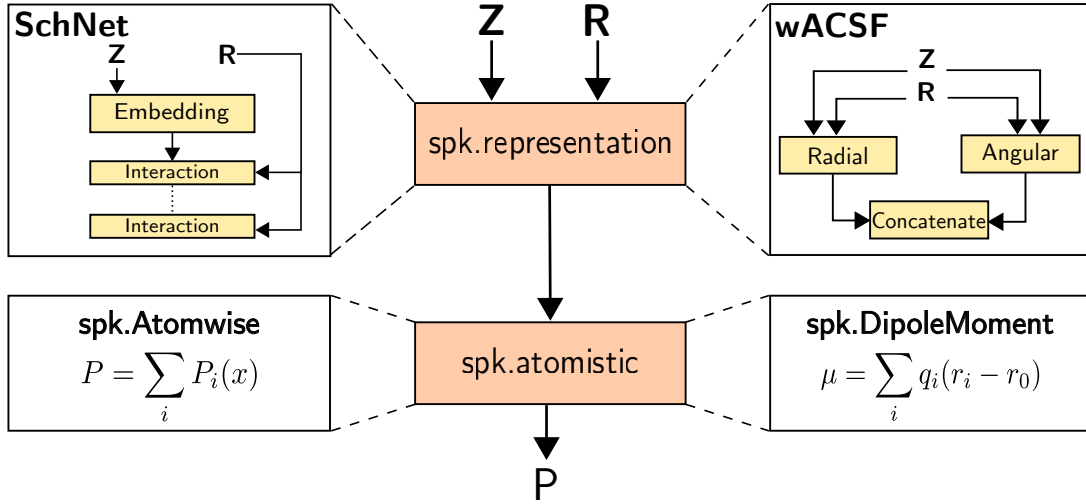


FIG. 1: Basic building blocks of a model predicting the property  $P$  from the positions  $R$  and atomic numbers  $\mathbf{Z}$  of the atomistic system. We use the abbreviation **spk** for the **schnetpack** package. All *representation* and *prediction* blocks are collected in the **spk.representation** and **spk.atomistic** package respectively. The right and left panels illustrate various choices for these building blocks.

choices for these components in detail.

### A. Representations

An atomistic system containing  $n$  atoms can be described by its atomic numbers  $\mathbf{Z} = (Z_1, \dots, Z_n)$  and positions  $R = (\mathbf{r}_1, \dots, \mathbf{r}_n)$ . The interatomic distances are given as  $r_{ij} = \|\mathbf{r}_i - \mathbf{r}_j\|$ . In the following, we will briefly describe the currently implemented representations, i.e. (w)ACSF [24] and SchNet [27]. For further details, refer to the original publications.

#### 1. (w)ACSF

Behler–Parrinello network potentials [19] have proven very useful for systems as diverse as small molecules, metal and molecular clusters, bulk materials, surfaces, water and solid-liquid interfaces (for a recent review, see [30]). Due to this impressive number of applications, Behler–Parrinello networks are now firmly established as a highly successful neural network architecture for atomistic systems.

For these networks, so-called atom-centered symmetry functions (ACSFs) form the representation of the atomistic system. Contrary to the approach taken by SchNet, where features are learned from the data, ACSFs need to be determined before training. Hence, using symmetry functions can be advantageous in situations where the available training data is insufficient to learn suitable representations in an end-to-end fashion. On the other hand, introducing rigid hand-crafted features might reduce the generality of the model. In the following, we will briefly review ACSFs and a variant called weighted

ACSFs, or wACSFs for short. We refer to References [29] and [24] for a more detailed discussion.

ACSFs describe the local chemical environment around a central atom via a combination of radial and angular distribution functions.

*a. Radial Symmetry Functions:* Radial ACSF descriptors take the form:

$$G_{i,\alpha}^{rad} = \sum_{j \neq i}^N g(Z_j) e^{-\gamma_\alpha (r_{ij} - \mu_\alpha)^2} f(r_{ij}), \quad (1)$$

where  $i$  is the central atom and the sum runs over all neighboring atoms  $j$ .  $\gamma_\alpha$  and  $\mu_\alpha$  are parameters which modulate the widths and centers of the Gaussians. Typically, a set of  $n_{\text{rad}}$  radial symmetry functions with different parameter combinations  $\alpha \in \{1, \dots, n_{\text{rad}}\}$  are used. In SchNetPack, suitable  $\gamma_\alpha$  and  $\mu_\alpha$  are determined automatically via an equidistant grid between zero and a spacial cutoff  $r_c$ , adopting the empirical parametrization strategy detailed in Reference [24].

A cutoff function  $f$  ensures that only atoms close to the central atom  $i$  enter the sum and is given by

$$f(r_{ij}) = \begin{cases} \frac{1}{2} \left( \cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1 \right), & r_{ij} \leq r_c, \\ 0, & \text{else.} \end{cases} \quad (2)$$

For convenience, we will use the notation  $f_{ij} = f(r_{ij})$  in the following. Finally,  $g(Z_j)$  is an element-dependent weighting function. In ACSFs,  $g(Z_j)$  takes the form

$$g(Z_j) = \delta_{Z_j, Z_a} = \begin{cases} 1 & \text{if } Z_j = Z_a \\ 0 & \text{else.} \end{cases} \quad (3)$$

Hence, radial ACSFs are always defined between the central atom and a neighbor belonging to a specific chemical element.

*b. Angular Symmetry Functions:* information about the angles between atoms are encoded by the  $n_a$  angular symmetry functions

$$G_{i,\alpha}^{ang} = 2^{1-\zeta_\alpha} \sum_{j \neq i, k > j}^N g(Z_j, Z_k) (1 + \lambda \theta_{ijk})^{\zeta_\alpha} \times e^{-\gamma_\alpha (r_{ij}^2 + r_{ik}^2 + r_{jk}^2)} f_{ij} f_{ik} f_{jk}, \quad (4)$$

where  $\theta_{ijk}$  is the angle spanned between atoms  $i$ ,  $j$  and  $k$ . The parameter  $\lambda$  takes the values  $\lambda = \pm 1$  which shifts the maximum of the angular terms between 0 and  $\pi$ . The variable  $\zeta_\alpha$  is a hyperparameter controlling the width around this maximum.  $\gamma_\alpha$  once again controls the width of the Gaussian functions. As with radial ACSFs, a set of  $n_{ang}$  angular functions differing in their parametrization patterns  $\alpha \in \{1, \dots, n_{ang}\}$  is chosen to describe the local environment. For angular ACSFs, the weighting function  $g(Z_k, Z_j)$  can be expressed as

$$g(Z_k, Z_j) = \frac{1}{2} (\delta_{Z_j Z_a} \delta_{Z_k Z_b} + \delta_{Z_j Z_b} \delta_{Z_k Z_a}), \quad (5)$$

which counts the contributions of neighboring atoms  $j$  and  $k$  belonging to a specific pair of elements (e.g. O-H or O-O).

Due to the choice of  $g$ , ACSFs always are defined for pairs (radial) or triples (angular) of elements and at least one parametrized function  $G_{i,\alpha}$  has to be provided for each of these combinations. As a consequence, the number of ACSFs grows quadratically with the number of different chemical species. This can lead to an impractical number of ACSFs for systems containing more than four elements (e.g. QM9).

Recently, alternative weighting functions have been proposed which circumvent the above issue. In these so-called weighted ACSFs (wACSFs), the radial weighting function is chosen as  $g(Z_j) = Z_j$  while the angular function is set to  $g(Z_k, Z_j) = Z_k Z_j$ . Through this simple reparametrization, the number of required symmetry functions becomes independent of the actual number of elements present in the system, leading to more compact descriptors. SchNetPack uses wACSFs as the standard descriptor for Behler–Parrinello potentials.

Irrespective of the choice for the weighing  $g$ , both radial and angular symmetry functions are concatenated as a final step to form the representation for the atomistic system, i.e.

$$X_i = \left( G_{i,1}^{rad}, \dots, G_{i,n_{rad}}^{rad}, G_{i,1}^{ang} |_{\lambda=\pm 1}, \dots, G_{i,n_{ang}}^{ang} |_{\lambda=\pm 1} \right). \quad (6)$$

This representation  $X_i$  can then serve as input for prediction block of the atomistic network.

## 2. SchNet

SchNet is an end-to-end deep neural network architecture based on continuous-filter convolutions [15, 27]. It

follows the deep tensor neural network framework [25], i.e. atom-wise representations are constructed by starting from embedding vectors that characterize the atom type before introducing the configuration of the system by a series of interaction blocks.

Convolutional layers in deep learning usually act on discretized signals such as images. Continuous-filter convolutions are a generalization thereof for input signals that are not aligned on a grid, such as atoms at arbitrary positions. Contrary to (w)ACSF networks which are based on rigid hand-crafted features, SchNet adapts the representation of the atomistic system to the training data. More precisely, SchNet is a multi-layer neural network which consists of an embedding layer and several interaction blocks, as shown in the top left panel of Figure 1. We describe its components in more detail in the following:

*a. Atom Embeddings:* Using an embedding layer, each atom type  $Z_i$  is represented by feature vectors  $\mathbf{x}_i^0 \in \mathbb{R}^F$  which we collect in a matrix  $X^0 = (\mathbf{x}_1^0, \dots, \mathbf{x}_n^0)$ . The feature dimension is denoted by  $F$ . The embedding layer is initialized randomly and adapted during training. In all other layers of SchNet, atoms are described analogously and we denote the features of layer  $l$  by  $X^l = (\mathbf{x}_1^l, \dots, \mathbf{x}_n^l)$  with  $\mathbf{x}_i^l \in \mathbb{R}^F$ .

*b. Interaction Blocks:* Using the features  $X^l$  and positions  $R$ , this building block computes interactions which additively refine the previous representation analogue to ResNet blocks [31]. To incorporate the influence of neighboring atoms, continuous-filter convolutions are applied which are defined as follows:

$$\mathbf{x}_i^{l+1} = (X^l * W^l) \equiv \sum_{j \in \text{nbh}(i)} \mathbf{x}_j^l \odot W^l(r_{ij}). \quad (7)$$

By  $\odot$  we denote element-wise multiplication and  $\text{nbh}(i)$  are the neighbors of atom  $i$ . In particular for larger systems, it is recommended to introduce a radial cutoff. For our experiments, we use a distance cutoff of 5 Å.

Here, the filter is not a parameter tensor as in standard convolutional layers, but a filter-generating neural network  $W^l : \mathbb{R} \rightarrow \mathbb{R}^F$  which maps atomic distances to filter values. The filter generator takes atom positions expanded on a grid of radial basis functions which are closely related to the radial symmetry functions (1) of (w)ACSF. For its precise architecture, we refer to the original publications [15, 27].

Several atom-wise layers, i.e. fully-connected layers

$$\mathbf{x}_i^{l+1} = W^l \mathbf{x}_i^l + \mathbf{b}^l \quad (8)$$

that are applied to each atom  $i$  separately, recombine the features within each atom representation. Note that the weights  $W^l$  and biases  $\mathbf{b}^l$  are independent of  $i$  and are therefore the same for all atom features  $\mathbf{x}_i^l$ . Thus the number of parameters of atom-wise layers is independent of the number of atoms  $n$ .

In summary, SchNet obtains a latent representation of the atomistic system by first using an embedding layer to

obtain features  $X^0$ . These features are then processed by  $L$  interaction blocks which results in the latent representation  $X^L$  which can be passed to the prediction block. We will sketch the possibilities for the architectures of these prediction blocks in the following section.

### B. Prediction Blocks

As discussed in the last sections, both SchNet and (w)ACSF provide representations  $X_i$  with  $i \in \{1, \dots, n\}$  for an atomistic system with  $n$  atoms. These representations are then processed by a prediction block to obtain the desired properties of the atomistic system. There are various choices for prediction blocks depending on the property of interest. Usually, prediction blocks consist of several atom-wise layers (8) with non-linearities, which reduce the feature dimension, followed by a property-dependent aggregation across atoms.

The most common choice are **Atomwise** prediction blocks, which express a desired molecular property  $P$  as a sum of atom-wise contributions

$$P = \sum_{i=1}^n p(\mathbf{x}_i). \quad (9)$$

While this is a suitable model for extensive properties such as the energy, intensive properties, which do not grow with the number of atoms  $n$  of the atomistic system, are instead expressed as the average over contributions.

**Atomwise** prediction blocks are suitable for many properties, however property-specific prediction blocks may be used to incorporate prior knowledge into the model. The **DipoleMoment** prediction block expresses the dipole moment  $\mu$  as

$$\mu = \sum_{i=1}^n q(\mathbf{x}_i)(\mathbf{r}_i - \mathbf{r}_0), \quad (10)$$

where  $q : \mathbb{R}^F \rightarrow \mathbb{R}$  can be interpreted as latent atomic charges and  $\mathbf{r}_0$  denotes the center of mass of the system.

The **ElementalAtomwise** prediction block is different from **Atomwise** in that instead of applying the same network to all the atom features  $X_i$ , it uses separate networks for different chemical elements. This is particularly useful for (w)ACSF representations. Analogously, **ElementalDipoleMoment** is defined for the dipole moment.

## III. DATA PIPELINE AND TRAINING

One of the main aims of SchNetPack is to accelerate the development and application of atomistic neural networks. To this end, SchNetPack contains a number of classes which provide access to standard benchmark datasets and manage the training process. Figure 2 summarizes this.

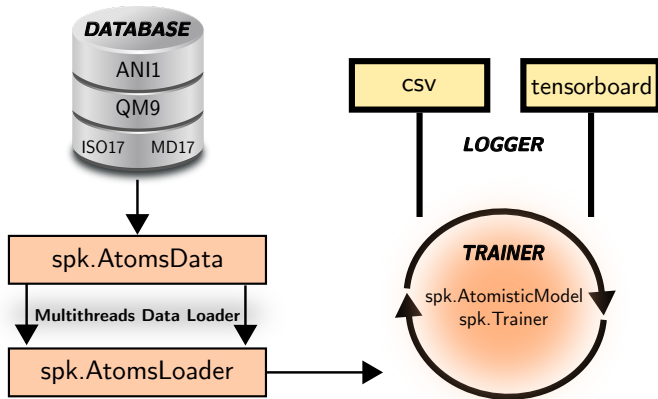


FIG. 2: Setup for training models in SchNetPack. Note that we denote the various choices for dataset classes mentioned in Section III by their common `spk.AtomsData` base class.

The dataset classes automatically download the relevant data, if not already present on disk, and use the standard ASE package [32] to store them in an SQLite database. In particular, this means that we use the conventions and units of the ASE package in SchNetPack, e.g. energies and lengths are in units of eV and Å. Currently, SchNetPack includes the following dataset classes:

- `schnetpack.datasets.QM9`: class for the QM9 dataset [33, 34] for 133,885 organic molecules with up to nine heavy atoms from C, O, N and F.
- `schnetpack.datasets.ANI1`: functionality to access ANI-1 dataset [35] which consists of more than 20 million conformations for 57454 small organic molecules from C, O and N.
- `schnetpack.datasets.ISO17`: class for ISO17 dataset [25, 27, 34] for molecular dynamics of  $C_7O_2H_{10}$  isomers. It contains 129 isomers with 5000 conformational geometries and their corresponding energies and forces.
- `schnetpack.datasets.MD17`: class for MD17 dataset [10, 25] for molecular dynamics of small molecules containing molecular forces.
- `schnetpack.datasets.MaterialsProject`: provides access to the Materials Project [36] repository of bulk crystal containing atom types ranging across the whole periodic table up to  $Z = 94$ .

We also provide a **AtomsLoader** class for feeding a model with (a subset of) a dataset during training using multiple threads. This class also calculates relevant statistics such as mean and standard deviation.

For convenience, a **Trainer** class is included in SchNetPack which manages the training process of the model. This class evaluates the model’s performance on a validation set, provides functionality for early stopping and

various learning rate schedules as well as checkpointing and logging. For the latter, one can choose between csv files and Tensorboard [37] which is a powerful web-based visualization interface. SchNetPack supports training on multiple GPUs for which we use the standard PyTorch implementation.

As we will show in the example discussed in Section V, the classes presented in this section allow us to efficiently train atomistic neural networks and evaluate their performance using a very compact amount of code.

#### IV. IMPLEMENTATION DETAILS

SchNetPack is implemented in Python using the PyTorch ( $\geq 0.4$ ) deep learning library [38]. Calculations that do not require automatic differentiation are performed using Numpy [39]. SchNetPack is tightly integrated with the Atomic Simulation Environment (ASE) [32] which is used to persist configurations of atomistic systems. We also provide an interface to the ASE calculator class which allows to easily incorporate SchNetPack models into ASE workflows, such as performing molecular dynamics. Logging the training progress to Tensorboard [37] is facilitated by tensorboardX [40]. Some of the datasets come in the HDF5 binary file format which we parse with the h5py package [41]. SchNetPack can be easily installed using pip [42]. The code for SchNetPack can be found on GitHub [43].

#### V. EXAMPLE: TRAINING IN SCHNETPACK

```

1  import schnetpack as spk
2  import schnetpack.atomistic as atm
3  import schnetpack.representation as rep
4  import torch
5  from torch.optim import Adam
6  import torch.nn.functional as F
7  from schnetpack.datasets import *
8
9  # load qm9 dataset and download if necessary
10 data = QM9("qm9/", properties=[QM9.U0])
11
12 # split in train and val
13 train, val, test = data.create_splits(10000,
14                                     1000)
15 loader = spk.AtomsLoader(train,
16                           batch_size=100,
17                           num_workers=4)
18 val_loader = spk.AtomsLoader(val)
19
20 # create model
21 reps = rep.SchNet()
22 output = atm.Atomwise()
23 model = atm.AtomisticModel(reps, output)
24
25 # create trainer

```

```

26 opt = Adam(model.parameters(), lr=1e-4)
27 loss = lambda b,p: F.mse_loss(p["y"],b[QM9.U0])
28 trainer = spk.Trainer("output/", model, loss,
29                       opt, loader, val_loader)
30
31 # start training
32 trainer.train(torch.device("cpu"))

```

Listing 1: Minimal code example for training a SchNet model on the QM9 dataset with SchNetPack.

Listing 1 is a minimal example of how to train a model with SchNet representation to predict the total energy  $U_0$  on QM9. Training and validation sets with 10k and 1k datapoints are used and the data is loaded asynchronously using four worker threads.

In order to train on a different dataset, one has to only change line 10 in Listing 1. In the example of ANI-1, it will read

```
data = ANI1("ani1/", properties=[ANI1.energy])
```

Similarly, one can straightforwardly change the representation to wACSF by replacing line 21 by

```
reps = rep.BehlerSFBBlock()
```

In this case however, it is advantageous to use a ElementalAtomwise output network by changing line 22 to

```
output = atm.ElementalAtomwise(reps.n_symfuncs)
```

These examples can also be found in the SchNetPack source directory in the examples subdirectory.

#### VI. EXAMPLE: SCHNETPACK FOR CHEMISTS

In addition to the above features, SchNetPack provides an interface to the ASE Calculator class. This makes it possible to use SchNetPack models with the calculation tools available in ASE, such as geometry optimization, normal mode analysis and molecular dynamics simulations.

The ASE interface is provided via the AseDriver class in the molecular\_dynamics module.

```

1  import schnetpack.molecular_dynamics as md
2
3  # Load trained model
4  model = md.load_model(model_directory)
5
6  ml_calculator = md.AseInterface(
7      path_to_molecule, model,
8      simulation_directory
9  )
10
11 # Optimize structure
12 ml_calculator.optimize()
13

```

```

14 # Compute numerical normal models
15 ml_calculator.compute_normal_modes()
16
17 # Setup and run molecular dynamics
18 ml_calculator.init_md()
19 ml_calculator.run_md()

```

Listing 2: Minimal code example for performing ASE calculations with a trained SchNetPack model stored in `model_directory`.

Listing 2 shows an example on how trained models are loaded into the calculator and used for computation. For convenience, SchNetPack provides the script `schnetpack_molecular_dynamics.py` which can be used to perform various simulations out of the box. To demonstrate the above features, SchNetPack was used to predict the power spectrum of the keto form of malondialdehyde via molecular dynamics simulations (shown in Figure 3). The machine learning models are able to re-

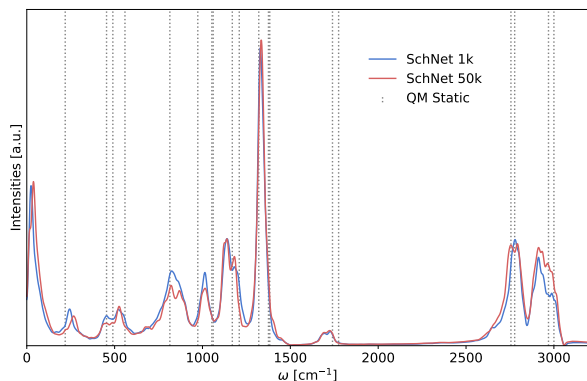


FIG. 3: Power spectra of malondialdehyde at 300 K, using SchNets trained on 1000 and 50000 data points taken from the MD17 database. The harmonic normal mode vibrations obtained with the electronic structure reference are shown in grey.

produce the peak positions accurately, even when trained on the smaller data set, demonstrating the efficacy of the force training procedure. Particularly impressive are the fine details observed in the spectra. For example, the two models are able to resolve the structure of the peak at  $1700\text{ cm}^{-1}$  which is due to the symmetric and asymmetric stretching vibrations of the two carbonyl groups.

SchNet simulations of malondialdehyde take approximately 11 milliseconds per timestep on a Tesla P100 GPU. This corresponds to a speedup of almost three orders of magnitude compared to the original electronic structure reference computations. In the present setup, Behler–Parrinello networks show a comparable performance to SchNet which indicates that both models do not yet exhaust the full capacity of the GPU for molecules of this size. It can be expected that ACSF based models are more efficient when simulating larger systems and also when using CPUs instead of GPUs.

## VII. RESULTS

In this section, we present results on QM9, ANI-1, MD17 and Material Project datasets obtained with SchNetPack. A summary of the test set performance of both Behler–Parrinello (ACSF and wACSF) and SchNet models can be found in Table I. The reported results are the average of three models trained on different splits of the same size. The Python scripts with which we obtained these results using a Tesla-P100 GPU can be found in the scripts subdirectory of SchNetPack. We refer to Appendix A for further details on the experiments.

Although Behler–Parrinello networks produce reliable results for a wide range of experiments, they are consistently outperformed by the SchNet architecture. Due to its end-to-end nature, SchNet is able to infer efficient molecular representations in a data driven fashion which leads to an improved flexibility compared to the rigid handcrafted features used in Behler–Parrinello potentials (ACSF and wACSFs). The expressive power of SchNet models is enhanced further by their deep architecture, compared to the shallow atomistic networks used in Behler–Parrinello models. These features are also advantageous for learning molecular forces for which derivatives of the energy prediction are required for training. A good example are the results obtained for the molecules malonaldehyde and acetylsalicylic acid taken from the MD17 dataset. Here, SchNet outperforms the other models even on small training sets. SchNet achieves chemically accurate performance for data sets containing a wealth of different molecular configurations (ANI-1), as well as for compounds incorporating a wide range of chemical elements, demonstrating its high utility.

The prime advantage of Behler–Parrinello models is their reduced computational cost compared to SchNet, which is expected to be beneficial e.g. for molecular dynamics simulations of large molecules. Moreover, it should be noted that all ACSF and wACSF models presented here use the empirical scheme introduced in Reference [24]. Their performance can be improved by careful fine-tuning of the descriptors. However, such a procedure is typically tedious, especially considering the excellent out of the box performance of SchNet.

An interesting effect can be observed when comparing the performance of standard ACSFs to the recently suggested wACSFs. In tasks which focus on modeling structurally and chemically diverse datasets (QM9), wACSF produce better results. However in problems for which small variations of the molecular structure need to be resolved (MD17), ACSFs outperform wACSFs. The reason for this behavior is the loss of spatial resolution of wACSF which is a direct consequence of the improved elemental resolution. Whether this problem can be circumvented by learning elemental weights in a similar manner as in SchNet will be the focus of future research.

SchNet achieves chemically accurate prediction on the ANI-1 dataset. The ANI-1 neural network potential [23], which is based on Behler–Parrinello networks, reported a



TABLE I: Summary of performance on test set. By  $N$ , we denote the size of the combined train and validation set.

Dataset	Property	Unit	Model	MAE	RMSE	time
Malondialdehyde (N=1k)	energy	$\text{kcal mol}^{-1}$	SchNet	0.08	0.11	2.5h
			ACSF	0.30	0.40	0.6h
			wACSF	1.16	1.52	0.6h
	atomic forces	$\text{kcal mol}^{-1} \text{ \AA}^{-1}$	SchNet	0.13	0.16	2.5h
			ACSF	1.08	1.59	0.6h
			wACSF	3.27	4.53	0.6h
Malondialdehyde (N=50k)	energy	$\text{kcal mol}^{-1}$	SchNet	0.07	0.09	13.5h
			ACSF	0.09	0.11	6h
			wACSF	0.69	0.88	6h
	atomic forces	$\text{kcal mol}^{-1} \text{ \AA}^{-1}$	SchNet	0.05	0.09	13.5h
			ACSF	0.26	0.42	6h
			wACSF	1.84	2.51	6h
Acetylsalicylic acid (N=1k)	energy	$\text{kcal mol}^{-1}$	SchNet	0.38	0.52	2.5h
			ACSF	0.79	1.03	0.7h
			wACSF	2.11	2.69	0.7h
	atomic forces	$\text{kcal mol}^{-1} \text{ \AA}^{-1}$	SchNet	1.17	1.68	2.5h
			ACSF	1.92	2.75	0.7h
			wACSF	4.80	6.81	0.7h
Acetylsalicylic acid (N=50k)	energy	$\text{kcal mol}^{-1}$	SchNet	0.11	0.14	2d 11.5h
			ACSF	0.40	0.53	1d 6h
			wACSF	1.20	2.69	1d 6h
	atomic forces	$\text{kcal mol}^{-1} \text{ \AA}^{-1}$	SchNet	0.14	0.19	2d 11.5h
			ACSF	0.88	1.26	1d 6h
			wACSF	2.31	3.14	1d 6h
QM9 (N=110k)	$U_0$	$\text{kcal mol}^{-1}$	SchNet	0.26	0.54	12h
			ACSF	0.49	0.92	8h
			wACSF	0.43	0.81	6h
	dipole moment	Debye	SchNet	0.020	0.038	13h
			ACSF	0.064	0.100	8h
			wACSF	0.064	0.095	8h
ANI-1 (N=10.1M)	energy	$\text{kcal mol}^{-1}$	SchNet	0.55	0.89	9d 7h <sup>a</sup>
ANI-1 (N=19.8M)	energy	$\text{kcal mol}^{-1}$	SchNet	0.47	0.77	12d 15h <sup>b</sup>
Materials Project (N=62k)	formation energy	eV / atom	SchNet	0.041	0.088	1d 14h

<sup>a</sup> We used four Tesla P100 GPUs for data-parallel training.

<sup>b</sup> We used two Tesla P100 GPUs for data-parallel training.

RMSE of  $1.2 \text{ kcal mol}^{-1}$  using 80% of the ANI-1 dataset for training and 10% for validation. Using SchNet, we already obtain a RMSE of  $0.89 \text{ kcal mol}^{-1}$  using a training set of 10 million reference examples. Raising our splits up to 80% of the whole dataset for training and 10% for validation and testing, we obtain a MAE of  $0.47 \text{ kcal mol}^{-1}$  and a RMSE of  $0.77 \text{ kcal mol}^{-1}$ .

## VIII. CONCLUSIONS

SchNetPack is a framework for neural networks of atomistic systems which simplifies accessing standard benchmark datasets, training models of different architectures and evaluating their performance. It provides an interface to combine it with the functionality of the ASE package such as molecular dynamics simulations. We plan on extending SchNetPack further in the future by adding more datasets, advanced training mechanisms such as active sampling, support for additional quantum-

mechanical observables and further neural network architectures. We expect this unification and simplification to be of great value for the community as it allows to concentrate on the design of the neural network models as well as to easily compare different architectures.

## ACKNOWLEDGEMENTS

This work was supported by the Federal Ministry of Education and Research (BMBF) for the Berlin Big Data Center BBDC (01IS14013A). Additional support was provided by the European Unions Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement NO 792572, the BK21 program funded by Korean National Research Foundation grant (No. 2012-005741). This research was also supported by Institute for Information & Communications Technology Promotion and funded by the Korea government (MSIT) (No. 2017-0-00451, No. 2017-0-01779). A.T. ac-

knowledges support from the European Research Council

(ERC-CoG grant BeStMo). Correspondence to KTS, AT and KRM.

- 
- [1] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Physical review letters* **104**, 136403 (2010).
  - [2] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).
  - [3] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller, and E. Gross, *Physical Review B* **89**, 205118 (2014).
  - [4] J. Behler, *International Journal of Quantum Chemistry* **115**, 1032 (2015).
  - [5] H. Huo and M. Rupp, *arXiv preprint arXiv:1704.06439* (2017).
  - [6] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, *The Journal of Chemical Physics* **148**, 241717 (2018).
  - [7] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, *Physical Chemistry Chemical Physics* **18**, 13754 (2016).
  - [8] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, *Proceedings of the National Academy of Sciences* **113**, 8368 (2016).
  - [9] M. Gastegger, J. Behler, and P. Marquetand, *Chemical science* **8**, 6924 (2017).
  - [10] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, *Science advances* **3**, e1603015 (2017).
  - [11] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, *Journal of chemical theory and computation* **13**, 5255 (2017).
  - [12] E. V. Podryabinkin and A. V. Shapeev, *Computational Materials Science* **140**, 171 (2017).
  - [13] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, *Nature communications* **8**, 872 (2017).
  - [14] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, *Science advances* **3**, e1701816 (2017).
  - [15] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, *The Journal of Chemical Physics* **148**, 241722 (2018).
  - [16] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, *arXiv preprint arXiv:1802.09238* (2018).
  - [17] A. Ziletti, D. Kumar, M. Scheffler, and L. M. Ghiringhelli, *Nature communications* **9**, 2775 (2018).
  - [18] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, *Physical Review Materials* **2**, 013808 (2018).
  - [19] J. Behler and M. Parrinello, *Physical Review Letters* **98**, 146401 (2007).
  - [20] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, and K.-R. Müller, in *Advances in Neural Information Processing Systems* (2012) pp. 440–448.
  - [21] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, *New Journal of Physics* **15**, 095003 (2013).
  - [22] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, *Physical review letters* **120**, 143001 (2018).
  - [23] J. S. Smith, O. Isayev, and A. E. Roitberg, *Chemical science* **8**, 3192 (2017).
  - [24] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsényi, and P. Marquetand, *The Journal of Chemical Physics* **148**, 241709 (2018).
  - [25] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, *Nature communications* **8**, 13890 (2017).
  - [26] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, *arXiv preprint arXiv:1704.01212* (2017).
  - [27] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, in *Advances in Neural Information Processing Systems* (2017) pp. 991–1001.
  - [28] N. Lubbers, J. S. Smith, and K. Barros, *The Journal of Chemical Physics* **148**, 241715 (2018).
  - [29] J. Behler, *The Journal of Chemical Physics* **134**, 074106 (2011).
  - [30] J. Behler, *Angewandte Chemie International Edition* **56**, 12828 (2017).
  - [31] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 770–778.
  - [32] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Duak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rossgaard, J. Schitz, O. Schtt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, *Journal of Physics: Condensed Matter* **29**, 273002 (2017).
  - [33] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, *Journal of Chemical Information and Modeling* **52**, 2864 (2012), PMID: 23088335, <https://doi.org/10.1021/ci300415d>.
  - [34] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, *Scientific Data* **1** (2014).
  - [35] J. S. Smith, O. Isayev, and A. E. Roitberg, *Scientific Data* **4**, 170193 (2017).
  - [36] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, *Apl Materials* **1**, 011002 (2013).
  - [37] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” (2015), software available from tensorflow.org.
  - [38] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and



- A. Lerer, (2017).
- [39] E. Jones, T. Oliphant, P. Peterson, *et al.*, “SciPy: Open source scientific tools for Python,” (2001).
  - [40] T.-W. Huang, “tensorboardX,” <https://github.com/lanpa/tensorboardX>.
  - [41] A. Collette, *Python and HDF5* (O’Reilly, 2013).
  - [42] To install run the following command: `pip install schnetpack`.
  - [43] Code can be found here: <https://github.com/atomistic-machine-learning/schnetpack>.
  - [44] L. Prechelt, in *Neural Networks: Tricks of the trade* (Springer, 1998) pp. 55–69.

## Appendix A: Details on Experiments

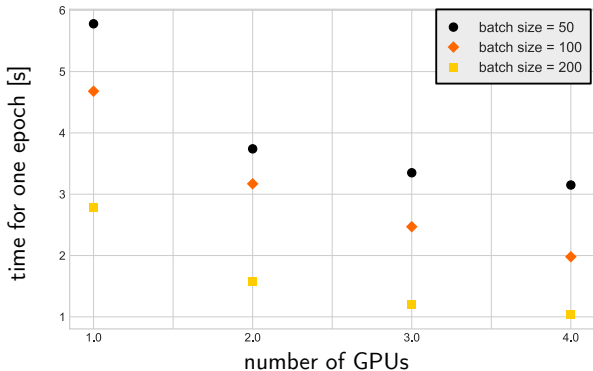


FIG. 4: Time required to go through one training epoch using 110k QM9 molecules for training.

We train three models on different splits which are summarized in Table II. The reported errors and run-times are their average. We use a Tesla P100 GPU for training. The scripts which include all the choices for the hyperparameters are contained in the scripts subdirectory of SchNetPack. One of the main reasons why we

can train SchNet models faster than in the original publications [15, 27] is that we are using a different learning rate schedule. In the original publications, the learning rate was lowered by a factor of 0.96 every 100k iterations. In our experiments however, the learning rate is reduced by a decay factor if the validation loss has not improved over a given number of epochs. We refer to Tables III to V for details on learning rates, schedules and symmetry function compositions. While both learning rate schedules lead to comparable results in the long run, our new setup converges significantly faster.

We can further speed up training by taking advantage of the support for multiple GPUs in SchNetPack as is demonstrated in Figure 4. In our experiments, we only did so for the ANI-1 dataset.

Finding the optimal point to stop the training process is a well-known problem in the optimization of neural networks [44]. It is worth noting that a significant amount of the training time in our experiments is spent on fine-tuning the accuracy of the prediction. Figure 5 demonstrates this for the example of training SchNet on 50k reference calculations for acetylsalicylic acid. In this case, approximately 50% of the training time is spent on the last 0.02 kcal/mol improvement. This also means that if this minor improvement is not required, the training can already be stopped after half of the reported training time when the learning progress flattens out. We plan to implement more advanced learning rate schedules and stopping criteria in future versions to alleviate this issue.

TABLE II: Setup of experiments

Dataset	Train	Validation	Test
Malondialdehyde (N=1k)	950	50	992,237
Malondialdehyde (N=50k)	49,000	1,000	943,237
Acetylsalicylic acid (N=1k)	950	50	210,762
Acetylsalicylic acid (N=50k)	49,000	1,000	161,762
QM9	109,000	1,000	20,813
ANI-1 (N=10.1M)	10,000,000	100,000	11,957,374
ANI-1 (N=19.8M)	17,600,000	2,200,000	2,257,374
Materials Project	60,000	2,000	21,623

TABLE III: Setup for SchNet training

Dataset	Learning Rate	Decay Factor	Minimal Learning Rate	Patience	$\rho$	Batch Size
QM9	0.0001	0.5	1e-06	25	–	100
ANI-1	0.0001	0.5	1e-06	6	–	400
MD17 (1k)	0.0001	0.5	1e-06	150	0.1	100
MD17 (50k)	0.0001	0.5	1e-06	50	0.1	100
Materials Project	0.001	0.5	1e-06	25	–	32

TABLE IV: Setup for Behler–Parrinello training

Dataset	Learning Rate	Decay Factor	Minimal Learning Rate	Patience	$\rho$	Batch Size
QM9	0.001	0.5	1e-06	25	-	100
MD17 (1k)	0.01	0.5	1e-06	20	0.1	20
MD17 (50k)	0.01	0.5	1e-06	20	0.1	100

TABLE V: Setup for ACSFs and wACSFs used in the Behler–Parrinello models. The symmetry functions were standardized in all experiments.

Dataset	Type	$n_{\text{rad}}$	$n_{\text{ang}}$	$n_{\text{SF}}$	centered radial
QM9	ACSF	5	3	115	-
	wACSF	22	5	32	-
MD17 (1k)	ACSF	5	3	51	+
	wACSF	15	18	51	+
MD17 (50k)	ACSF	5	3	51	+
	wACSF	15	18	51	+

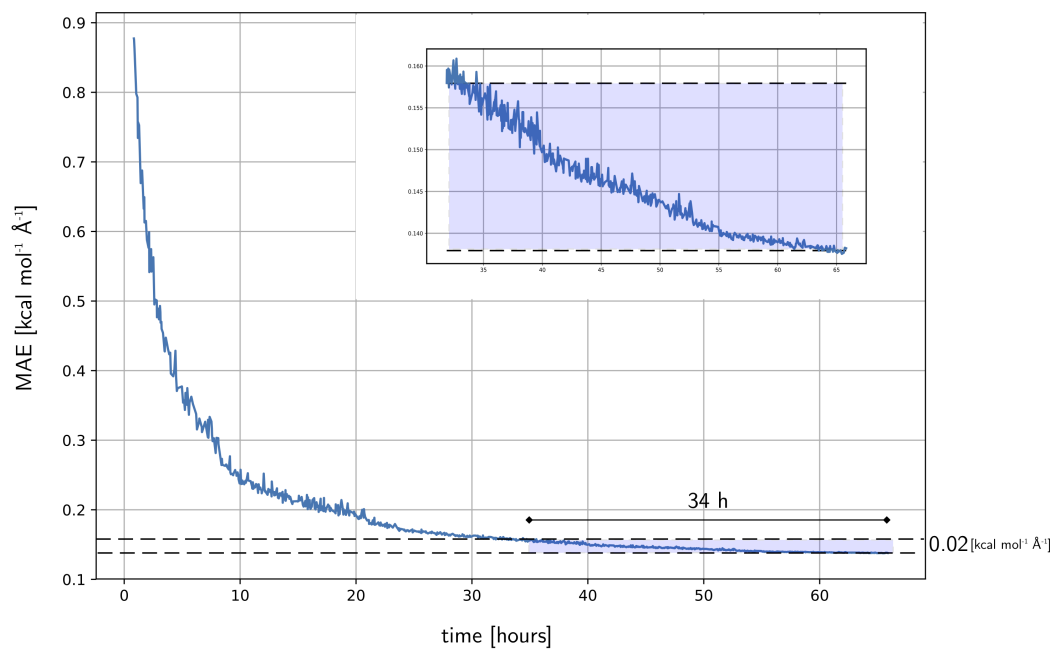


FIG. 5: Force training of SchNet on acetylsalicylic acid ( $N=50k$ ). Note that approximately 50% of training time is spend on fine-tuning the last  $\sim 0.02 \text{ kcal mol}^{-1} \text{\AA}^{-1}$ .