Project Machine Learning
SchNet: A continuous-filter convolutional neural network for modeling quantum interactions
— Milestone 1: Data Sets and Prototype —

Ariane Oesch, Christian Kasim Loan, Trung Duc Ha

November 24, 2023

In this first milestone report, our goal is to understand the paper by Schütt et al. [1] and the data, as well as develop some basic baseline methods to solve the task. First, we will describe the data and its domain. Then we will move on to explain some baseline methods we used, and finally we will discuss the results. You can find our source code here https://github.com/DevinTDHa/PML-SchNet .

1 Data Set Overview

In this section, we look at the data sets that Schütt et al. used to train their proposed network SchNet [1]. All data sets used to train SchNet can be loaded as PyTorch tensors using the SchNetPack [2] library. Just like Schütt et al., we perform no normalization or preprocessing since we want to capture minuscule state changes in molecules. Working with this kind of data is hard because of the vast number of possible chemical properties, i.e. high dimensionality of the problem and the changing shape of data because molecules can have varying amounts of atoms.

Since the datasets are generated and cleaned there are no missing values or garbage data, so we do not perform any data cleaning.

1.1 QM9

QM9 is a 133,885 molecule subset of the GDB-17 [3] chemical universe which has 166 Billion organic molecules. The approximately 134k molecules are made up of 9 heavy atoms of types C, H, O, N, F (Carbon, Hydrogen, Oxygen, Nitrogen, and Fluorine). The goal is to predict the target variable U_0, internal energy at 0 K, as kcal/mol from the molecular and atomic properties in Table 1. Since molecules are in equilibrium, the forces are zero by definition do not need to be predicted.

1.2 MD17

MD17 is a database of simulations for Molecular Dynamic (MD) trajectories. It consists of ten data sets, one for each of the following molecules: Benzene, Aspirin, Uracil, Naphthalene, Salicylic Acid, Malonaldehyde, Ethanol, Toluene, Paracetamol and Azobenzene. All trajectories are calculated at a temperature of 500 K and a resolution of 0.5 fs. One data set consists of a trajectory of tens of thousands of data points.

For each data point, it provides us the energy of the molecule in the corresponding configuration, the forces acting on atoms and the 3-dimensional coordinates of the atoms of the molecule. The atom states do not correspond to equilibria.

No.	Property	Unit	Description
1	tag	-	'gdb9' string to facilitate extraction
2	i	-	Consecutive, 1-based integer identifier
3	A	GHz	Rotational constant
4	B	GHz	Rotational constant
5	c	GHz	Rotational constant
6	μ	D	Dipole moment
7	a	a_0^3	Isotropic polarizability
8	ϵ_{HOMO}	Ha	Energy of HOMO
9	ϵ_{LUMO}	Ha	Energy of LUMO
10	ϵ_{gap}	Ha	Gap ($\epsilon_{\text{LUMO}} - \epsilon_{\text{HOMO}}$)
11	$\langle R^2 \rangle$	a_0^2	Electronic spatial extent
12	zpve	Ha	Zero point vibrational energy
13	U_0	Ha	Internal energy at 0 K
14	u	Ha	Internal energy at 298.15 K
15	H	Ha	Enthalpy at 298.15 K
16	G	Ha	Free energy at 298.15 K
17	c_v	$\frac{\text{cal}}{\text{molK}}$	Heat capacity at 298.15 K
18	_n_atoms	-	Number of Atoms
19	_atomic_numbers	-	X,Y,Z coordinates of atoms. 298.15 K

Table 1: Description of properties in the QM9 data set. U_0 is the label, `_n_atoms` and `_atomic_numbers` are used as features for the model

Molecule	Number of points
Benzene	627,983
Uracil	133,770
Naphthalene	326,250
Aspirin	211,762
Salicylic acid	320,231
Malonaldehyde	993,237
Ethanol	555,092
Toluene	442,790
Paracetamol	106,490
Azobenzene	99,999

Table 2: Data points for different molecules in MD17

When we look at the distribution of energies in a data set, we get a curve close to a Gaussian. In Figure 1 we can see an example for ethanol. The observation is valid for all the molecules included in MD17. Therefore, it seems that most of the spatial configurations of the molecule give energy close to the average energy of the molecule, which could perhaps correspond to the most stable configuration.

Furthermore, having data that follows a Gaussian distribution can facilitate learning and model convergence and also reduce systematic errors due to outliers. It also makes it easier to interpret the results in comparison to the initial distribution.

1.3 ISO17

The SchNet paper also introduces the ISO17 data set [1]. It can be seen as a mix of the QM9 and MD17 data sets. It consists of 129 isomers of the molecule $C_7O_2H_{10}$. The included isomers were drawn randomly from the largest ones of the QM9 data set. For each, the data set includes short simulated MD trajectories. In addition, it contains molecules in equilibrium, i.e. a subset of the molecules have zero forces. The mix of these molecules makes it the most challenging data set to work with. In total, ISO17 consists of 5000 conformations for each of the 129 molecules, so 645,000 examples to predict the energy and inter-atomic forces. The data is partitioned, as can be seen in Table 4.

Items	Unit	Description
_idx	–	Integer identifier
energy	kcal/mol	Internal energy
forces	kcal/mol/Å	Array of forces acting on each atom
_n_atoms	–	Number of atoms
_atomic_numbers	–	Number of charges in each atom
_positions	Å	Array of 3-dimensional coordinates of atoms
_cell	–	Unit cell
_pbc	–	Periodic boundary conditions

Table 3: Features of MD17. Energy and Forces are the labels. We use the number of atoms in the molecule, atomic charge, and 3D coordinates of the atoms as features for the model.

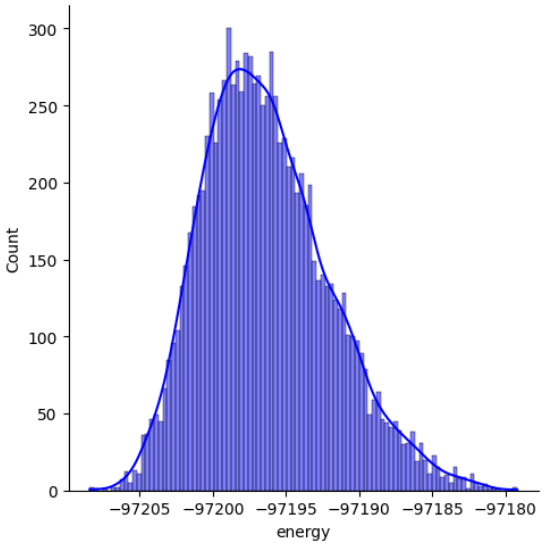


Figure 1: Distribution of the energy for the ethanol molecule

In Figure 2 we can see the distribution of energy values for the reference data set. The curve looks like the sum of two Gaussians, one with an average energy around −11504 kcal/mol and the other with an average energy around −11502 kcal/mol. Therefore, there might be two different sub-populations with different levels of energy of isomers in the reference data set.

In Figure 3, we can see the distribution of the interatomic forces for the X, Y, and Z coordinates. It appears that they are following a Gaussian distribution with mean zero.

1.4 Data Loader and Visualization

The data loader is implemented by wrapping SchNetPack [2] data set classes with our load data function, which gives easy access to train and test generators for QM9, MD17, and ISO17 and provides a parameter to specify which molecule class to load for MD17. No additional transformations are performed.

For the visualization of a data point, we implement a function `show`. The atom's positions with their respective numbers and the energy of the molecule are taken as input to produce a three-dimensional plot of the molecule state. Optionally, we can provide the forces acting on each atom to also be displayed. An example of a state of the MD17 aspirin molecule can be seen in Figure 4[1].

[1]The data points from MD17 are a time series of simulated MD. Therefore it is possible to create an animation of such plots. An example can be seen here.

Fold	Description
reference	80% of steps of 80% of Molecular Dynamics trajectories
reference_eq	only equilibrium conformations of the molecules
test_within	the 20% steps that are unseen in the reference fold
test_other	the remaining 20% of MD trajectories
test_eq	the equilibrium conformations of test trajectories

Table 4: Partition of ISO17 data

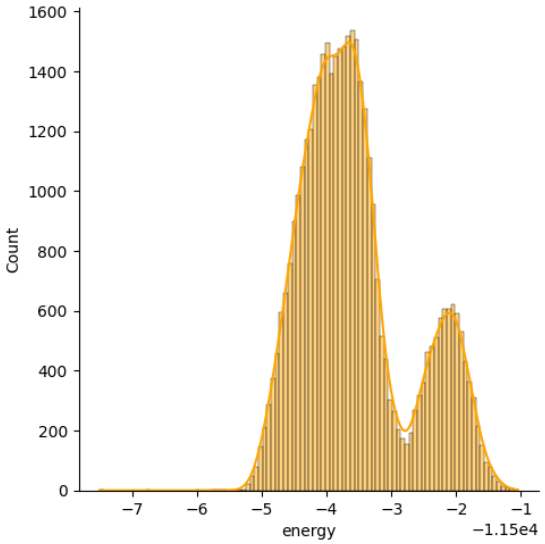


Figure 2: Distribution of the energy in the ISO17 reference data

2 Baseline Method and Evaluation

For the baseline method, we decided to go for simple fully connected neural networks. The goal was to see what a naive approach to prediction could yield. Naturally, this does not consider the quantum-mechanical limitations that the model needs to adhere to for actual prediction tasks. We create a model that is capable of accommodating molecules that have a *variable* of number of atoms. This flexibility is crucial for dealing with the diverse molecular structures present in these data sets such as QM9.

We introduce two types of feature vectors. First, the nuclear charges of the atoms are embedded in a fixed 8-dimensional space. As the number of atoms can vary, this enables us to treat varying molecules in the same feature space. This is done by treating each vector of nuclear charges as a tuple to index a specific embedding. The SchNet neural network also maps the atom charges in this fashion.

We create a feature vector of the atom positions by propagating them through two hidden linear layers equipped with a ReLU activation to create 16-dimensional "spatial" embeddings. As we have linear layers, these spatial embeddings can handle molecules with arbitrary amounts of atoms, since each atom is embedded individually. This makes our model able to handle inputs of varying length. Both feature vectors are concatenated and passed to a final linear layer with a single output to predict the energy.

2.1 Loss Functions and Model Training

In total, we have 3 different tasks for which we can construct a separate loss function. First, we can measure the mean absolute error from the predicted energy and the target energy, as it is also used by Schütt et al. Second, we can apply the same loss function to measure the target forces and the predicted forces. We derive the predicted forces by taking the partial derivative of the predicted energy w.r.t. the atom positions. Lastly, we implement the loss function of

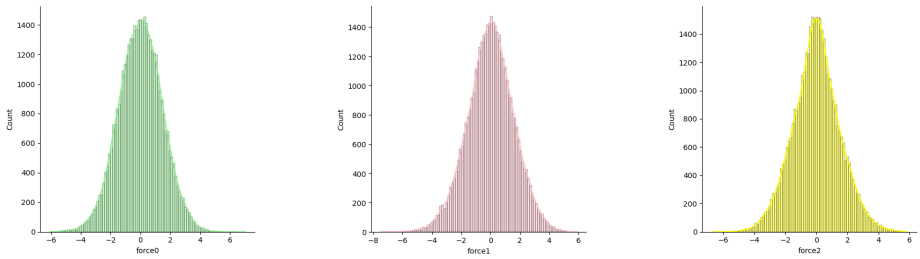


Figure 3: Distributions of the X, Y and Z dimensions of forces in ISO17 reference data

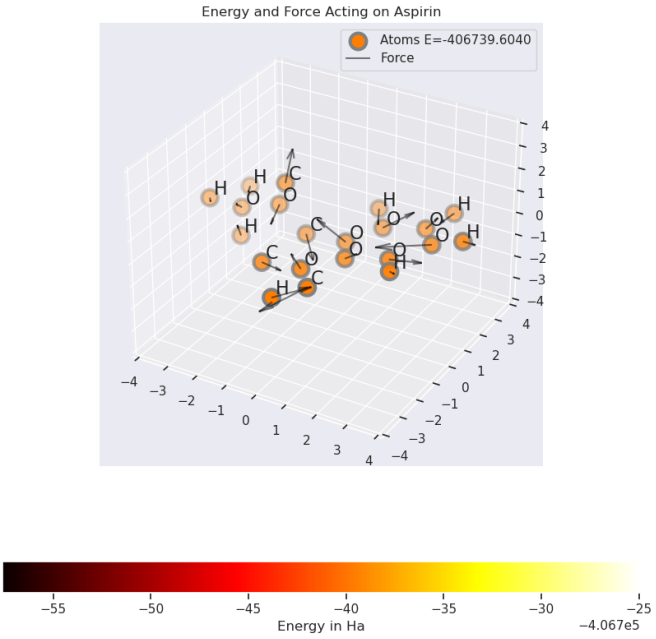


Figure 4: 3D Visualization of Aspirin

SchNet, which is defined by both energy and the forces acting on each atom. It is defined as the sum of distances between the predicted energy and its derived force[1]:

$$\ell(\hat{E}, (E, \mathbf{F}_1, \ldots, \mathbf{F}_n)) = \rho \|E - \hat{E}\|^2 + \frac{1}{n} \sum_{i=0}^{n} \left\| \mathbf{F}_i - \left(-\frac{\partial \hat{E}}{\partial \mathbf{r}_i} \right) \right\|^2 \tag{1}$$

In Eq. 2.1 $\hat{E}$ is the predicted energy, E is the target energy, ρ is a parameter that was empirically estimated at 0.01 [1] and $\mathbf{F}_i$ is the force acting on atom i with its respective position $\mathbf{r}_i$. The predicted energy corresponds to the output of the network.

We train a different model for each task and each data set. As previously described, we do not pre-process the data to recreate the conditions for SchNet. As we will see in the results, this makes the problem especially difficult. The magnitudes of the initial network and the target energies are very large.

We train the models for one epoch with 10000 train data points, 100 test data points, in mini-batches of size 32.

2.2 Evaluation

An overview of the baseline results can be seen in Table 5. It is apparent that the amount of training was not enough to achieve a sufficiently low loss. This is because of the different magnitudes in the prediction and target energies; As the weights of the network are initialized with

a Gaussian distribution with mean 0, the network will take a long time to reach the magnitudes of the target labels (up to -10^5).

However, the loss for the force predictions are comparatively low, even reaching levels of the baseline in the paper by Schütt et al for the ISO17 data set. We were able to see in the analysis in Section 1.3, that the forces roughly follow a Gaussian distribution around zero. Hence, we were able to reach lower validation losses in that task.

To our surprise, the QM9 model is converging. We would have expected them to perform worse than MD17, since the molecules can have dynamic shapes, which should make the learning task harder.

We note that the training parameters did not enable the loss of the models to converge to a stable state (with the exception of force prediction of ISO17). We attempted model training with a higher number of training points and epochs and noticed a further decrease in loss, indicating further potential in the baseline models.

Data Set	Energy Loss	Force Loss	Energy+Force Loss
QM9	5.64	n.a.	n.a.
MD17 - Aspirin	2709.31	19.53	25767678
MD17 - Azobenzene	155617.57	21.74	299110688
MD17 - Benzene	126340.88	10.94	499045536
MD17 - Ethanol	25041.22	18.02	58639376
MD17 - Malonaldehyde	41752.62	19.90	45739288
MD17 - Naphthalene	76627.39	19.52	77312560
MD17 - Paracetamol	144221.46	20.02	1221430656
MD17 - Salicylic Acid	39732.83	20.26	43561572
MD17 - Toluene	79319.87	19.52	89748176
MD17 - Uracil	76933.86	20.05	114911304
ISO17	129.48	0.82	28243.96

Table 5: Losses on test data for energy, force, and energy+force tasks 10,000 train and 100 test samples

3 Discussion

3.1 Challenges of the data sets

Although QM9 is relatively large, it offers a limited diversity of molecules. It contains around 133,000 organic molecules, which is quite large, but given the infinite diversity of possible organic molecules, this sampling remains finite and does not capture all possible chemical diversity. However, it remains a challenge as a learning task, since we have many different molecules with different numbers of atoms.

Furthermore, it focuses mainly on small and medium-sized organic molecules. It would be difficult to represent larger or more complex molecules in such detail. This can present challenges for the generalization of the model to types of molecules not present in the data set.

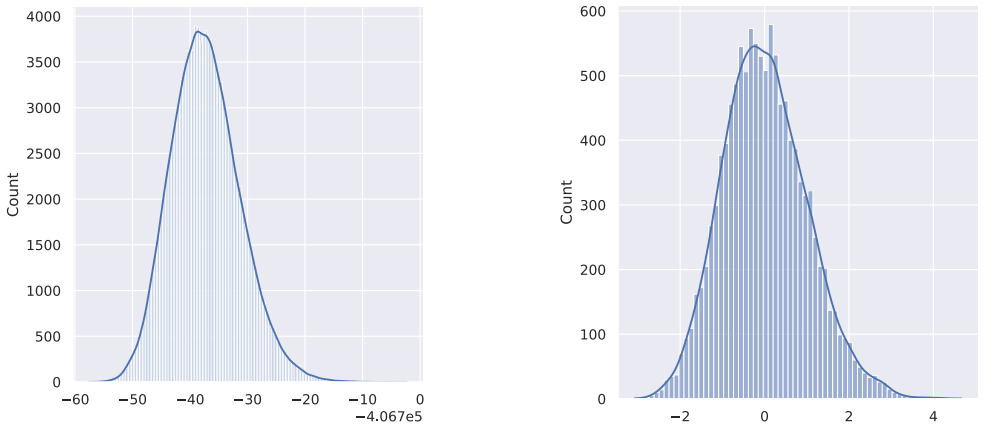
The main challenge of the MD17 data set lies in the dynamic nature of the data. In fact, the data are time trajectories of 10 molecules. This means that energy and force measurements can depend not only on atomic positions at a given moment but also on subsequent interactions and dynamics. Therefore, the prediction of molecular properties can be more complex in the case of MD17, whose data is evolutionary in nature, than when using static data as in QM9. The same applies to ISO17, whose data are derived from MD trajectories. ISO17 is the most challenging one because it also includes chemical changes, and therefore the data is more complex and more difficult to predict.

Additionally, in all data sets the label can have very large values, making loss larger and convergence slower.

3.1.1 Normalization of Data

As the magnitude of the prediction and target labels are very far apart, we tried to adjust it before training as a separate experiment. This procedure was not done by SchNet.

We take the MD trajectories of aspirin as the basis. We choose to center and normalize the training data so that the values behave better in the network. Training consists of the same procedure as the baseline. The result is a model that predicts a *Gaussian distribution* seen in Figure 5 that is similar to the analysis of Section 1.2. Compared to the other baselines, the training as well as the final validation loss is fairly low. As we are initializing the network with Gaussian distributed values, the network only needs to learn small variations of the target distribution.



(a) Distribution of Energy for Aspirin

(b) Distribution of Energy for the Baseline

Figure 5: Comparison of the energy distribution for the data and the baseline model

3.2 Use-Cases

De Novo design of new drugs, materials, and chemical compounds is one of the main applications that can be built on top of data sets like QM9, MD17, and ISO17 [4]. It describes the process of creating something from scratch, without relying on existing models or frameworks. In the context of drugs, the molecules should interact with biological targets such as proteins, and DNA or treat disease. Material design focuses on creating new materials with desired properties such as strength, flexibility, or conductivity. Generally, factors like molecular shape, electronic properties, and potential biological activity need to be considered. Typically, a human query would be to find molecules with certain properties that satisfy product demands.

3.3 Future Work

Visualizing decision boundaries of activation functions and explaining predictions would help to develop a better understanding of the data and model. Additionally plotting bonds of molecules can aid in understanding the data. Finally, we would like to apply similar structured problems from different domains like astronomy and financial arbitrage graphs.

References

[1] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," no. arXiv:1706.08566, Dec. 2017. arXiv: `1706.08566 [physics, stat]`. (visited on 11/16/2023).

[2] K. T. Schütt, S. S. P. Hessmann, N. W. A. Gebauer, J. Lederer, and M. Gastegger, "SchNetPack 2.0: A neural network toolbox for atomistic machine learning," *The Journal of Chemical Physics*, vol. 158, no. 14, p. 144 801, Apr. 2023, ISSN: 0021-9606. DOI: 10.1063/5.0138367. eprint: https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0138367/16825487/144801\textbackslash_1\textbackslash_5.0138367.pdf.

[3] L. Ruddigkeit, R. van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2864–2875, Nov. 2012, ISSN: 1549-9596. DOI: 10.1021/ci300415d. (visited on 11/20/2023).

[4] M. Sun, S. Zhao, C. Gilvary, O. Elemento, J. Zhou, and F. Wang, "Graph convolutional networks for computational drug development and discovery," *Briefings in Bioinformatics*, vol. 21, no. 3, pp. 919–935, May 2020. DOI: 10.1093/bib/bbz042.