AI4Science: Neural Networks for Molecular Property Prediction

Zhiming Xu
zx2rw@virginia.edu

UNIVERSITY of VIRGINIA | ENGINEERING

School of Engineering and Applied Science
University of Virginia

September 17, 2021

Table of Contents

Atom, Molecule & Chemical Bond

- *Atom*: the smallest unit of ordinary matter that forms a chemical element, composed of a nucleus and one or more electrons.
- *Molecule*: an electrically neural group of two or more atoms held together by chemical bonds.
- *Chemical bond*: an attractive force between atoms, ions, or molecules that enables the formation of chemical compounds.

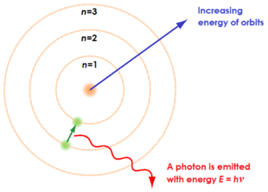

Figure 1: Bohr's model.

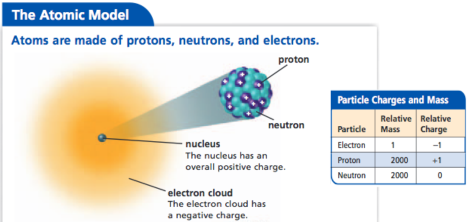

Figure 2: Quantum mechanics' model.

- Molecules can take different shapes, depending on the chemical bonds as well as non-bond forces, such as electrostatic attraction/repulsion.
- For chemical bonds, they can have different lengths and form various angles. The following figures show them in an ammonia and a methane molecule (both from Wikipedia).

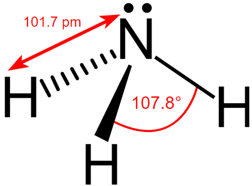

Figure 3: Shape of NH_3

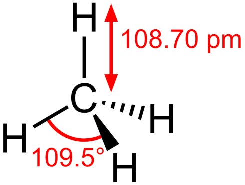

Figure 4: Shape of CH_4.

Table of Contents

Quantum Mechanics

- Since atoms are too small, Newtonian mechanics do not work on it. They have the *wave-particle duality*, and their behavior can be described by *wave functions*.
- Specifically, suppose that a quantum system, such as the electron of a hydrogen, is represented by the wave function Ψ, then we have (*time-dependent*) *Schrödinger equation*

$$
\begin{aligned}
i\hbar \frac{\mathrm{d}}{\mathrm{d}t} |\Psi(t)\rangle &= \hat{H} |\Psi(t)\rangle, \\
\hat{H} &= \left(-\frac{\hbar^2}{2m} \frac{\partial^2}{\partial^2 x} + V(x,t) \right).
\end{aligned}
$$

The probability of finding this electron in position x at time t, i.e., $\Pr(x,t)$, equals the square of the wave function's modulus $|\Psi(x,t)|^2$.

Time-independent Schrödinger Equation

If we only consider a stationary quantum system that does not change over time, then the derivative w.r.t time t should be 0.

$$i\hbar \frac{\mathrm{d}}{\mathrm{d}t}|\Psi(t)\rangle = 0.$$

Therefore, the right side of time-dependent Schrödinger equation also equals 0

$$\hat{H}|\Psi(t)\rangle = \left(-\frac{\hbar^2}{2m}\frac{\partial^2}{\partial^2 x} + V(x,t)\right)|\Psi(t)\rangle = 0.$$

Time t becomes irrelevant here and can be eliminated. As a result, we have the (*time-independent*) *Schrödinger equation*

$$\hat{\mathrm{H}}\,|\Psi\rangle = E|\Psi\rangle.$$

Time-independent Schrödinger Equation (cont.)

Remarks on *time-independent Schrödinger equation*

$$\hat{\mathrm{H}}\,|\Psi\rangle = E|\Psi\rangle.$$

- It is an eigenvalue equation. Specifically, Ψ is the eigenfunction of the linear operator $\hat{\mathrm{H}}$, with corresponding eigenvalue(s) E.
- It is intractable in current computational technology except for a single hydrogen atom. However, if $\hat{\mathrm{H}}$ is time-independent, the wave function Ψ can be written as $\psi(\mathbf{r})\psi(t)$, and solved in certain cases.
- It is linear. If ψ_1 and ψ_2 are solutions to it, then any linear combination of them, $\psi = \alpha\psi_1 + \beta\psi_2$, is also a solution.
- It implies that modeling a molecule as still, rigid spheres (atoms) connected by fixed-length edges (bonds) is inaccurate.

Motivations to Use NNs for Molecules

- The exact calculation of Schrödinger equation is prohibitively hard. Many theories have already been proposed to give approximate solutions, such as density function theory.
- Neural networks are good at approximating functions. Therefore, they can be used to learn equations reflecting underlying physics from data, and hence substitute them.

Table of Contents

Input

- Nuclear charges $\mathbf{Z}$.
- Pairwise distances $\mathbf{D}$.

Structure

- Atom embedding.
- Distance expansion.
- Interaction.
- Individual contribution.
- Summation.

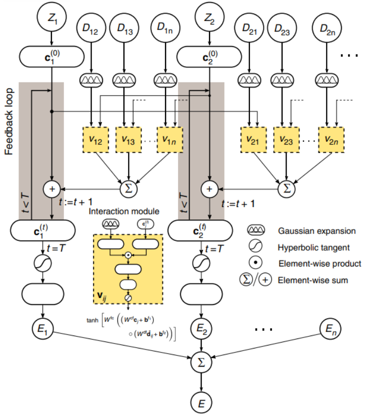

Figure 5: Overall framework of DTNN.

- Atom embedding.
  Randomly initialized vector for each kind of elements.

$$\mathbf{c}_i^{(0)} = \mathbf{c}_{\mathbf{Z}_i} \in \mathbb{R}^B$$

- Gaussian expansion of the atom-wise distances[1].

$$\hat{\mathbf{d}}_{ij} = \left[\exp\left(-\frac{(\mathbf{D}_{ij} - (\mu_{\min} + k\Delta\mu))^2}{2\sigma^2}\right)\right]_{k\in\{0,1,\cdots,\mu_{\max}/\Delta\mu\}}$$

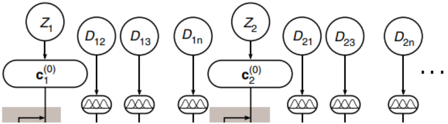

Figure 6: Atom embedding and distance expansion.

[1] This kind of functions defined only on distance is called *radial basis function* (RBF).

- Interaction (T passes in a row).

$$\mathbf{c}_i^{(t+1)} = \mathbf{c}_i^{(t)} + \sum_{j \neq i} \mathbf{v}_{ij}.$$

$\mathbf{v}_{ij}$ is the message passed to atom i from j in the form of

$$\mathbf{v}_{ij} = \tanh\left[\mathbf{W}^{\text{cf}}\left(\left(\mathbf{W}^{\text{fc}}\mathbf{c}_j + \mathbf{b}^{\text{f}_1}\right) \circ \left(\mathbf{W}^{\text{df}}\hat{\mathbf{d}}_{ij} + \mathbf{b}^{\text{b}_2}\right)\right)\right].$$

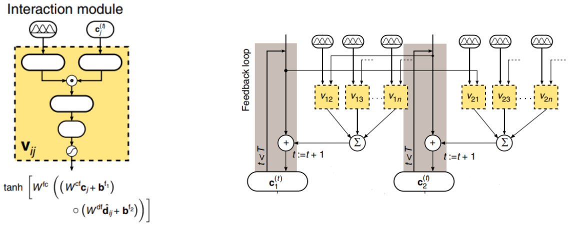

Figure 7: Interaction module of DTNN. It loops for T times.

DTNN – Aggregation & Prediction

- Individual contribution.

$$\mathbf{o}_i = \tanh\left(\mathbf{W}^{\text{out}_1}\mathbf{c}_i^{(T)} + \mathbf{b}^{\text{out}_1}\right)$$
$$\hat{E}_i = \mathbf{W}^{\text{out}_2}\mathbf{o}_i + \mathbf{b}^{\text{out}_2}$$

  Additionally, to scale the output range, $\hat{E}_i$ predicts the shifted value. To bring it back, $E_i = E_\sigma \hat{E}_i + E_\mu$.

- Summation to obtain the total molecular energy.

$$E = \sum_i E_i$$

Table of Contents

Input

- Nuclear charges $\mathbf{Z}$.
- Positions $\mathbf{R}$.

Structure

- Atom embedding.
- Atom-wise layers.
- Interaction.
- Filter generation.
- Property prediction.

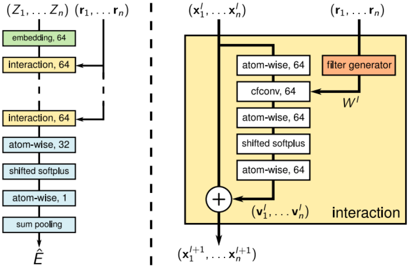

Figure 8: Overall framework of SchNet.

SchNet – Atom Embedding

- Atom embedding.
  Randomly initialized vector for each kind of elements.

$$\mathbf{c}_i^{(0)} = \mathbf{c}_{\mathbf{Z}_i} \in \mathbb{R}^B$$

- Atom-wise layers.

$$\mathbf{c}_i^{(l+1)} = \mathbf{W}^{(l)} \mathbf{c}_i^{(l)} + \mathbf{b}^{(l)}$$

SchNet – Interaction

- Interaction.

$$\mathbf{x}_i^{(l+1)} = \left(\mathbf{X}^{(l)} \star \mathbf{W}^{(l)}\right)_i = \sum_{j=0}^{n_{\text{atoms}}} \mathbf{x}_j^{(l)} \circ \mathbf{W}^{(l)} \left(\mathbf{r}_j - \mathbf{r}_i\right).$$

  Instead of a learnable tensor, the filter is a neural network $\mathbb{R}^3 \to \mathbb{R}^F$ with parameter matrix $\mathbf{W}^{(l)}$.

- Filter-generating networks.
  - *Rotational invariance*: use pairwise distances instead of relative positions and expand them into Gaussians

$$e_k\left(\mathbf{r}_j - \mathbf{r}_i\right) = \exp\left(-\gamma\left(\|\mathbf{r}_j - \mathbf{r}_i\| - \mu_k\right)^2\right).$$

  - *Periodic boundary conditions*: for atoms with PBCs, $\mathbf{x}_i$ should be invariant w.r.t. all periodic repetitions, $\mathbf{x}_i = \mathbf{x}_{ib} = \mathbf{x}_{ib} = \cdots$ for repeated unit cells $a, b, \cdots$.

SchNet – Incorporate PBC into Filter

Filter satisfying PBCs

Given a filter $\tilde{\mathbf{W}}^{(l)}\left(\mathbf{r}_{jb} - \mathbf{r}_{ia}\right)$ over all atoms with $\|\mathbf{r}_{jb} - \mathbf{r}_{ia}\| < r_{\text{cut}}$, where all i's forms a set $\mathcal{N}$, the convolution operator works as follows

$$\mathbf{x}_i^{(l+1)} = \mathbf{x}_{im}^{(l+1)} = \frac{1}{|\mathcal{N}|} \sum_{\substack{j,n \\ \mathbf{r}_{jn}}} \mathbf{x}_{jn}^{(l)} \circ \tilde{\mathbf{W}}^{(l)}\left(\mathbf{r}_{jn} - \mathbf{r}_{im}\right)$$

$$= \frac{1}{|\mathcal{N}|} \sum_{j} \mathbf{x}_{j}^{(l)} \circ \underbrace{\left(\sum_{n} \tilde{\mathbf{W}}^{(l)}\left(\mathbf{r}_{jn} - \mathbf{r}_{im}\right)\right)}_{\mathbf{W}}.$$

- The filter depends on the PBCs of the atomic system.
- $\frac{1}{|\mathcal{N}|}$ serves as a normalization.

Visualize filters w/ and w/o PBC.

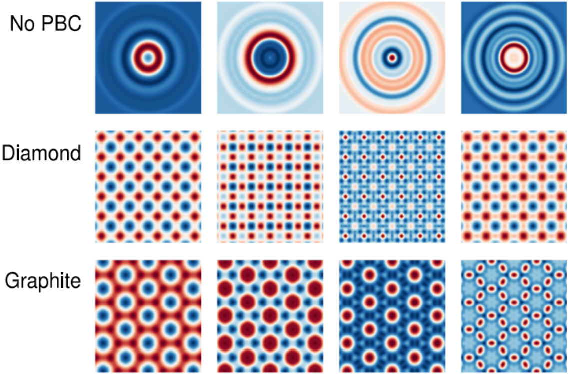

Figure 9: The first line shows filters that are only rotation-invariant, while the next two lines show filters aware of periodic boundaries.

SchNet – Activation Function & Prediction

- Activation function
  Shifted softplus function is used because of its zero at 0 and its infinite continuity.
  $$\mathrm{ssp}(x) = \ln\left(\frac{e^x + 1}{2}\right).$$
- Property prediction

$$\text{Atom } i\text{'s contribution: } \tilde{P}_i = \mathrm{ssp}\left(\mathbf{W}^{\mathrm{out}}\mathbf{x}_i^{(L)} + \mathbf{b}^{\mathrm{out}}\right)$$
$$\text{In total: } \tilde{P} = \sum_i \tilde{P}_i$$

SchNet – Training Objective

- Special case in prediction.
  When predicting atomic forces, SchNet predicts the energy and then differentiate it w.r.t. atoms' positions.

$$\tilde{\mathbf{F}}\left(\mathbf{Z}_1, \cdots, \mathbf{Z}_n, \mathbf{r}_1, \cdots, \mathbf{r}_n\right) = -\frac{\partial \tilde{E}}{\partial \mathbf{r}}\left(\mathbf{Z}_1, \cdots, \mathbf{Z}_n, \mathbf{r}_1, \cdots, \mathbf{r}_n\right).$$

- Training objective
  - *Predict property P*:

$$\mathcal{L}\left(\tilde{P}, P\right) = \left\|P - \tilde{P}\right\|.$$

  - *Predict energies and forces in molecular dynamics*:

$$\begin{aligned}&\mathcal{L}\left((\tilde{E}, \tilde{\mathbf{F}}_1, \cdots, \tilde{\mathbf{F}}_n), (E, \mathbf{F}_1, \cdots, \mathbf{F}_n)\right) \\ &= \rho\left\|E - \tilde{E}\right\|^2 + \frac{1}{n_{\text{atoms}}}\sum_{i=0}^{n_{\text{atoms}}}\left\|\mathbf{F}_i - \left(-\frac{\partial \tilde{E}}{\partial \mathbf{R}_i}\right)\right\|^2.\end{aligned}$$

Table of Contents

Input

- Nuclear charges $\mathbf{Z}$.
- Positions $\mathbf{R}$.

Structure

- Atom embedding.
- Atom-wise layers w/ residual.
- Interaction.
- Output.
- Property prediction.

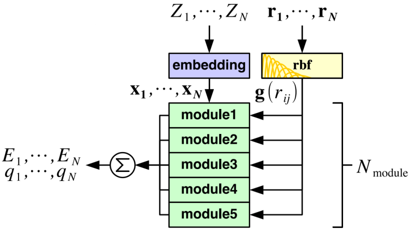

Figure 10: Overall framework of PhysNet.

PhysNet – Atom Embedding

- Atom embedding.
- Atom-wise layer w/ residual

$$\mathbf{c}_i^{(l+1)} = \mathbf{c}_i^{(l)} + \sigma\left(\mathbf{W}^{(l)}\mathbf{c}_i^{(l)} + \mathbf{b}^{(l)}\right).$$

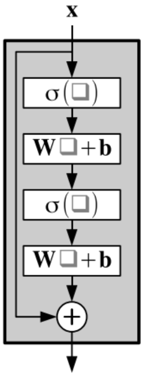

Figure 11: Residual layers after atom embedding in PhysNet.

PhysNet – Interaction

Interaction

The interaction, i.e., filtering and message-passing is

$$\mathbf{x}_i^{(l+1)} = \mathbf{u}^{(l)} \circ \mathbf{x}_i^{(l)} + \mathbf{W}^{(l)} \sigma \left(\mathbf{v}_i^{(l)} \right) + \mathbf{b}^{(l)}.$$

- $\mathbf{u}^{(l)}$ is similar to a memory gate.
- $\mathbf{v}_i^{(l)}$ is the message prototype $\tilde{\mathbf{v}}_i^{(l)}$ after several residual blocks.

$$\tilde{\mathbf{v}}_i^{(l)} = \sigma \left(\mathbf{W_I}^{(l)} \sigma \left(\mathbf{x}_i^{(l)} \right) + \mathbf{b_I}^{(l)} \right) + \sum_{j \neq i} \underbrace{\mathbf{G}^{(l)} \overbrace{\mathbf{g}\left(r_{ij}\right)}^{\text{radial basis}}}_{\text{Attention mask}} \circ \sigma \left(\mathbf{W_J}^{(l)} \sigma \left(\mathbf{x}_j^{(l)} \right) + \mathbf{b_J}^{(l)} \right).$$

PhysNet – RBF

PhysNet's radial basis function.

$$\mathbf{g}(r_{ij}) = [g_1(r_{ij}), \cdots, g_K(r_{ij})]^\top$$

$$g_k(r_{ij}) = \phi(r_{ij}) \cdot \exp\left(-\beta\left(\exp(-r_{ij}) - \mu_k\right)^2\right)$$

$$\phi(r_{ij}) = \begin{cases} 1 - 6\left(\frac{r_{ij}}{r_{\text{cut}}}\right)^5 + 15\left(\frac{r_{ij}}{r_{\text{cut}}}\right)^4 - 10\left(\frac{r_{ij}}{r_{\text{cut}}}\right)^3, & r_{ij} < r_{\text{cut}} \\ 0, & r_{ij} \geq r_{\text{cut}} \end{cases}$$

$\phi(r_{ij})$ aims to ensure continuity when r_{ij} approaches r_{cut}.

PhysNet – Output & Prediction

- Output block.
  For each module m, the atomic features pass through several residual layers, and then through a linear layer

$$\mathbf{y}_i^m = \mathbf{W}_{\text{out}}^m \sigma\left(\mathbf{x}_i^l\right) + \mathbf{b}_{\text{out}}^m$$

- Property prediction.
  Sum each module's atomic features and account for scale and shift.

$$\mathbf{y}_i = \mathbf{s}_{\mathbf{Z}_i} \cdot \left(\sum_{m=1}^{N_{\text{module}}} \mathbf{y}_i^m \right) + \mathbf{c}_{\mathbf{Z}_i}$$

  Final prediction of total energy in a system is

$$E = \sum_i^{N_{\text{atoms}}} E_i$$

- Account for long-range interaction beyond cutoff c_{cut}.

$$E = \sum_{i=1}^{N_{\text{atoms}}} E_i + k_e \sum_{i=1}^{N_{\text{atoms}}} \sum_{j>i}^{N_{\text{atoms}}} \tilde{q}_i \tilde{q}_j \chi(r_{ij}) + E_{D3}.$$

  $\chi(r_{ij})$ is an envelope of cutoff function $\phi(r_{ij})$, and E_{D3} is a result from DFT-D3 or learned by NN.

- Correct partial charges $\tilde{q}_i$.

$$\tilde{q}_i = q_i - \frac{1}{N_{\text{atoms}}} \left(\sum_{j=1}^{N_{\text{atoms}}} q_j - Q \right).$$

Table of Contents

Input

- Nuclear charges $\mathbf{Z}$.
- Pairwise distances $\mathbf{D}$.

Structure

- RBF & SBF.
- Atom embedding.
- Interaction.
- Output.

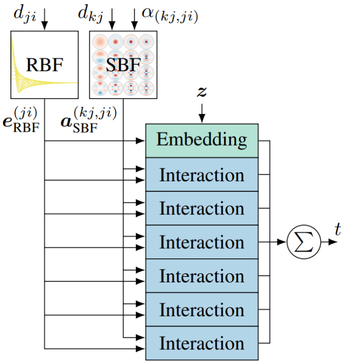

Figure 12: Overall framework of DimeNet.

DimeNet – Interaction

Interaction module that considers angles.

- Directional message passing of DimeNet

$$\mathbf{x}_{ji}^{(l+1)} = f_{\text{update}}\left(\mathbf{x}_{ji}^{(l)}, \sum_{k \in \mathcal{N}_j \setminus \{i\}} f_{\text{int}}\left(\mathbf{x}_{kj}^{(l)}, \mathbf{e}_{\text{RBF}}^{(ji)}, \alpha_{\text{SBF}}^{(kj,ji)}\right)\right).$$

- Both RBF and SBF derive from a solution set of a special case of Schrödinger equation. This solution set in a spherical coordinate systems (called *spherical harmonics*) is

$$\Psi(d, \alpha, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \left(a_{lm} j_l(kd) + b_{lm} y_l(kd)\right) Y_l^m(\alpha, \phi).$$

DimeNet – SBF & RBF

- For SBF, a 2D basis is needed for d_{kj} and $\alpha_{(kj,ji)}$, therefore, m is set to 0. After normalization, it becomes[2]

$$\tilde{\alpha}_{\mathsf{SBF},ln}(d,\alpha) = \sqrt{\frac{2}{c^3} j_{j+1}^2(z_{ln})} j_l \left(\frac{z_{ln}}{c} d\right) Y_l^0(\alpha).$$

- For RBF, it should only have a single variable d, therefore, both l and m are set to 0. After normalization and using $j_0(d) = \dfrac{\sin d}{d}$

$$\tilde{e}_{\mathsf{RBF},n}(d) = \sqrt{\frac{2}{c}} \frac{\sin\left(\frac{n\pi}{c} d\right)}{d}.$$

- In practice, an envelope function $u(d)$ is introduced to ensure the continuity at the cutoff: $\alpha = u \cdot \tilde{\alpha}, e = u \cdot \tilde{e}$.

[2] $y_l(\cdot)$ is a divergent function, and it is eliminated by setting b_{lm} to 0.

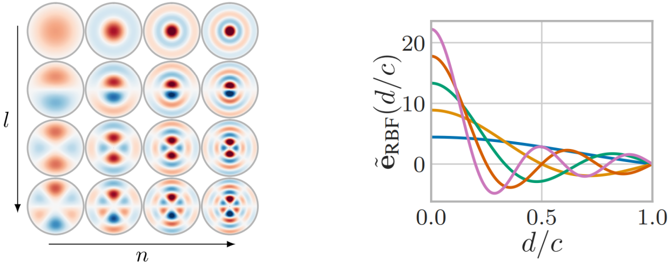

Figure 13: Visualize spherical basis $\tilde{\alpha}_{\mathsf{SBF},ln}(d,\alpha)$ and radial basis $\tilde{e}_{\mathsf{RBF},n}(d)$.

DimeNet – Message-Passing

- For the first layer

$$\mathbf{m}_j i^{(1)} = \sigma\left(\left[\mathbf{h}_j^{(0)} \| \mathbf{h}_i^{(0)} \| \mathbf{e}_{\mathsf{RBF}}^{(ji)}\right] \mathbf{W} + \mathbf{b}\right).$$

- For subsequent layers

$$\tilde{\mathbf{m}}_{ji}^{(l+1)} = \sigma\left(\mathbf{W}\mathbf{m}_{ji}^{(l)}\right) + \sum_{k \in \mathcal{N}_j \setminus \{i\}} \left(\mathbf{W}\alpha_{\mathsf{SBF}}^{(kj,ji)}\right)^\top \mathbf{W}\left(\mathbf{e}_{\mathsf{RBF}}^{(ji)}\mathbf{W} \circ \mathbf{m}_{kj}^{(l)}\right)$$

$$\mathbf{m}_{ji}^{(l+1)} = \mathrm{Residual}\left(\tilde{\mathbf{m}}_{ji}^{(l)}, \mathbf{m}_{ji}^{(l)}\right)$$

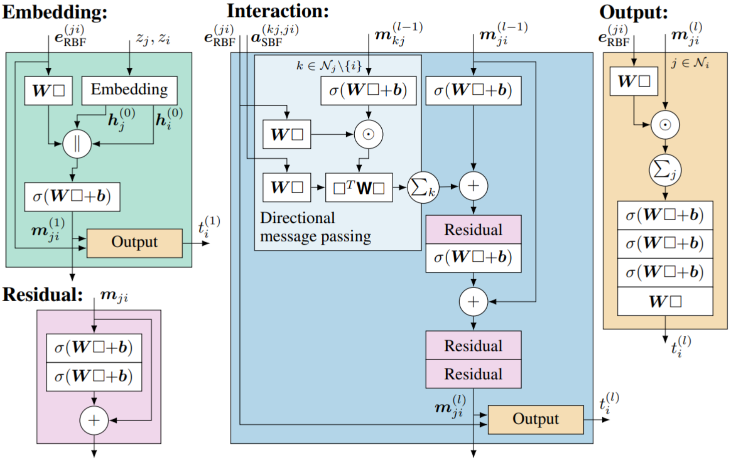

Figure 14: Each module's operations in DimeNet.

Table of Contents

Compare Building Blocks

Model / Component	DTNN	SchNet	PhysNet	DimeNet
Atom embedding	Randomly initialized acc. to nuclear charge	w/ linear layers	w/ residual layers	w/ linear & RBF
RBF	A series of Gaussians w/ same std. and evenly separated mean	Gaussians w/ scaling	Gaussians w/ scaling and continuity	spherical harmonics $\alpha_{\mathsf{SBF}}(d, \alpha)$ and continuity
Filter	Linear layer on RBF	Linear, w/ PBC awareness	Learned attention mask	w/ 2D SBF
Output	Sum each atom's contribution	Sum each atom's contribution	w/ correction for long-range interaction	Sum each atom's contribution in each layer
Similarities	1. Each type of element has a distinct, learnable embedding. 2. Atom only interacts with neighbors within cutoff range. 3. Molecular property is the summation of each atom's contribution.			

Table 1: Comparing the differences and similarities of different models.

Table of Contents

Performances Comparison

Target	Unit	PPGN	SchNet	PhysNet	MEGNet-s	Cormorant	**DimeNet**
μ	D	0.047	0.033	0.0529	0.05	0.13	**0.0286**
α	${a_0}^3$	0.131	0.235	0.0615	0.081	0.092	**0.0469**
ϵ_{HOMO}	meV	40.3	41	32.9	43	36	**27.8**
ϵ_{LUMO}	meV	32.7	34	24.7	44	36	**19.7**
$\Delta\epsilon$	meV	60.0	63	42.5	66	60	**34.8**
$\langle R^2 \rangle$	${a_0}^2$	0.592	**0.073**	0.765	0.302	0.673	0.331
ZPVE	meV	3.12	1.7	1.39	1.43	1.98	**1.29**
U_0	meV	36.8	14	8.15	12	28	**8.02**
U	meV	36.8	19	8.34	13	-	**7.89**
H	meV	36.3	14	8.42	12	-	**8.11**
G	meV	36.4	14	9.40	12	-	**8.98**
c_{v}	$\frac{\text{cal}}{\text{mol K}}$	0.055	0.033	0.0280	0.029	0.031	**0.0249**
std. MAE	%	1.84	1.76	1.37	1.80	2.14	**1.05**
logMAE	-	-4.64	-5.17	-5.35	-5.17	-4.75	$\mathbf{-5.57}$

Table 2: Mean absolute error (MAE) on QM9 dataset [4]. The prediction targets are 11 physical quantities of a molecule.

Thank You for Your Attention

Q & A

References

[1] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks.," *Nature Communications*, vol. 8, no. 1, pp. 13 890–13 890, 2017.

[2] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller, "Schnet - a deep learning architecture for molecules and materials.," *Journal of Chemical Physics*, vol. 148, no. 24, pp. 241 722–241 722, 2018.

[3] O. T. Unke and M. Meuwly, "Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges.," *Journal of Chemical Theory and Computation*, vol. 15, no. 6, pp. 3678–3693, 2019.

[4] J. Klicpera, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in *ICLR 2020 : Eighth International Conference on Learning Representations*, 2020.