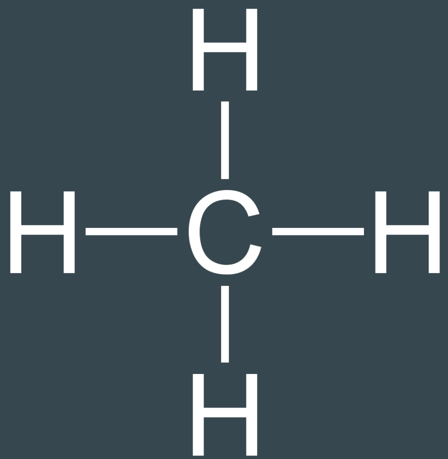
XYZ Files



Methane

```
5
free=-1545.46 (Comment)
C 5.18160188 -0.90711087 -2.80992509
H 0.64223971  0.28323184  1.01144250
H 0.59174445 -1.01258058 -0.19836824
H 0.60046186  0.68235075 -0.71605884
H 0.90046186  0.48235075 -0.21605884
```

Coulomb Matrix

```
6
free=-3360.46 (Comment)
C 5.18160188 -0.90711087 -2.80992509
C 6.18160188 -2.90711087 -3.80992509
H 0.64223971  0.28323184  1.01144250
H 0.59174445 -1.01258058 -0.19836824
H 0.60046186  0.68235075 -0.71605884
H 0.90046186  0.48235075 -0.21605884
```

$$C_{ij} = \begin{cases} \frac{1}{2} Z_i^{2.4}, & i = j \\ \frac{Z_i Z_j}{\|R_i - R_j\|_2}, & i \neq j. \end{cases}$$

$$\mathbf{C} = \begin{array}{c} \\ \mathrm{H} \\ \mathrm{H} \\ \mathrm{C} \\ \mathrm{C} \\ \mathrm{H} \\ \mathrm{H} \end{array} \begin{array}{c} \begin{array}{cccccc} \mathrm{H} & \mathrm{H} & \mathrm{C} & \mathrm{C} & \mathrm{H} & \mathrm{H} \end{array} \\ \left[\begin{array}{cccccc} 0.5 & 0.3 & 2.9 & 1.5 & 0.2 & 0.2 \\ 0.3 & 0.5 & 2.9 & 1.5 & 0.2 & 0.2 \\ 2.9 & 2.9 & 36.9 & 14.3 & 1.5 & 1.5 \\ 1.5 & 1.5 & 14.3 & 36.9 & 2.9 & 2.9 \\ 0.2 & 0.2 & 1.5 & 2.9 & 0.5 & 0.3 \\ 0.2 & 0.2 & 1.5 & 2.9 & 0.3 & 0.5 \end{array} \right] \end{array}$$

Given

some representation of molecules.

- XYZ files.
- Coulomb Matrices*
- Smiles**

...etc.

Predict

some properties.

- Atomization energy
- Highest occupied molecular orbital (HOMO)
- ...

- Energies (Corresponding to Molecular Orbitals)
- Absorption spectrum

* Rupp 2015 Machine learning for quantum mechanics in a nutshell (eq. 26)
**https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system

Given
some representation of molecules.

- Smiles*
- Coulomb Matrices**
- XYZ files.

```
4
free=-1545.46
C 5.18160188 -0.90711087 -2.80992509
H 0.64223971  0.28323184  1.01144250
H 0.59174445 -1.01258058 -0.19836824
H 0.60046186  0.68235075 -0.71605884
```

...etc.

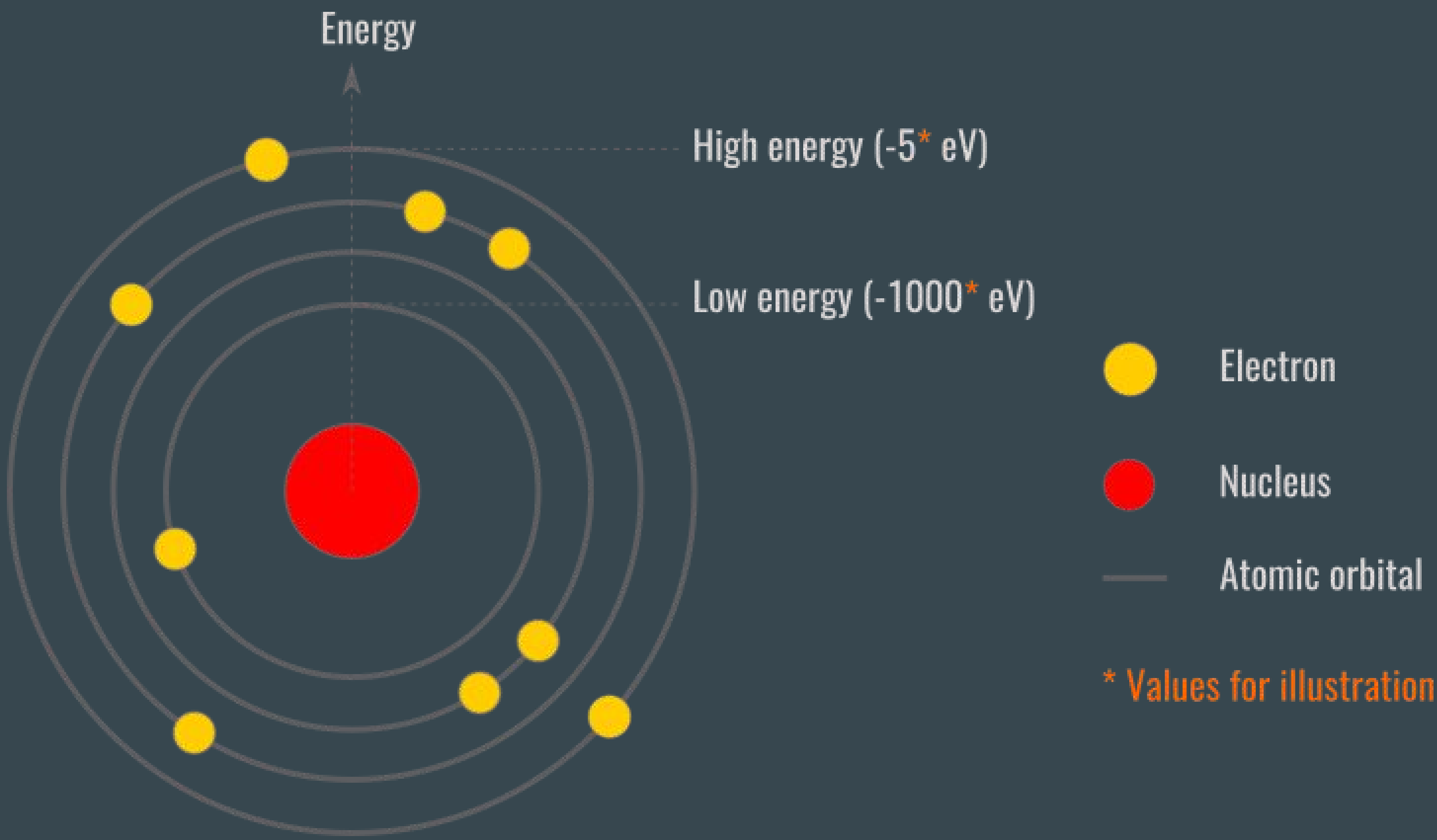
Predict
some properties.

- Atomization energy
- Highest occupied molecular orbital (HOMO)
- ...

- Energies (Corresponding to Molecular Orbitals)
- Absorption spectrum

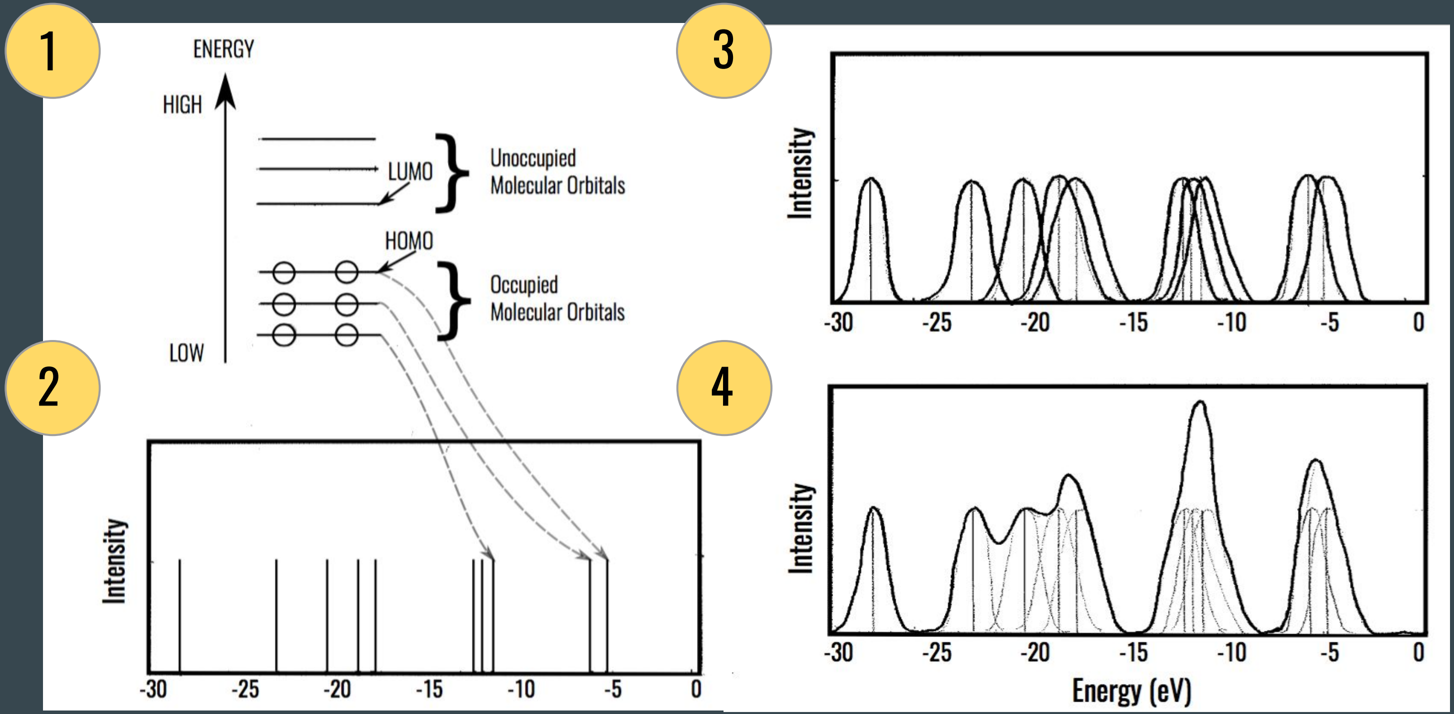
*https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system
** Rupp 2015 Machine learning for quantum mechanics in a nutshell (eq. 26)

Intuition from atoms



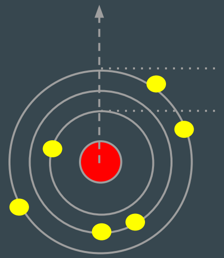
* Values for illustration only

Extending to molecules



Target values - HOMO energies

Intuition from Atoms



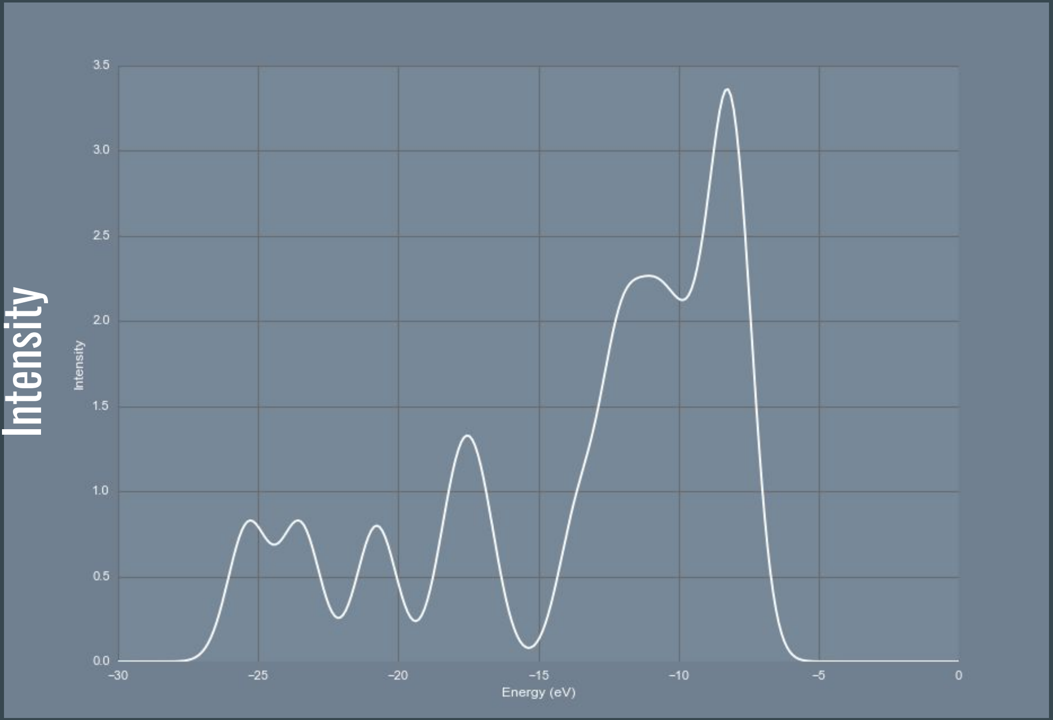
For each molecule:
We want to predict a set of 16 real numbers corresponding to energies of these molecular orbitals.

-19.75084	-18.45148	-16.5375	-14.57497	-13.96967	-11.51794	-10.60673	-10.40516	-9.80747	-9.45303	-8.99068	-8.45312	-7.89553	-7.76155	-7.55804	-7.52253
-19.96160	-17.98181	-16.68139	-15.22888	-13.48942	-11.29259	-10.76136	-10.37269	-9.90364	-9.19086	-9.08365	-8.57864	-8.26906	-7.66870	-7.56287	-7.43001
-17.73987	-16.7965	-16.05788	-13.32784	-13.08767	-12.23344	-11.24781	-11.01421	-9.26942	-8.66231	-8.45719	-8.05627	-7.63661	-7.29622	-7.22952	-7.13946
-18.76251	-17.28378	-14.79647	-13.78733	-13.20733	-12.39633	-11.21562	-10.15360	-9.81649	-9.76154	-8.30788	-8.11097	-7.84390	-7.76135	-6.63684	-6.25399
-18.11031	-14.71805	-14.71805	-13.28189	-11.62577	-10.58315	-10.46336	-10.46336	-8.99491	-8.99488	-8.98559	-8.01946	-8.01945	-7.34948	-7.00147	-7.00146
-17.14543	-16.77362	-14.03449	-12.84359	-11.59662	-11.02884	-10.38074	-9.912060	-9.44403	-9.27453	-9.07982	-8.92919	-8.31617	-7.98388	-6.46692	-6.13883
-17.60407	-15.07804	-14.82862	-13.20740	-11.19026	-11.04740	-10.33376	-9.865510	-9.78466	-9.42230	-8.58497	-8.40987	-7.70107	-7.37607	-6.90260	-6.81396

* For illustration only

Target values - Absorption spectra

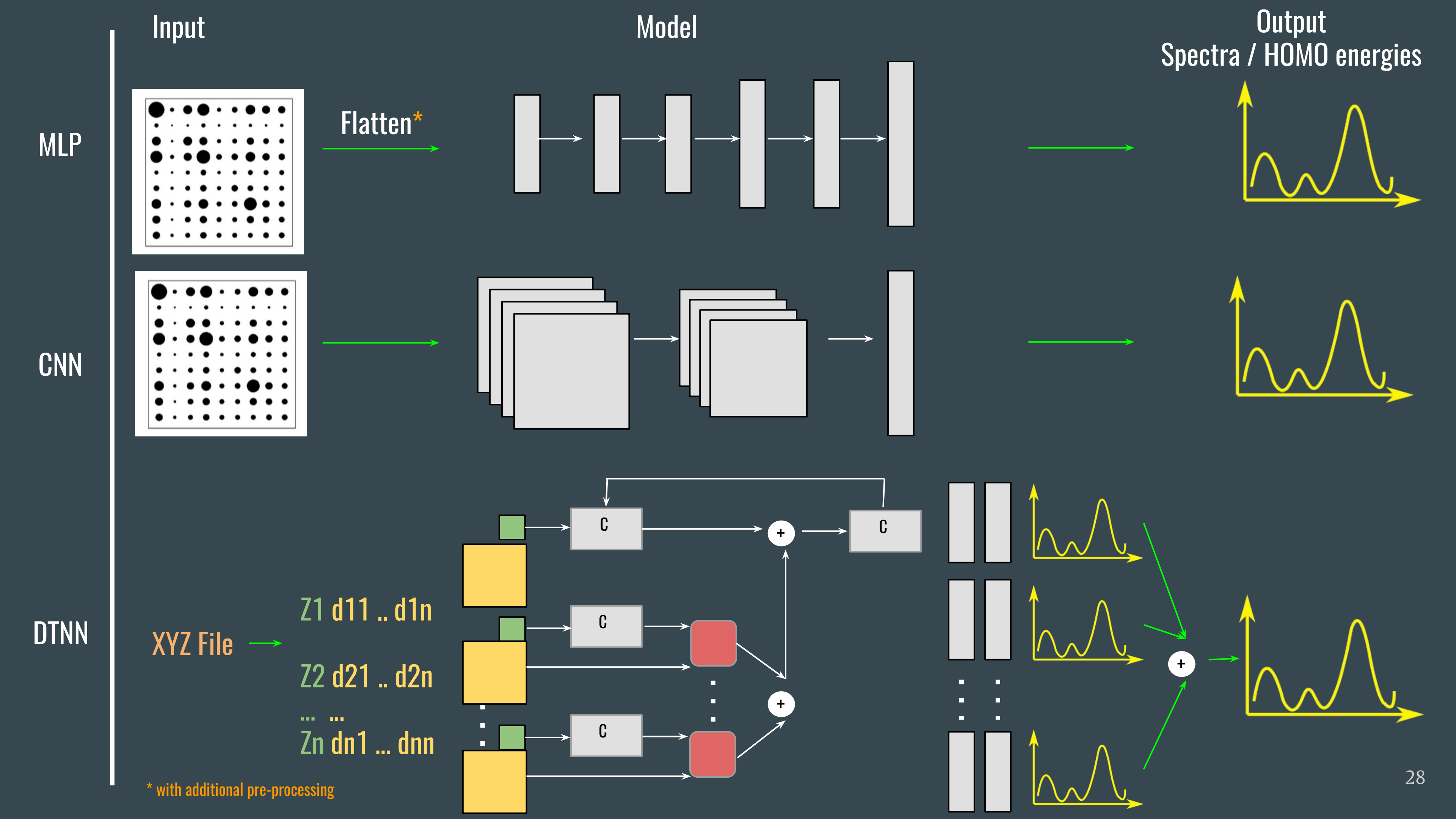
Intensity



Energy (eV)

For each molecule:
We also want to predict a set of 300 real numbers corresponding to the spectrum discretized at 300 points.

Outline



Input
Model
Output
Spectra / HOMO energies
MLP
Flatten*
CNN
DTNN
XYZ File
Z1 d11 .. d1n
Z2 d21 .. d2n
... ...
Zn dn1 ... dnn
C
C
C
C
+
+
+
* with additional pre-processing

Outline

Introduction

Current Methods

Need for Machine Learning

The data

Our work

Results

Conclusion

Quantitative Results

Datasets →	6K		132K	
Model (Input) ↓	16 HOMO energies (eV)	Spectrum	16 HOMO energies (eV)	Spectrum (3 run summary)
MLP (Coulomb matrix)	0.317 ± 0.000	NA	NA	NA
CNN (Coulomb matrix)	0.304 ± 0.006	0.282 ± 0.002	0.231 ± 0.002	0.199 ± 0.000
DTNN (XYZ file)	**0.251 ± 0.024**	**0.210 ± 0.000**	**0.186 ± 0.002**	**0.145 ± 0.000**

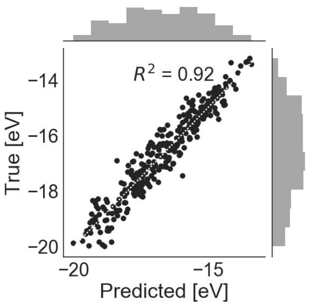
state of the art

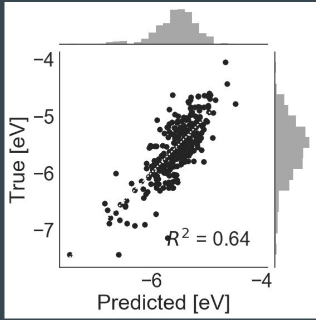
Qualitative Results
HOMO - Energies

6K dataset

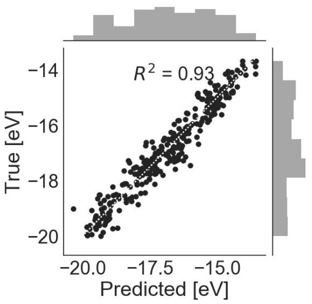
Best prediction (HOMO-15) | Worst prediction (HOMO)

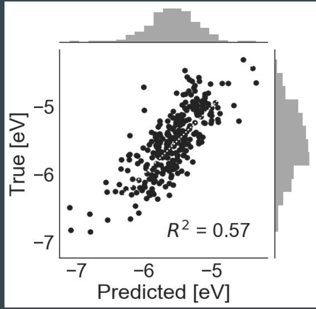
MLP



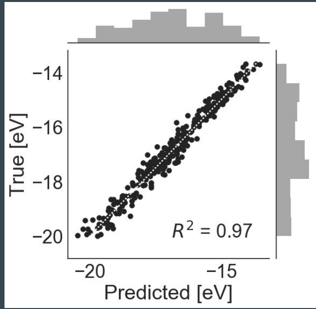


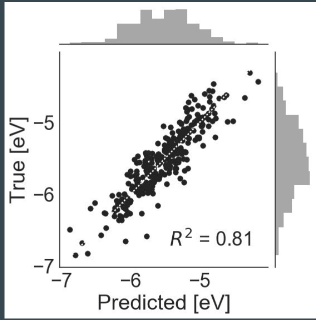
CNN

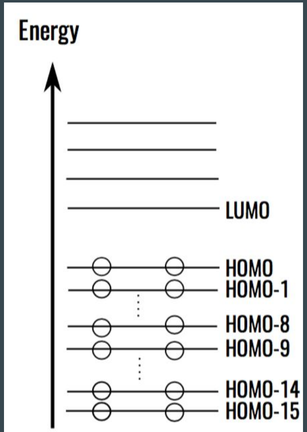




DTNN





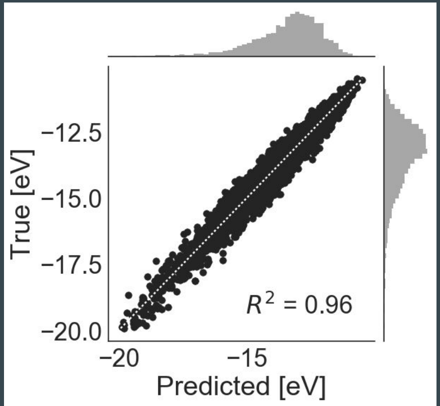


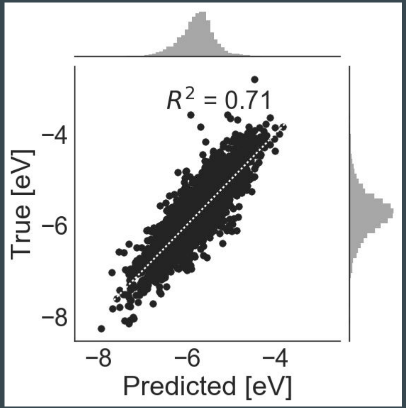
132K dataset

Best prediction (HOMO-15)

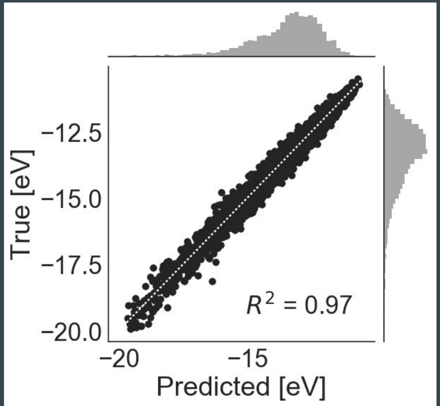
Worst prediction (HOMO)

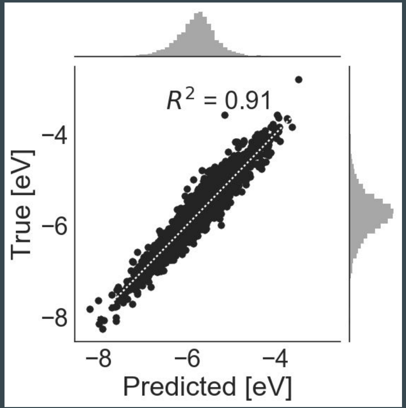
CNN

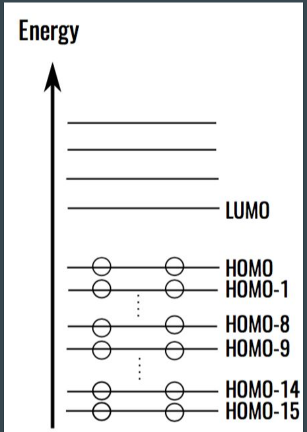




DTNN







Qualitative Results
Spectra

Spectra predictions for the first time* using machine learning

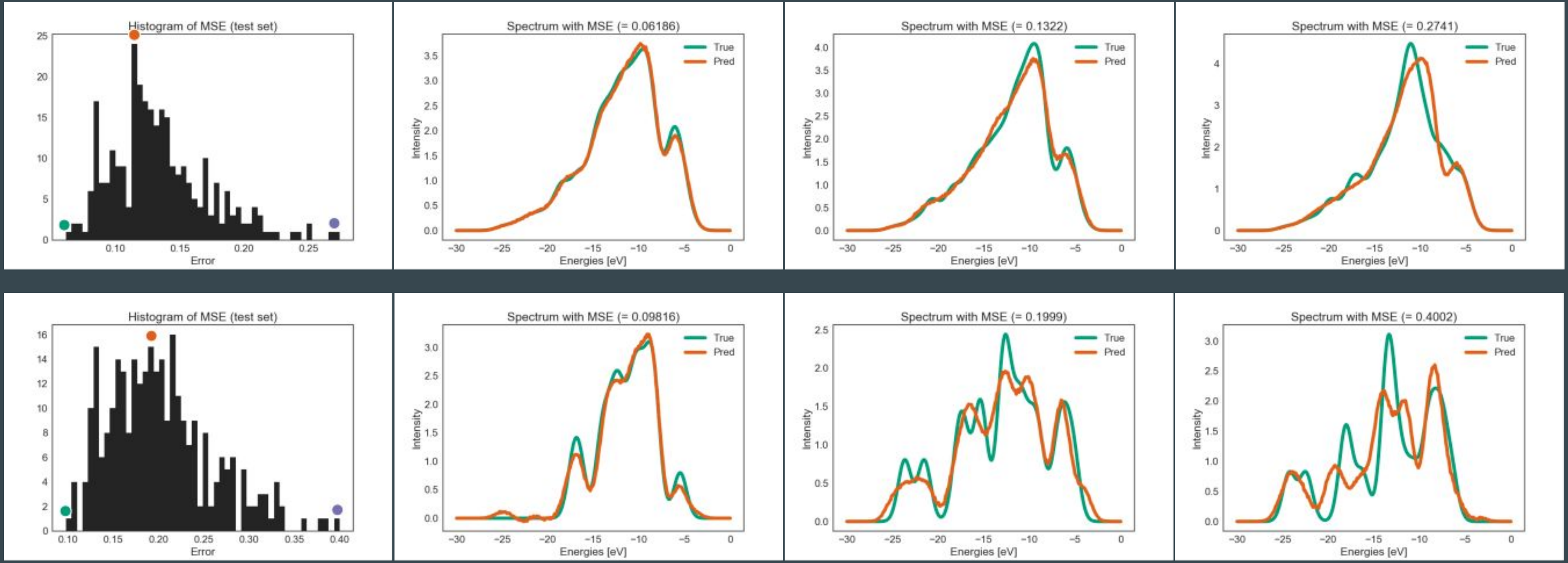
* based on our literature review

6K dataset

Best | Average | Worst

CNN

DTNN

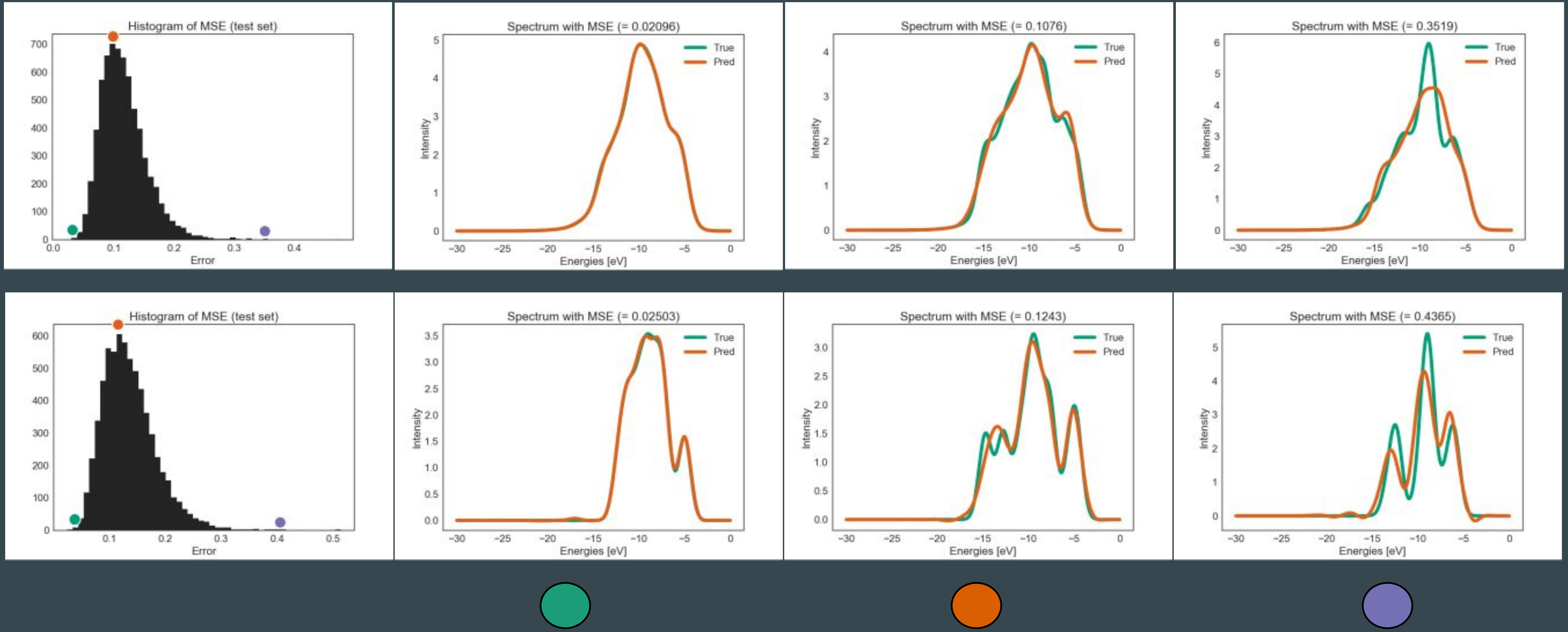


132K dataset

Best | Average | Worst

CNN

DTNN



Outline

Conclusion

- Search for novel materials has a significant societal impact.
- Current simulation based methods slow*.
- Machine learning (ML) based methods promising.
- Further work needed to improve ML predictions.

* For the intended application

Outline

Introduction

Current Methods

Need for Machine Learning

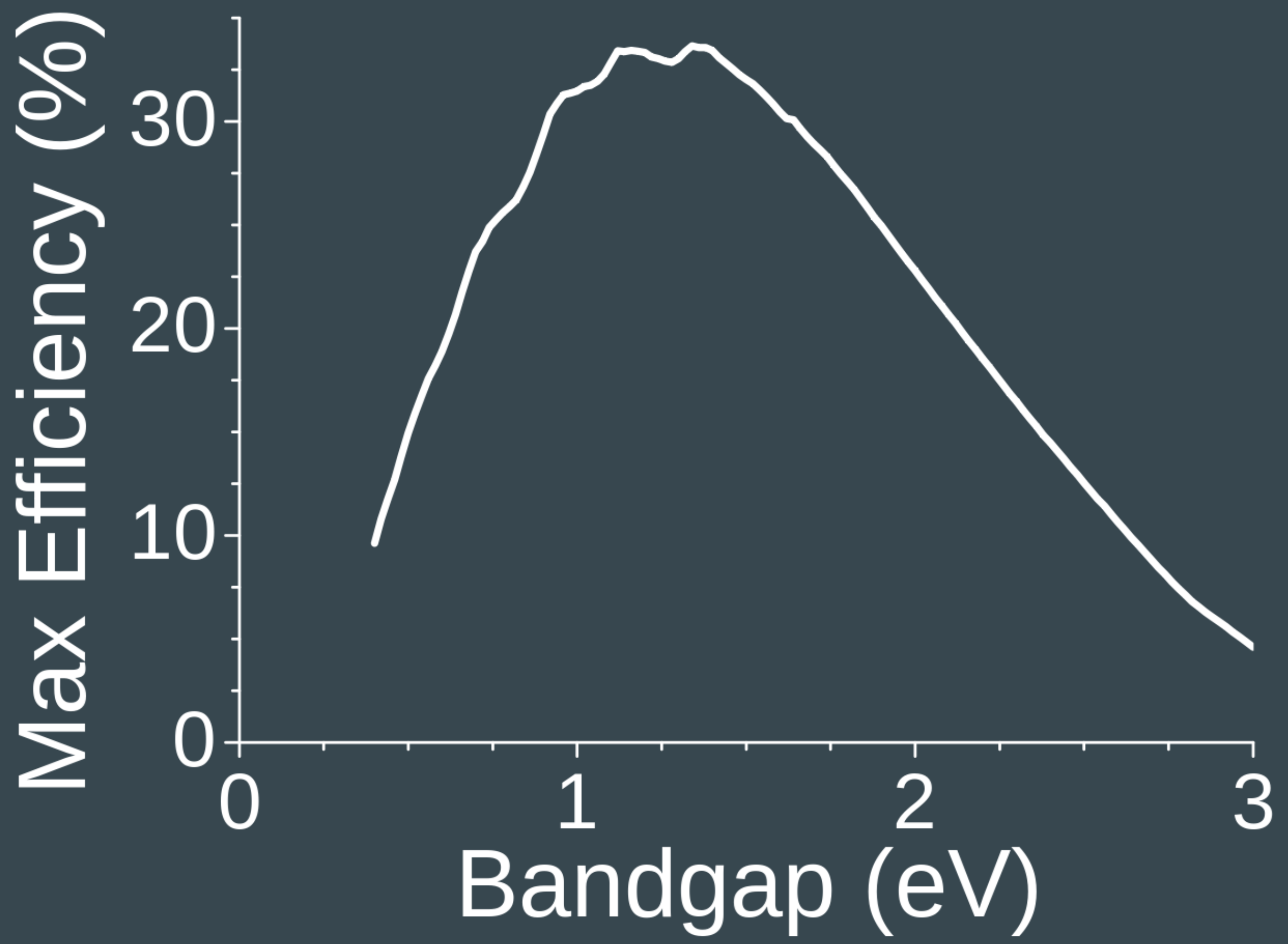
The data

Our work

Results

Conclusion

Thank You !

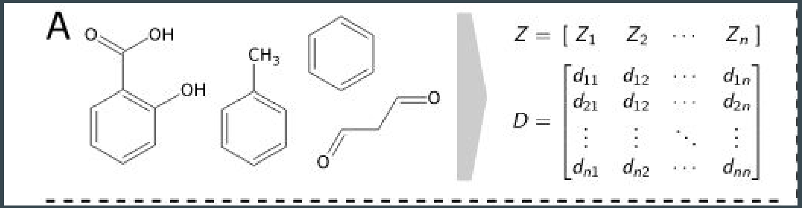


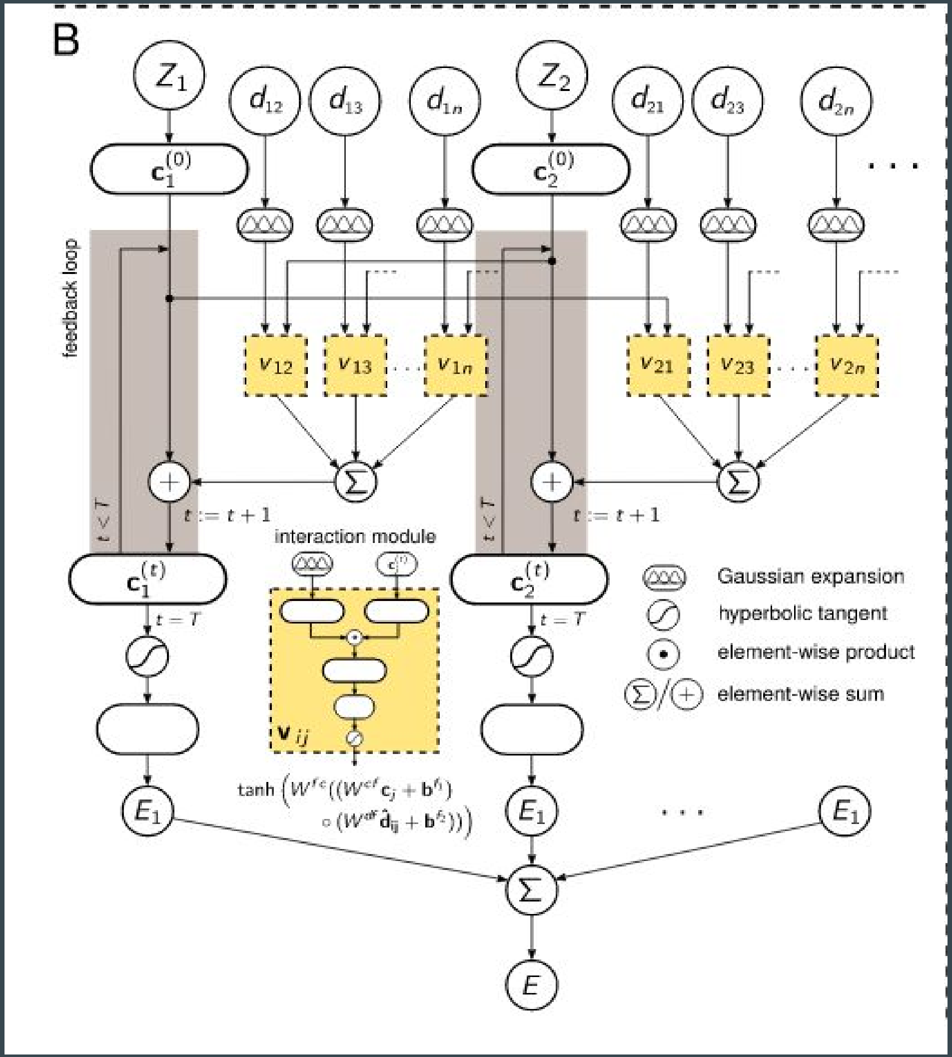
https://en.wikipedia.org/wiki/Solar_cell_efficiency

Our work

Deep Tensor Neural Network*

The total energy E for a molecule composed of N atoms can be written as a sum over N atomic energy contributions Ei.

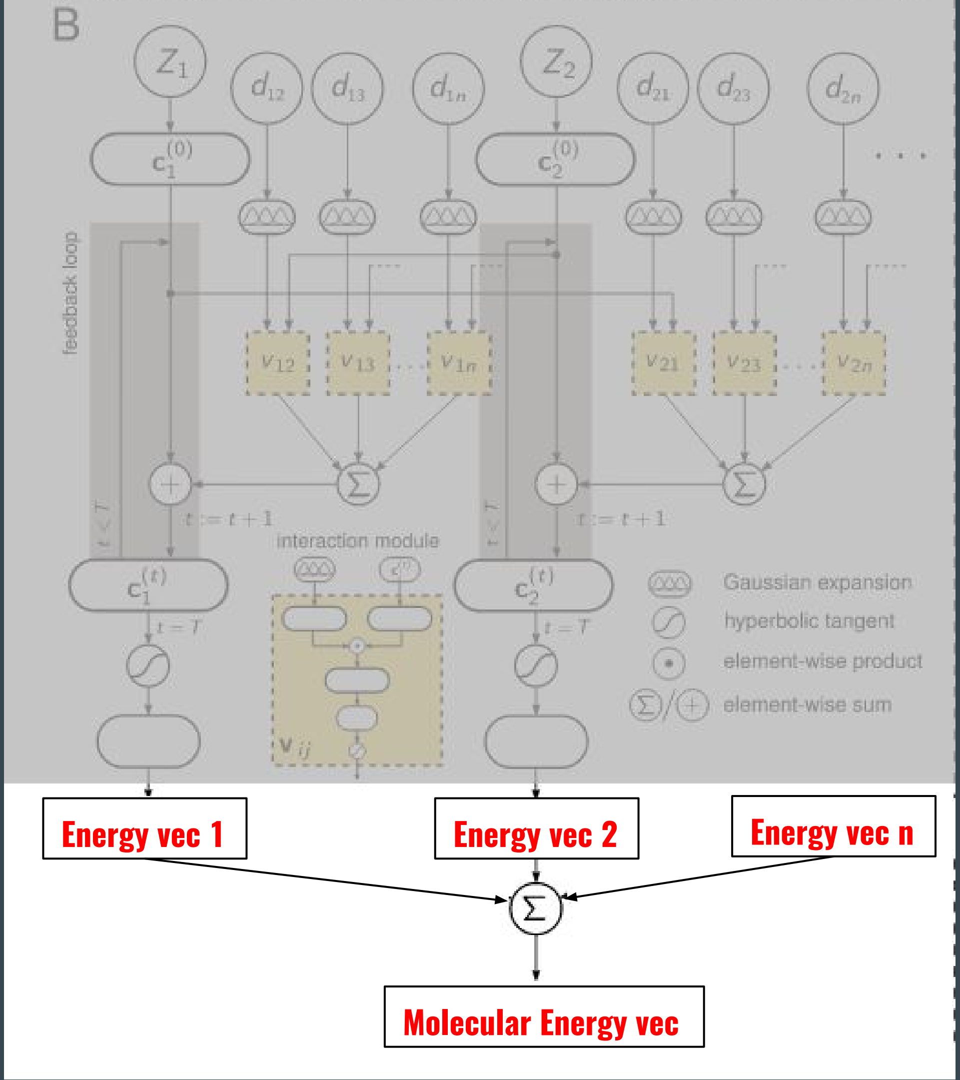




* Quantum-Chemical Insights from Deep Tensor Neural Networks. Schütt et.al 2017

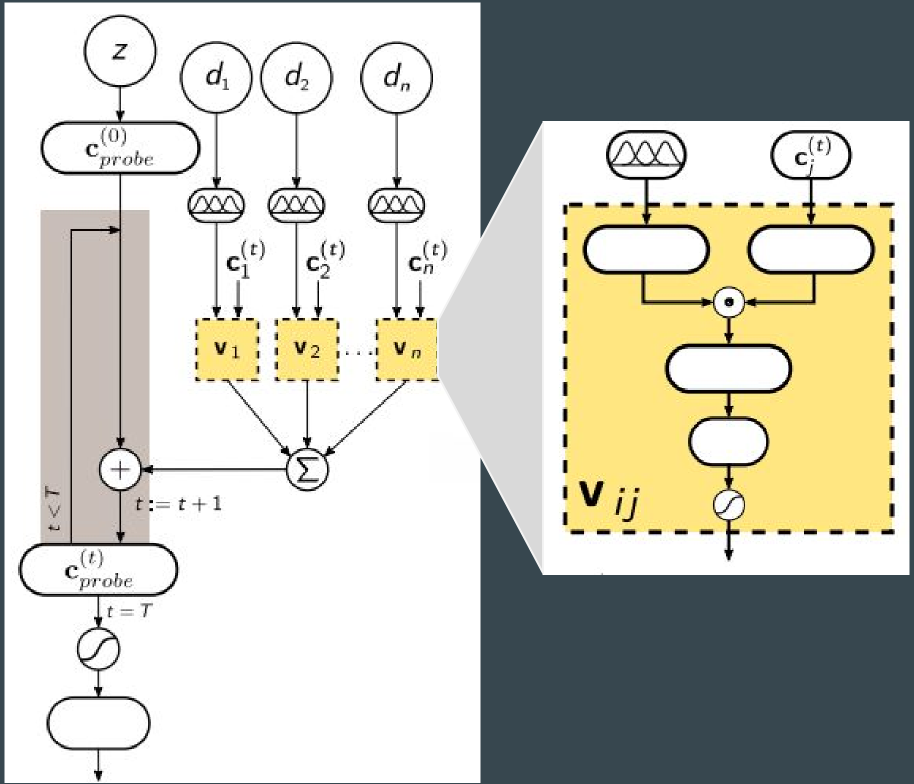
We make a small change to the network.

Similar Intuition :
Molecular energy vector* = Sum of Individual atomic contributions.



* Vector of 16 real values. Similar modification made to predict the spectrum.

Our work



1. Assign initial atomic descriptors.

$$\mathbf{c}_i^{(0)} = \mathbf{c}_{Z_i} \in R^B \qquad \mathbf{c}_Z \sim \mathcal{N}(0, 1/\sqrt{B}) \qquad B = 30$$

2. Gaussian feature expansion of the interatomic distances (in experiments $\Delta\mu = \sigma = 0.2$)

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

3. Perform T interaction passes

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

$$\mathbf{v}_{ij} = \tanh\left(W^{fc}((W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df}\hat{\mathbf{d}}_{\mathbf{ij}} + \mathbf{b}^{f_2}))\right), \in R^B$$

Why is this model called Deep Tensor neural net ?

$$v_{ijk} = \tanh\left(\mathbf{c}_j^{(t)} V_k \hat{\mathbf{d}}_{ij} + (W^c \mathbf{c}_j^{(t)})_k + (W^d \hat{\mathbf{d}}_{ij})_k + b_k\right)$$

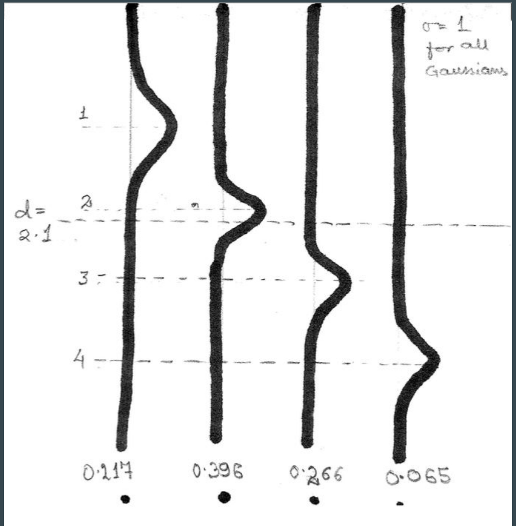


Figure 1

Gaussian feature expansion

$$\underset{\uparrow \atop \mathbb{R}^{30\times a}}{W^{fc}}\Big(\underset{\uparrow \atop \mathbb{R}^{a\times 30}}{W^{cf}} c_j + \underset{\uparrow \atop \mathbb{R}^{30}}{b^{f_1}}\Big)$$

$$\underset{\uparrow \atop \mathbb{R}^{30\times a}}{W^{fc}}\Big(\underset{\uparrow \atop \mathbb{R}^{a\times G}}{W^{df}} \hat{d}_{ij} + \underset{\uparrow \atop \mathbb{R}^{G}}{b^{f_2}}\Big)$$

e.g. if $a = 2 \Rightarrow$ low rank factorization

Figure 2

Low Rank Factorization To Reduce number of parameters

$$\underset{\uparrow \atop \mathbb{R}^{30}}{c_j^{(t)}}\; V_K \;\underset{\uparrow \atop \mathbb{R}^{G}}{\hat{d}_{ij}} + W^c \underset{\uparrow \atop \mathbb{R}^{30}}{c_j^{(t)}} + W^d \underset{\uparrow \atop \mathbb{R}^{G}}{\hat{d}_{ij}}$$

output dimension of $V_{ij} \equiv$ dimension of c.

$\therefore\; V_K \in \mathbb{R}^{30\times 30\times G}$.

Figure 3

Where is the Tensor ?

Qualitative - HOMO energy