Machine learning under physical constraints
Final projet

Sixin Zhang
(sixin.zhang@toulouse-inp.fr)

Kaggle project: regression of molecular energy

- Problem: predict the molecular energy in 3d space based on its geometric structure.

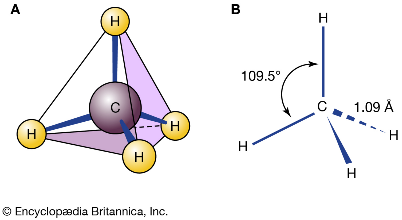


Figure: Image from `https://www.britannica.com/science/methane`.

Kaggle participation link: `https://www.kaggle.com/t/f1caef4be2ae4a82861dfd798f6b91c6`

Molecule energy regression

- Analyze 3d structures with physical constraints: invariant properties.

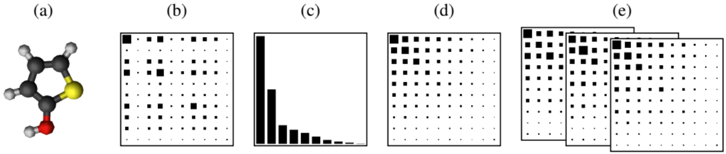


Figure 1: Different representations of the same molecule: (a) raw molecule with Cartesian coordinates and associated charges, (b) original (non-sorted) Coulomb matrix as computed by Equation 1, (c) eigenspectrum of the Coulomb matrix, (d) sorted Coulomb matrix, (e) set of randomly sorted Coulomb matrices.

From: Learning Invariant Representations of Molecules for Atomization Energy Prediction, Montavon1 et al. 2012

- Challenge: How to choose $\Phi(x)$?

Scattering in 3d

- Excellent performance with scattering + Multi-linear regression (T-Scat)

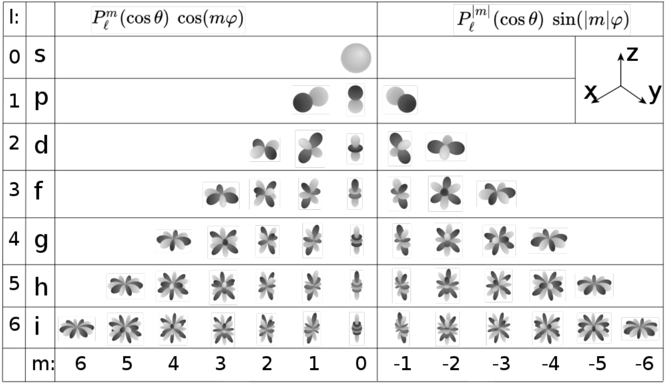
TABLE I. Prediction errors for molecular energies of the QM9 dataset in kcal/mol. From right to left: the scattering with linear regression, the scattering with trilinear regression, Neural Message Passing and Coulomb Matrices.

	L-Scat	T-Scat	NMP	CM	SchNet
U_0	1.89	0.50	0.45	2.95	0.31
U	2.4	0.51	0.45	2.99	
H	1.9	0.51	0.39	2.99	
G	1.87	0.51	0.44	2.97	
μ	0.63	0.34	0.030	0.45	
α	0.52	0.16	0.092	0.43	
ϵ_{HOMO}	4.08	1.97	0.99	3.06	
ϵ_{LUMO}	5.39	1.76	0.87	4.22	
ϵ_{gap}	7	2.73	1.60	5.28	
$\langle R^2 \rangle$	6.67	0.41	0.18	3.39	
zpve	0.004	0.002	0.0015	0.0048	
C_v	0.10	0.049	0.04	0.12	

From: Solid Harmonic Wavelet Scattering for Predictions of Molecule Properties, Eickenberg et al. 2018

Solid harmonic wavelets

- Let spherical harmonics on a unit sphere in 3d be $Y_\ell^m(\theta, \psi)$ for $\theta \in [0, \pi]$ and $\psi \in [0, 2\pi]$, $0 \leq \ell \leq L - 1$ and $-\ell \leq m \leq \ell$.



From wikipedia: Spherical harmonics

Solid harmonic wavelets

- Construct solid harmonic wavelets from spherical harmonics,

$$\psi_\ell^m(u) \propto e^{-|u|^2/2}|u|^\ell Y_\ell^m(u/|u|).$$

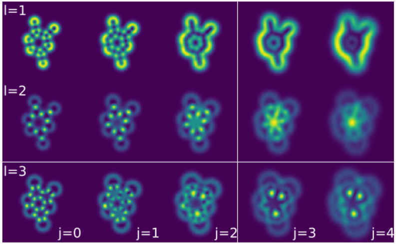
- Solid harmonic wavelets are constructed by dilating ψ_ℓ^m,

$$\psi_{j,\ell}^m(u) = 2^{-3j}\psi_\ell^m(2^{-j}u), \quad 0 \leq j \leq J-1.$$

Solid harmonic wavelet coefficients

- Order 1: $p_1 = (j, \ell)$

$$U_{p_1} x(u) = \left(\sum_{m=-\ell}^{\ell} |x \star \psi_{j,\ell}^{m}(u)|^2 \right)^{1/2}$$



From: Solid Harmonic Wavelet Scattering for Predictions of Molecule Properties, Eickenberg et al. 2018

Solid harmonic wavelet coefficients

- Order 2: $p_2 = (p_1, j_2)$

$$U_{p_2}x(u) = \left(\sum_{m=1}^{\ell} |U_{p_1}x \star \psi^m_{j_2,\ell}(u)|^2\right)^{1/2}$$

- Compute invariants from $U_{p_1}x$ and $U_{p_2}x$ by integrating over u with some exponent $q > 0$,

$$\Phi(x) = \left\{\int |U_{p_1}x(u)|^q du, \int |U_{p_2}x(u)|^q du\right\}_{p_1,p_2}.$$

- $\Phi(x)$ is invariant to translation and rotation of x in 3d.

Multi-linear regression

- Beyond Linear regression to capture interactions in $\Phi(x)$.
- General form, order (r, l)

$$f(x) = b + \sum_{i=1}^{l} (v_i \prod_{k=1}^{r} (\langle \Phi(x), w_i^{(k)} \rangle + c_i^{(k)})$$

- Linear regression case: $r = l = 1$.
- The parameters can be optimized with SGD.