Machine learning for drug design, medicine and quantum chemistry : an introduction

Convolutional neural networks and solid harmonic wavelet transform

Robert Benda

CERMICS

November 13, 2019

Outline

1. Machine learning for drug design and medicine : general introduction
   - Examples of applications
   - Molecular descriptors

2. Solid harmonic wavelet scattering for predictions of molecule properties (based on Refs. [1, 2])
   - General idea of the method
   - Wavelet scattering coefficients
   - Comparison to usual force fields
   - Why does it work so well : comparison to the multipolar expansion in polarizable force fields methods

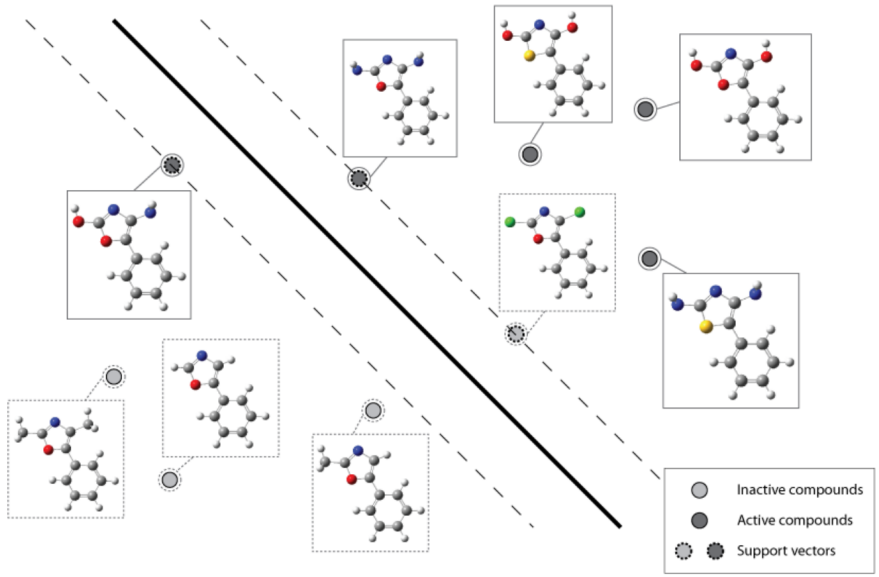

Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, Stéphane Mallat, and Louis Thiry.
Solid harmonic wavelet scattering for predictions of molecule properties.
The Journal of Chemical Physics, 148(24):241732, jun 2018.


Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, and Stephane Mallat.
Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3d electronic densities.
In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6540–6549. Curran Associates, Inc., 2017.

Outline

Examples of applications (1)

- Compound classification (*e.g.* with Support Vector Machines) according to their *functions* or *activities*



J.C. Gertrudes, *Machine Learning Techniques and Drug Design*, Current Medicinal Chemistry, **2012**, 19, 4289-4297.

Examples of applications (2)

- Compound classification (*e.g.* with SVMs) according to their *functions* or *activities*

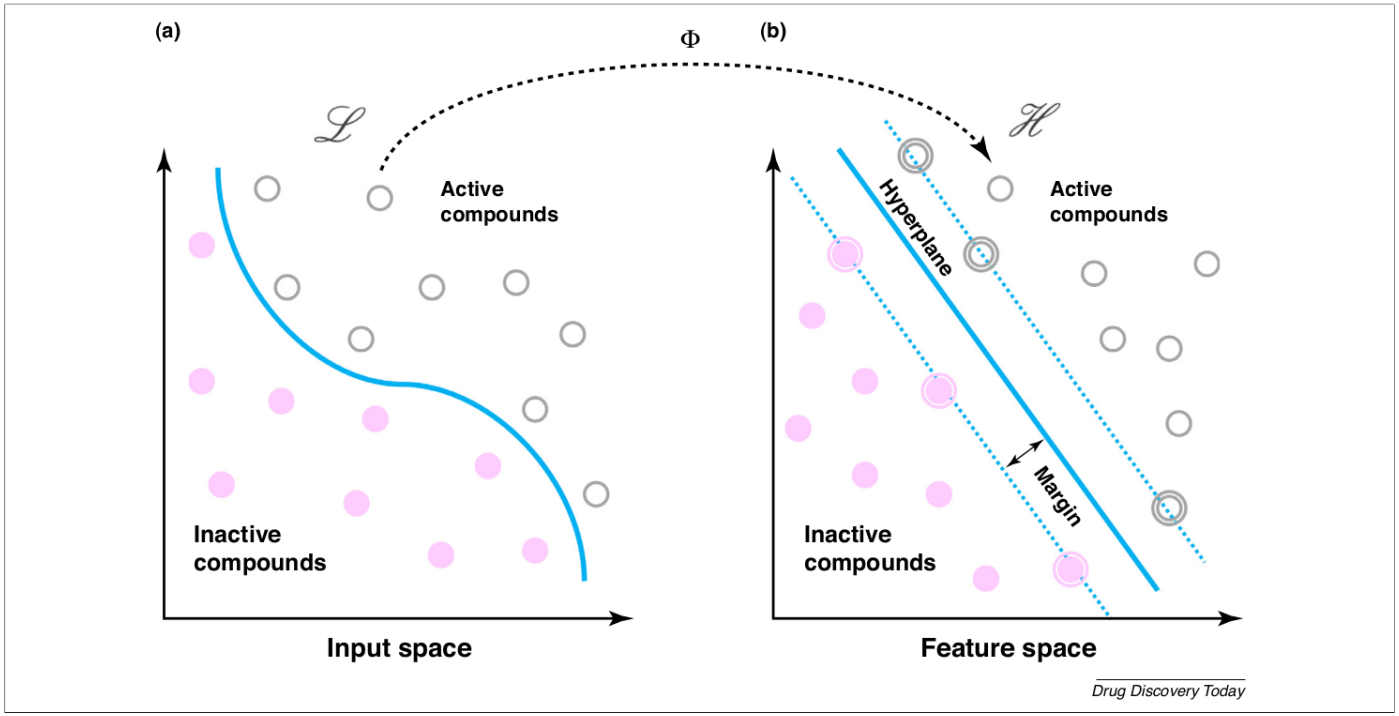


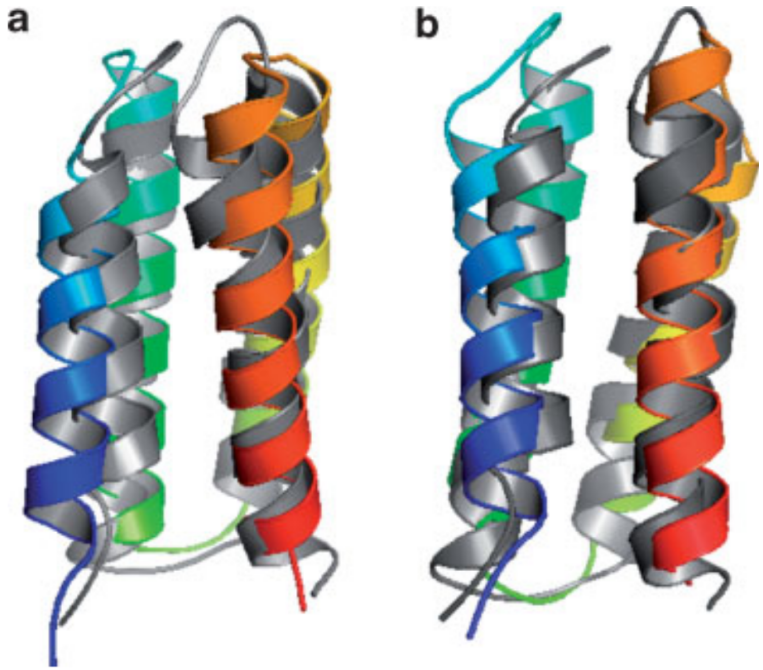
FIGURE 1

Projection into high-dimensional feature space. Using a mapping function Φ, active (empty gray points) and inactive (filled pink points) compounds that are not linearly separable in low-dimensional input space $\mathcal{L}$ **(a)** are projected into high-dimensional feature space $\mathcal{H}$ **(b)** and separated by the maximum-margin hyperplane. Points intercepted by the dotted line are called 'support vectors' (circled points).

A. Lavecchia, *Machine-learning approaches in drug discovery: methods and applications*, Drug Discovery Today, Vol. 20, Nb 3 (**2015**).

Examples of applications (3)

- Protein structure prediction (*e.g.* folding from sequence of amino-acids)



C.A. Floudas, *Computational Methods in Protein Structure Prediction*, Biotechnology and Bioengineering, Vol. 97, No. 2, June 1, **2007**.

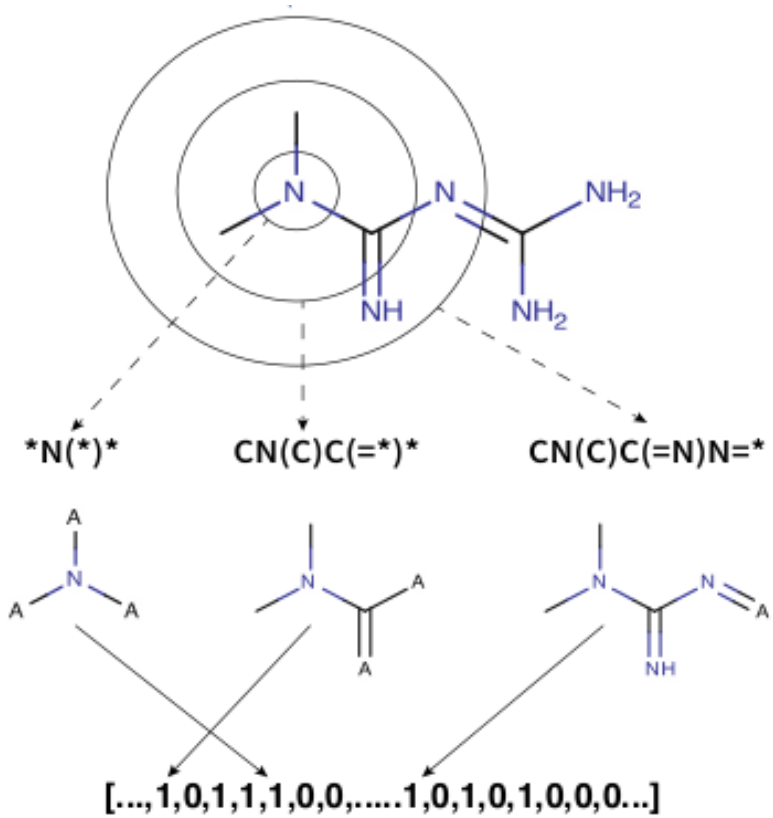
Examples of applications (4)

- QSAR (Quantitative Structure-Activity Relationship) analysis
- "Virtual screening" (search in the (huge) compound space to meet a given target property : *e.g. drugs*)
- Relationship between genome variation (*e.g.* mutation) and disease risk (genotype-phenotype relation, predictive medicine, etc.)

Molecular *descriptors*

- Numerical representation necessary (*e.g.* for input in a neural network) : "common language" to represent molecules.
- Chemical graph theory / chemical fingerprint (*e.g.* by fragments).
- Example : Coulomb matrix

$$C_{ij} = \begin{cases} 0.5 Z_i^{2.4} \text{if i=j} \\ \frac{Z_i Z_j}{|\vec{r_i} - \vec{r_j}|} \text{otherwise} \end{cases}$$



Yu-Chen Lo *et al.*, *Machine learning in chemoinformatics and drug discovery*, Drug Discovery Today, Vol. 23, Nb 8 (**2018**).

Outline

1. Machine learning for drug design and medicine : general introduction

2. Solid harmonic wavelet scattering for predictions of molecule properties (based on Refs. [1, 2])
   - General idea of the method
   - Wavelet scattering coefficients
   - Comparison to usual force fields
   - Why does it work so well : comparison to the multipolar expansion in polarizable force fields methods

General idea : regression of scattering coefficients (1)

- $$\underbrace{\mathbf{x}}_{molecule} = \left\{ \left(\underbrace{r_k}_{position}, \underbrace{z_k}_{nuclear-charge} \right) \in \mathbb{R}^3 \times \mathbb{Z} \right\}_{k \in [\![1,N]\!]} \longmapsto f(\mathbf{x}) \in \mathbb{R}$$
- $f(\mathbf{x})$: physical property of interest of the molecule (*e.g.* total energy)
- $\mathbf{x}$ (all atomic positions and element types) : *possible* molecular descriptor (not unique : cf. drug design context)
- Approximation of $f(\mathbf{x})$ by multilinear regression :

$$\tilde{f}_{a_1,..,a_q}(\mathbf{x}) = b + \underbrace{\sum_{i=1}^{M} \underbrace{a_i}_{\in \mathbb{R}} \mathcal{C}_{\mathbf{p_i}}[\mathbf{x}]}_{linear} \left(+ \underbrace{\sum_{q=2}^{r} \left[\sum_{i_1<..<i_q} a_{i_1,..,i_q} \mathcal{C}_{\mathbf{p_{i_1}}}[\mathbf{x}]..\mathcal{C}_{\mathbf{p_{i_q}}}[\mathbf{x}] \right]}_{multilinear} \right) \quad (1)$$

- $\mathcal{C}_{\mathbf{p_i}}[\mathbf{x}]$: (solid harmonic wavelet) scattering coefficient *i.e.* relevant *molecular descriptor*.

General idea : regression of scattering coefficients (2)

- $(a_1^*, .., a_p^*) = \underset{(a_1,..,a_p)}{\arg\min} \underbrace{\sum_{\nu}}_{Training-set} \left| \tilde{f}_{a_1,..,a_q}(\mathbf{x}_\nu) - f(\mathbf{x}_\nu) \right|^2 : (locally)$

  optimal regression coefficients.
- Training set $\left\{ \left(\underbrace{\mathbf{x}_\nu}_{geometry}, \underbrace{f(\mathbf{x}_\nu)}_{energy} \right) \right\}_\nu$: from ab-initio (DFT)

  calculations.
- Validation set : similar type of data (but different from the training set)

General idea (3)

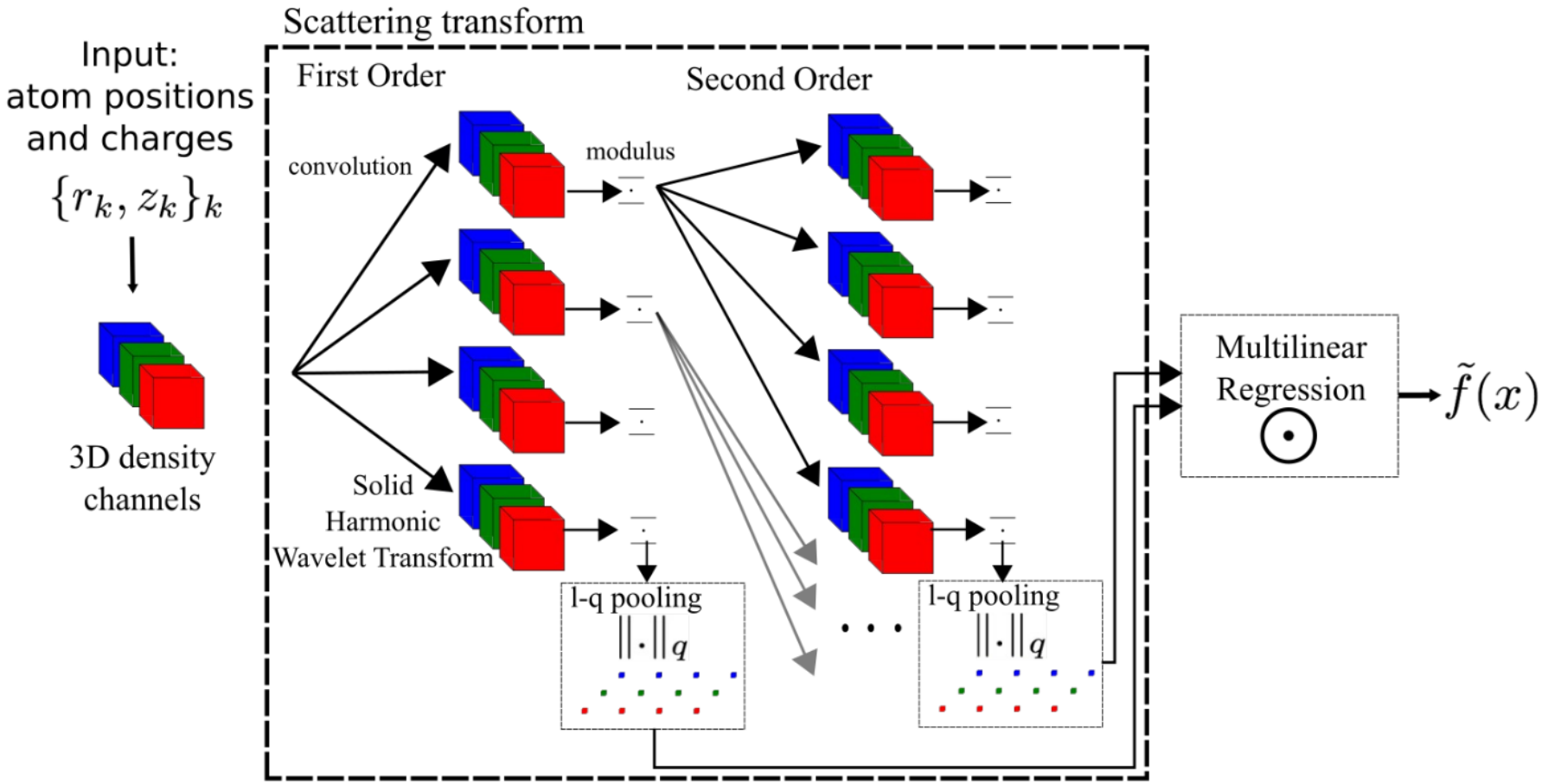


Figure: Convolutional network (two wavelet scattering transform steps) [1, 2]

Computation of wavelet scattering coefficients : outline (1)

- $\underbrace{x}_{geometry} \longrightarrow \underbrace{\rho_x(.)}_{density" channels"} \longrightarrow \underbrace{S\rho_x}_{wavelet-transform}$ (set of coefficients)
- Three density channels (if available in the dataset) : core / valence / bond density
- "Naïve" ("surrogate") core/valence density :

$$\rho_x^{core/valence}(\mathbf{u}) = \sum_k \gamma_k g(\mathbf{u} - \mathbf{r_k}) \tag{2}$$

- $g(\mathbf{v}) = Ke^{-\frac{\mathbf{v}^2}{2\sigma^2}}$, γ_k number of core/valence electrons at atom k (so that $\rho_x(.)$ integrates to $\sum_k \gamma_k$)
- **Invariance on permutation of atom indexes**.

Computation of wavelet scattering coefficients : outline (2)

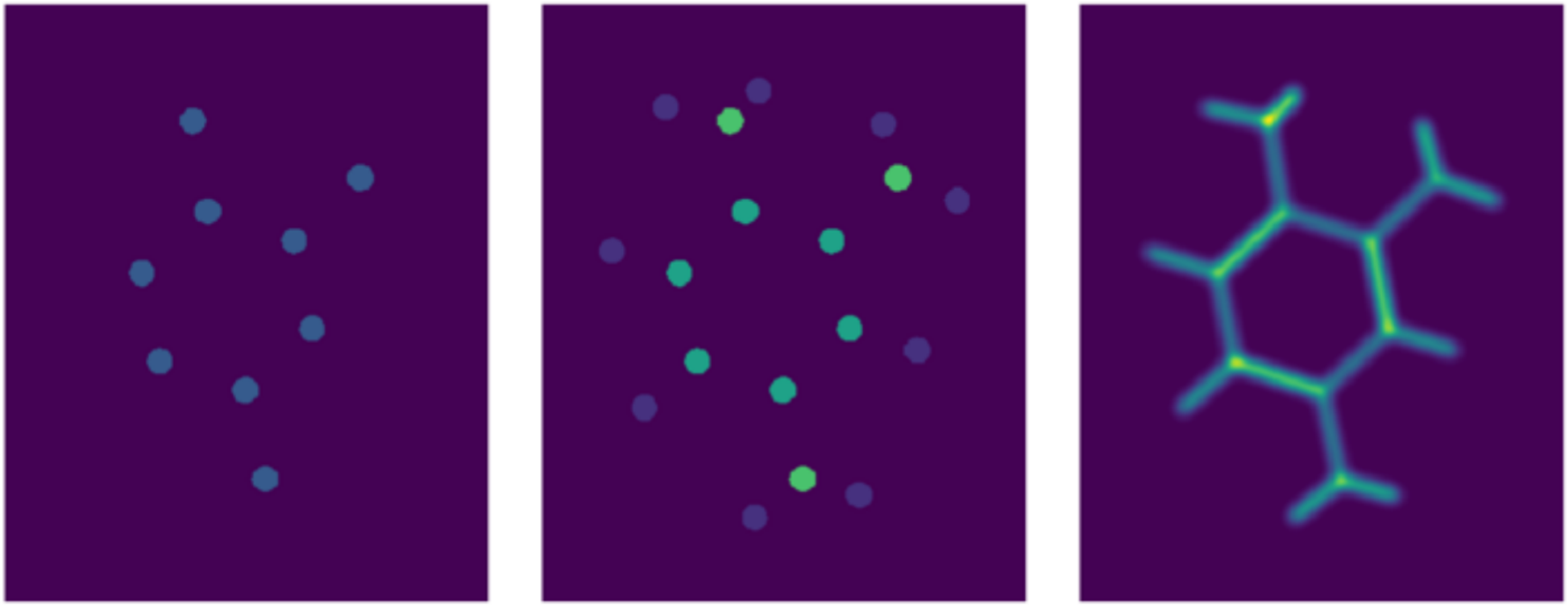


Figure: Three density "channels" (or "guesses") possibly used as input to the convolutional neural network (core, valence and bond channel) [1, 2] – the "full" density channel is also added

Computation of wavelet scattering coefficients : outline (3)

- Surrogate bond density :

$$\rho_x^{bonds}(\mathbf{u}) = C \sum_{i \leftrightarrow j}^{bonds} \frac{\gamma_{ij}}{|\mathbf{r_i} - \mathbf{r_j}|} e^{-\frac{d_{ij}(\mathbf{u})^2}{2d_0^2}} \tag{3}$$

- $d_{ij}(.)$ = distance to the bond (line) $i - j$
- γ_{ij} : number of electrons involved in the bond $i - j$
- No prior, precise knowledge of the electronic density required ("rough" guess).

Solid harmonic wavelets (1)

- Wavelets $\psi(.)$: localized both in time and in frequency.
- Wavelets coefficients / decomposition (over a family of wavelets scaled and translated from a "mother" wavelet) : analogous to Fourier decomposition.
- More efficient to represent signal with discontinuities than Fourier.

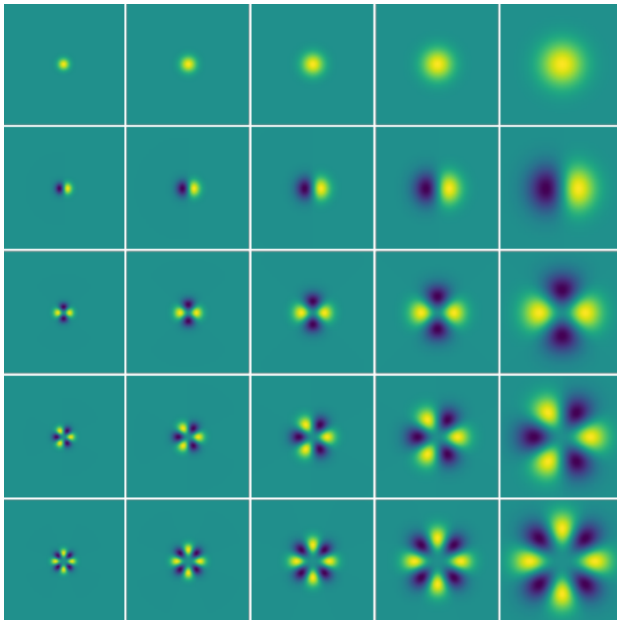


Figure: Real parts of 2D solid hamornic wavelets $\psi^{2D}_{l,j}$ ($\psi^{2D}_{l}(r,\theta) = \frac{1}{2\pi} e^{-\frac{1}{2}r^2} r^l e^{il\phi}$) : **angular momentum** $l = 0...4$ from top to bottom and **scale** $j = 0...4$ from left to right [2]

Solid harmonic wavelets (2)

- Solid harmonic wavelets in 3D :

$$\psi_l^m(\mathbf{u}) = \frac{1}{\left(\sqrt{2\pi}\right)^3} e^{-\frac{|\mathbf{u}|^2}{2}} |\mathbf{u}|^l Y_l^m\left(\frac{\mathbf{u}}{|\mathbf{u}|}\right) \in \mathbb{C} \tag{4}$$

- The Fourier transform of a wavelet is a wavelet :

$$\hat{\psi}_l^m(\omega) = (-i)^l e^{-\frac{1}{2}|\omega|^2} |\omega|^l Y_l^m\left(\frac{\omega}{|\omega|}\right) \tag{5}$$

- Wavelets *scaled* at scale 2^j :

$$\psi_{l,j}^m(\mathbf{u}) = \frac{1}{(2^j)^3} \psi_l^m\left(\frac{\mathbf{u}}{2^j}\right) \tag{6}$$

Wavelet scattering transform : step 1 (1)

- $$\underbrace{\rho(.)}_{density-channel} \underbrace{\longrightarrow}_{1^{st}\,convolution} \rho * \psi^m_{l,j} \underbrace{\longrightarrow}_{modulus-operator} U[j,l]\rho \longrightarrow S\rho[j,l,q]$$

$$U[j,l]\rho : \mathbf{u} \longmapsto \sqrt{\sum_{m=-l}^{l} \left| \left(\rho * \psi^m_{l,j} \right)(\mathbf{u}) \right|^2} \tag{7}$$

$$\boxed{S\rho[j,l,q] = \int_{\mathbb{R}^3} |U[j,l]\rho(\mathbf{u})|^q \, d\mathbf{u}} \in \mathbb{R} \tag{8}$$

- **First order** *solid harmonic wavelet scattering coefficients*.
- **Rotational invariant**.
- Coefficients $S\rho[j,l,q]$ for $q = 2$ exponent : encode pairwise interactions (*e.g.* Coulomb) ? *Why ?* No correlations $\rho(\mathbf{u})\rho(\mathbf{u}')$ ...
- Encode both short (small j) and long-range (large j) interactions contributions to the energy. *Why ?*

Wavelet scattering transform : step 1 (2)

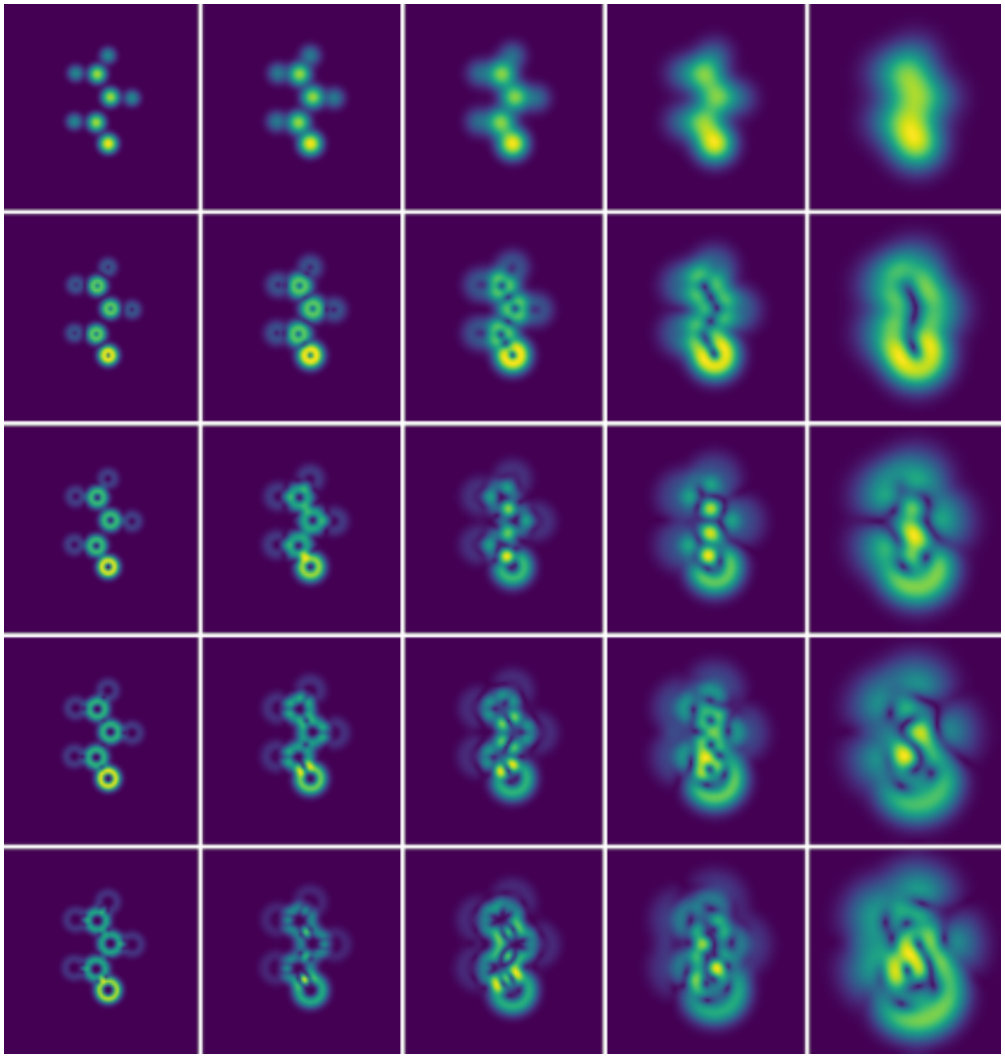


Figure: Solid harmonic wavelet scattering coefficient (moduli $U[j,l]\rho$) [2] : reminiscent of interference patterns / molecular orbitals ?

Wavelet scattering transform : step 1 (3)

- Modulus coefficients (functions) $U[j,l]\rho(.)$ are similar to multipole moments :

$$U[j,l]\rho(\mathbf{u}) = \sqrt{\sum_{m=-l}^{l} \left| \int_{\mathbb{R}^3} \rho(\mathbf{v}) \psi_{l,j}^{m}(\mathbf{u}-\mathbf{v}) d\mathbf{v} \right|^2} \tag{9}$$

$$U[j,l]\rho(\mathbf{u}) = \frac{1}{2^{3j}} \frac{1}{2^{lj}} \sqrt{\sum_{m=-l}^{l} \left| \underbrace{\int_{\mathbb{R}^3} |\mathbf{v}-\mathbf{u}|^l Y_l^m \left(\frac{\mathbf{u}-\mathbf{v}}{|\mathbf{u}-\mathbf{v}|} \right) \rho(\mathbf{v}) e^{-\frac{1}{2}\left(\frac{|\mathbf{u}-\mathbf{v}|}{2^j}\right)^2} d\mathbf{v}}_{\longrightarrow_{j\longrightarrow+\infty} \left(Q_l^m\right)_{\mathbf{u}}^{\mathbb{C}}} \right|^2} \tag{10}$$

- Multipole moments (of a charge distribution) of order l with respect to expansion center $\mathbf{u}$ (using **real** spherical harmonics...) :

$$\left(Q_l^m\right)_{\mathbf{u}} = \int_{\mathbb{R}^3} |\mathbf{v}-\mathbf{u}|^l Y_l^m \left(\frac{\mathbf{v}-\mathbf{u}}{|\mathbf{v}-\mathbf{u}|} \right) \rho(\mathbf{v}) d\mathbf{v} \in \mathbb{R} \tag{11}$$

Wavelet scattering transform : step 1 (4)

- In which sense modulus coefficients (functions) $U[j,l]\rho(.)$ are "*analogous to localized multipole moments*" [1] ?
- If $\forall w \in \{\mathbf{v}|\rho(\mathbf{v}) \neq 0\}, |\mathbf{u}-\mathbf{w}| << 2^j$ (*i.e.* at a large enough scale):

$$U[j,l]\rho(\mathbf{u}) \approx \frac{1}{2^{3j}} \frac{1}{2^{lj}} \sqrt{\sum_{m=-l}^{l} \left| \left(Q_l^m\right)_{\mathbf{u}}^{\mathbb{C}} \right|^2} \tag{12}$$

where $\left(Q_l^m\right)_{\mathbf{u}}^{\mathbb{C}}$ are defined analogously to (real) multipole moments but with complex spherical harmonics.

- Otherwise, for all j, $U[j,l]\rho(\mathbf{u})$ relates to the multipole moments (with respect to center $\mathbf{u}$) of the charge distribution $\rho(.)$ *localized by a gaussian*.

Wavelet scattering transform : step 2

- $U[j,l]\rho \underbrace{\longrightarrow}_{(*)} (U[j,l]\rho) * \psi^m_{l,j'} \underbrace{\longrightarrow}_{modulus} U[j',l]\,(U[j,l]\rho) =^{\mathcal{D}} U[j,j',l]\rho$
- Step $(*)$: second convolution, at larger scales ($j' > j$)

$$U[j,j',l]\rho(\mathbf{u}) = U[j',l]\,(U[j,l]\rho)\,(\mathbf{u}) = \left(\sum_{m=-l}^{l} \left| \left(U[j,l]\rho * \psi^m_{l,j'} \right)(\mathbf{u}) \right|^2 \right)^{\frac{1}{2}} \tag{13}$$

- Final step :

$$U[j,j',l]\rho \longrightarrow \boxed{S\rho[j,j',l,q] = \int_{\mathbb{R}^3} \left| U[j,j',l]\rho(\mathbf{u}) \right|^q d\mathbf{u}} \tag{14}$$

- **Second order** *solid harmonic wavelet scattering coefficients*.
- "Multiscale coupling" coefficient (substructures scales 2^j and $2^{j'}$) *interpreted as van der Waals (dispersion) interaction terms* $\propto C\frac{\alpha_1\alpha_2}{R^6}$...

Final regression

- Case of linear regression :

$$\tilde{f}_{\vec{p}}(\mathbf{x}) = b + \sum_{j}^{\text{Scale}} \sum_{l}^{\text{Moment}} \sum_{q} w_{l,q}^{j} \underbrace{S\rho[j,l,q]}_{local-terms+Coulomb?} + \sum_{j}^{\text{Scale}} \sum_{j'>j}^{\text{Scale}} \sum_{l}^{\text{Moment}} \sum_{q} w_{l,q}^{j,j'} \underbrace{S\rho[j,j',l,q]}_{induction+dispersion?} \quad (15)$$

- "*Analogous to perturbation expansion*" (cf. SAPT theory)
- $\vec{p} = \left\{ w_{l,q}^{j}, w_{l,q}^{j,j'} \right\}_{l=0..L, j=0..J, j<j', q=0..Q}$: set of parameters to optimize
- In fact : sum also on the three different density channels $\rho_x(.)$?
- Relatively small number of scattering coefficients.
- Bilinear regression : additional products of scattering coefficients $S\rho[j_1,l_1,q_1]S\rho[j_2,j_2',l_2,q_2]$, more coefficients to fit.
- Linear *vs.* bilinear *vs.* trilinear regression ? Overfitting ? What is the right "*functional form*" to assume ?

Comparison to usual "force fields"

- "Physical" functional form near equilibrium :

$$U_{\mathbf{p}}\left(\vec{r_1},..,\vec{r_N}\right) = \sum_{bonds} k_b\left(b-b_0\right)^2 + \sum_{angles} k_\theta\left(\theta-\theta_0\right)^2 + \sum_{dihedrals} k_\phi\left(1+\cos(n\phi-\delta)\right) + \sum_{impropers} k_\chi\left(\chi-\chi_0\right)^2 + \underbrace{\sum_{i<j|d(i,j)\geq M} C\frac{q_i q_j}{4\pi\epsilon r_{ij}}}_{electrostatic} + \underbrace{\sum_{i<j|d(i,j)\geq M} \epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{dispersion(vdW)} \quad (16)$$

- Set of parameters to optimize :

$$\mathbf{p} = \left(\left\{k_b^{ij}\right\}\left\{b_0^{ij}\right\}, \left\{k_\theta^{ijk}\right\}\left\{\theta_0^{ijk}\right\}, \left\{k_\phi^{ijkl}\right\}\left\{n^{ijkl}\right\}, \left\{k^{abcd}\right\}, \left\{\chi_0^{abcd}\right\}, \{q_i\}, \{\epsilon_{ij},\sigma_{ij}\}_{i<j|d(i,j)\geq 3}\right) \quad (17)$$

- Cheap computationnaly.
- Used to perform long simulations of molecules (molecular dynamics).

Why does multilinear regression of wavelet scattering coefficients work so well (1) ?

- Electrostatic (**long-range** interaction) energy terms in (*e.g.*) **polarizable** force fields (multipolar expansion of the interaction energy between two multipoles $[q_i, \vec{d}_i, \bar{\bar{Q}}^i]$ and $[q_j, \vec{d}_j, \bar{\bar{Q}}^j]$ at atoms i and j) :

$$\underbrace{\frac{q_i q_j}{\left|\vec{r}_i - \vec{r}_j\right|}}_{charge-charge} \underbrace{-\frac{\left(q_j \vec{d}_i - q_i \vec{d}_j\right).\left(\vec{r}_i - \vec{r}_j\right)}{2\left|\vec{r}_i - \vec{r}_j\right|^3}}_{charge-dipole} + \underbrace{\frac{\vec{d}_i \vec{d}_j}{\left|\vec{r}_i - \vec{r}_j\right|^3} - 3\frac{\left[\vec{d}_i.\left(\vec{r}_i - \vec{r}_j\right)\right]\left[\vec{d}_j.\left(\vec{r}_i - \vec{r}_j\right)\right]}{\left|\vec{r}_i - \vec{r}_j\right|^5}}_{dipole-dipole}$$
$$\underbrace{-\frac{q_i\, Tr\left(\bar{\bar{Q}}^j\right) + q_j\, Tr\left(\bar{\bar{Q}}^i\right)}{\left|\vec{r}_i - \vec{r}_j\right|^3} + 3\frac{q_j\left(\vec{r}_i - \vec{r}_j\right)^T \bar{\bar{Q}}^i\left(\vec{r}_i - \vec{r}_j\right) + q_i\left(\vec{r}_i - \vec{r}_j\right)^T \bar{\bar{Q}}^j\left(\vec{r}_i - \vec{r}_j\right)}{\left|\vec{r}_i - \vec{r}_j\right|^5}}_{charge-quadrupole} \quad (18)$$

- q_i (atomic charge), $\vec{d}_i$ (dipole), $\bar{\bar{Q}}^i$ (quadrupole) are (local) **molecular descriptors**.

Why does multilinear regression of wavelet scattering coefficients work so well (2) ?

- Equation 18 is an approximation (at angular moment of order 2) of :

$$\frac{1}{2}\left(\int \rho_i(\vec{r})V_j(\vec{r})d\vec{r} + \int \rho_j(\vec{r})V_i(\vec{r})d\vec{r}\right) \tag{19}$$

- **Bilinear** regression should work better than **linear** regression : products charge-dipole, dipole-dipole, charge-quadrupole in the "physical" expression of the electrostatic energy *i.e.* **products of two molecular descriptors** only ?
- **Trilinear** should not (physically) give better results than **bilinear** regression – at least concerning description of long-range electrostatic interactions ?

Discussion

- Transferability of the approximated function $\tilde{f}(\mathbf{x})$ to larger molecules ? Increase number of scales (j) ? ($\Longrightarrow$ more scattering terms)
- Transferability to other chemical contexts (*e.g.* chemical reaction $\Longrightarrow$ no training on the bond channel) ?
- Training data set : symmetric configurations of the molecule ? Or impose "by hand" the symmetries (e.g. by symmetry invariant coefficients) as here.
- Interpretation of the first/second order scattering coefficients as physical interaction terms (cf. short-range / long-range interactions) ?
- To what extent (and how) do scattering coefficients "*account for different types of interactions at different scales*" ?

Questions ?