Quantum Energy Regression using Scattering Transforms

Matthew Hirn   MATTHEW.HIRN@ENS.FR
École Normale Supérieure, Département d'Informatique, 45 rue d'Ulm, 75005 Paris, France

Nicolas Poilvert   NUP18@PSU.EDU
Millennium Science Complex, The Pennsylvania State University, University Park, PA 16801 USA

Stéphane Mallat   STEPHANE.MALLAT@ENS.FR
École Normale Supérieure, Département d'Informatique, 45 rue d'Ulm, 75005 Paris, France

Abstract

We present a novel approach to the regression of quantum mechanical energies based on a scattering transform of an intermediate electron density representation. A scattering transform is a deep convolution network computed with a cascade of multiscale wavelet transforms. It possesses appropriate invariant and stability properties for quantum energy regression. This new framework removes fundamental limitations of Coulomb matrix based energy regressions, and numerical experiments give state-of-the-art accuracy over planar organic molecules.

1. Introduction

Estimating the ground state energy of atoms and molecules is one of the most fundamental and studied topics in computational quantum mechanics. The traditional approach to this problem has been to devise clever ways to solve for the equations of quantum mechanics. Recently though, it has been proposed to attack the problem from a Machine Learning (ML) perspective (Rupp et al., 2012).

Most machine learning approaches are representing the molecular state as a Coulomb matrix of pairwise energy terms (Rupp et al., 2012; Hansen et al., 2013). An important limitation of a Coulomb representation is that it depends on an ordering of the atoms in the molecule. When the atom ordering is changed, the Coulomb matrix changes while the energy does not.

A first contribution of this paper is to introduce a new molecular representation in the form of a two or three di-

mensional signal. We define a one-to-one mapping between the molecular state and a real-valued positive function defined over $\mathbb{R}^2$ or $\mathbb{R}^3$, which has the physical interpretation of an approximate electron density. This first step circumvents the issue of atom ordering. In numerical applications, we restrict ourselves to *planar* molecules, with atoms lying on the same molecular plane. We will therefore use two-dimensional electron densities, but three dimensional extensions are calculated similarly.

Regression of a high dimensional functional requires the use of prior knowledge of its invariance and stability properties. For quantum energy regression, many invariance and stability properties of the energy function are known. Imposing these same properties onto a representation is important to obtain accurate regressions. Scattering transforms introduced by (Mallat, 2012) are examples of convolutional networks (Krizhevsky et al., 2012; Sermanet et al., 2013), computed with iterated wavelet transforms, which yield appropriate invariants. Our second contribution shows that these scattering transforms, successfully used for image classification (Bruna & Mallat, 2013; Oyallon & Mallat, 2014), can be used to regress quantum energies to state-of-the-art accuracy.

The paper is organized as follows. Section 2 discusses the properties of "good" molecular representations, and presents the current state-of-the-art along with its known limitations. Section 3 introduces the problem of regressing energies through sparse linear expansions over dictionaries of density functionals. Section 4 gives the mathematical details of a number of invariant representations used in this work. Section 5 describes the setup of the numerical experiments along with the values of all the numerical parameters used in generating the dictionaries in 4. Finally, Section 6 analyses the results of our experiments.

2. Molecular representations for energy regression

High dimensional regressions must take advantage of prior information and invariances of the function that is estimated in order to reduce the problem dimensionality.

We start by outlining properties that should be satisfied by a molecular representation in an energy regression context. Next, the current best-in-class representation based on Coulomb matrices will be presented, along with its known limitations.

2.1. Desirable properties for a molecular representation

We are interested in regressing molecular atomization energies. A molecule containing K atoms is entirely defined by its nuclear charges z_k and its nuclear position vectors p_k indexed by k. Denoting by x the state vector of a molecule, we have

$$x = \{(p_k, z_k) \in \mathbb{R}^3 \times \mathbb{R} : k = 1, \ldots, K\}.$$

Since the target value that we are trying to regress is a scalar representing a physical energy, we know that:

Permutation invariance The energy is invariant to the permutation of the indexation of atoms in the molecule.

Isometry invariance The energy is invariant to translations, rotations, and symmetries of the molecule and hence to any orthogonal operator.

Deformation stability The energy is differentiable with respect to the distances between atoms.

Multiscale interactions The energy has a multiscale structure, with highly energetic covalent bonds between neighboring atoms, and weaker energetic exchanges at larger distances, such as Van Der Waals interactions.

The deformation stability stems from the fact that a small deformation of the molecule induces a small modification of its energy. The primary difficulty is to construct a representation which satisfies these four properties, while simultaneously containing a rich enough set of descriptors to accurately regress the atomization energy of a diverse collection of molecules.

2.2. Coulomb matrix representation

Representations of distributions of points, which are invariant to orthogonal transformations and stable to deformations can be defined from pairwise distances between these points. This is the strategy adopted by the current state-of-the-art in molecular energy regression, which makes use of a so-called *Coulomb matrix* representation (Rupp et al., 2012; Hansen et al., 2013). Given a state vector x, the Coulomb matrix representation C is a function of the pairwise distances $|p_k - p_\ell|$ and of the charge products $z_k z_\ell$

$$C_{k,\ell} = \begin{cases} \frac{1}{2} z_k^{2.4}, & k = \ell, \\ \dfrac{z_k z_\ell}{|p_k - p_\ell|}, & k \neq \ell. \end{cases}$$

This representation thus satisfies the isometry invariance and deformation stability properties. However, it is not invariant to the permutation of the atom indices, which is a priori arbitrary. Although (Hansen et al., 2013) proposes many strategies to mitigate this problem, it remains a challenge to this day. The most successful strategy is to augment the data set by associating to each molecule several permutations of its Coulomb matrix. The final predicted energy is then the average of the predicted energy for each permutation. While this technique improves performance, the data augmentation can significantly increase the size of the data set. In the context of kernel ridge regression, which achieves some of the best reported numbers for Coulomb matrices, this means that the size of the kernel can be very large. Furthermore, the fact that the size of the Coulomb matrix depends on the number of atoms K in the molecule is another limitation. To remedy that issue, a fixed maximum size is set *a priori*, and small Coulomb matrices are padded with zeros on the remaining rows and columns. However, once the training phase is complete, this approach effectively sets an upper bound on the molecular size supported by the representation. Finally, while the Coulomb matrix representation features multiple molecular length scales, it treats them on an equal footing. In particular, it does not take advantage of the multiscale structure of the energy, that emphasizes some scales more than others. It is possible though that the highly nonlinear regressors that couple with Coulomb matrices make up for this shortcoming. Numerical results seem to confirm that intuition, as linear regression with Coulomb matrices is an order of magnitude worse.

3. Energy regression from electronic densities

Hohenberg and Kohn proved in (Hohenberg & Kohn, 1964) that the molecular energy E can be written as a functional of the electron density $\rho(u) \geq 0$ which specifies the density of electronic charge at every point $u \in \mathbb{R}^3$. The minimization of $E(\rho)$ over a set of electron densities ρ leads to the calculation of the ground state energy

$$f(x) = E(\rho_x) = \inf_\rho E(\rho).$$

Hohenberg and Kohn also proved that there is a one to one mapping between $\rho_x(u)$, the minimizing density, and x.

In this work, we consider neutral molecules for which the total electronic charge is equal to the sum of the protonic charges z_k so that $\int \rho_x(u)du = \sum_k z_k$. Computing ρ_x is as difficult as computing $E(\rho_x)$. The next section explains how to replace ρ_x by an approximate density while section 3.2 describes a sparse linear regression of E.

3.1. Electronic density approximations

We use the spirit of the electron density approach by representing x by a crude approximate density $\tilde{\rho}_x$ of ρ_x which also has a one to one mapping with x and satisfies $\int \tilde{\rho}_x(u)du = \sum_k z_k$. By construction, this approximate density is invariant to permutations of atom indices k and its expression is given by:

$$\tilde{\rho}_x(u) = \sum_{k=1}^{K} \rho_{\text{at}}^{a(k)}(u - p_k), \qquad (1)$$

which represents a simple linear superposition of isolated atomic densities. The notation $a(k)$ is a shorthand for the chemical nature of atom k which determines its nuclear charge z_k, and hence which atomic density should be substituted. Isolated atomic densities are pre-computed from Density Functional Theory calculations for every distinct atomic species present in a molecular database. The electron density model (1) only gives a crude approximation to ρ_x. It is a sum of independent atomic contributions and hence does not model any chemical effects like bond sharing. An example of an exact and approximate electron density is shown on Figure 1. The effect of bond sharing in ρ_x appears as higher density "bridges" between atoms. These bridges are almost entirely absent in the case of $\tilde{\rho}_x$.
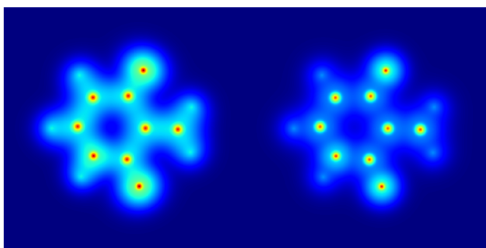

Figure 1. (left) Ground state electron density ρ_x and (right): Approximate electron density $\tilde{\rho}_x$ from 1.

In this work, molecules are bi-dimensional. As a consequence, we transform the three dimensional density in (1) into a two dimensional one, and use that later as our intermediate representation. The dimensionality reduction in the density is obtained by replacing each of the three dimensional atomic densities with two dimensional ones.

That last transformation is obtained by "condensing" the three dimensional charge of a spherical shell of radius r and width dr onto an annulus with the same radius and width.

The resulting approximate density representation $\tilde{\rho}_x$ is invariant to permutations of atom indices but it is not invariant to isometries nor is it multiscale. The regression of the molecular energy $E(\rho_x)$ is therefore not computed from $\tilde{\rho}_x$ but from a representation $\Phi(\tilde{\rho}_x)$ which satisfies these properties. Section 4 will explain how to construct such representations $\Phi(\tilde{\rho}_x) = \{\phi_k(\tilde{\rho}_x)\}_k$, while the next section explains how to use them and approximate the energy $E(\rho_x)$ from a sparse linear regression calculated from a data set of training examples:

$$\tilde{f}(x) = \tilde{E}(\tilde{\rho}_x) = \sum_k w_k \phi_k(\tilde{\rho}_x). \qquad (2)$$

3.2. Sparse Regression by Orthogonal Least Square

Given a training set of N molecular state vectors and associated energies $\{x_i, f(x_i)\}_{1 \le i \le N}$, we explain how to compute the sparse regression in (2). To simplify notations, we shall write $\phi_k(\tilde{\rho}_x) = \phi_k(x)$. A sparse M term regression is obtained by selecting M functionals $\{\phi_{k_m}\}_{1 \le m \le M}$ and computing an optimized linear regression

$$\tilde{f}_M(x) = \sum_{m=1}^{M} w_m \phi_{k_m}(x),$$

which minimizes the quadratic error on training examples

$$\sum_{i=1}^{M} \left| \sum_{m=1} w_m \phi_{k_m}(x_i) - f(x_i) \right|^2. \qquad (3)$$

These M functionals are selected with a greedy orthogonal least square forward selection algorithm (Chen et al., 1991). The procedure selects and orthogonalizes each functional, one at a time.

At the mth iteration it selects ϕ_{k_m}, and a Gram-Schmidt orthogonalization yields an orthonormalized $\phi_{k_m}^{\perp}$, which is uncorrelated relative to all previously selected functionals:

$$\forall n < m, \quad \sum_i \phi_{k_m}^{\perp}(x_i)\, \phi_{k_n}^{\perp}(x_i) = 0,$$

$$\text{and } \sum_i |\phi_{k_m}^{\perp}(x_i)|^2 = 1.$$

The dictionary element $\phi_{k_m}^{\perp}$ is selected so that the linear regression of f over $\{\phi_{k_n}^{\perp}\}_{1 \le n \le m}$ minimizes the quadratic error $\sum_i |\tilde{f}_m(x_i) - f(x_i)|^2$. This is equivalent to finding $\phi_{k_m}^{\perp}$ so that $\sum_i f(x_i) \phi_{k_m}^{\perp}(x_i)$ is maximized. The algorithm can be implemented with a QR factorization, as described in (Blumensath & Davies, 2007). The M-term regression can then be written as a function of the original

functionals ϕ_{k_m} but it is more easily expressed in the orthogonalized Gram-Schmidt basis $\{\phi_{k_m}^{\perp}\}_{m \leq M}$:

$$\tilde{f}_M(x) = \sum_{m=1}^{M} w_m^{\perp} \phi_{k_m}^{\perp}(x)$$

where each coefficient is the correlation:

$$w_m^{\perp} = \sum_i f(x_i)\, \phi_{k_m}^{\perp}(x_i)\ .$$

The parameter M is the dimension of the regression model. Increasing M reduces the bias error but also increases the variance error. The optimal M results from a bias-variance trade-off. It is estimated with a cross validation over training examples.

The bias error is the minimum approximation error of $f(x)$ from an M term linear combination of dictionary vectors $\{\phi_k(x) = \phi_k(\tilde{\rho}_x)\}_k$. This error is small if $f(x)$ is well approximated by an element of a space spanned by M terms of the dictionary. This can be true only if the functionals $\phi_k(\tilde{\rho}_x)$ have the same invariance and stability properties as $f(x)$. The next section explains how to construct such a dictionary.

4. Invariant representations

The central issue is to define a dictionary $\Phi(\tilde{\rho}_x) = \{\phi_k(\tilde{\rho}_x)\}_k$ which is invariant to isometries, stable to deformations, multiscale, and sufficiently rich to perform an accurate regression of the energy $f(x) = E(\rho_x)$. The numerical study is performed over planar molecules. We thus restrict $\tilde{\rho}_x(u)$ over the molecular plane $u \in \mathbb{R}^2$ and normalize it so that $\int_{\mathbb{R}^2} \tilde{\rho}_x(u)du = \sum_k z_k$. Invariance to isometries and stability to deformations is therefore defined in $\mathbb{R}^2$. To understand the challenge of defining such a representation we begin by defining Fourier and wavelet representations and explain the limitations of these two approaches. We then motivate the use of the Scattering representation introduced in (Mallat, 2012), to systematically construct stable invariants for regression. The extension to non-planar molecules in $\mathbb{R}^3$ involves no mathematical difficulty.

4.1. Fourier Invariants

Isometry invariant Fourier type representations based on the bispectrum and spherical harmonics are described in (Bartók et al., 2010; 2013), and are used to regress potential energy surfaces for the dynamics of single molecules. We present a Fourier representation of an arbitrary function $\rho(u)$, which is invariant to linear isometric transformations of u, and discuss some of the limitations of Fourier based representations.

Let $\hat{\rho}$ denote the Fourier transform of ρ:

$$\hat{\rho}(\omega) = \int_{\mathbb{R}^2} \rho(u)e^{-iu\cdot\omega}\,du.$$

The modulus of the Fourier transform is a translation invariant representation of ρ. To add rotation invariance, we take L^p averages over circles of radii $\gamma \in \mathbb{R}^+$:

$$\phi_{\gamma,p}^p(\rho) = \int_{|\omega|=\gamma} |\hat{\rho}(\omega)|^p\,d\omega. \qquad (4)$$

In order to obtain a finite dictionary, we evenly sample the radii up to a maximum radius R, and we build the dictionary out of L^1 and L^2 terms for $p = 1$ and $p = 2$:

$$\Phi_F(\rho) = \{\phi_0(\rho), \phi_{\gamma,1}(\rho), \phi_{\gamma,1}^2(\rho), \phi_{\gamma,2}^2(\rho)\}_\gamma.$$

We use L^1 and L^2 terms to capture linear and quadratic dependencies in the energy functional. In particular, one can prove that an important part of the exact Density Functional $E(\rho)$, namely the Hartree electron-electron repulsion functional, is a weighted sum of L^2 Fourier terms (4) for $p = 2$. These results extend for non-planar molecules by replacing the integrations in $\mathbb{R}^2$ by integrations in $\mathbb{R}^3$.

Numerical results from Table 1 provide the regression error obtained over a Fourier dictionary $\Phi_F(\tilde{\rho}_x)$ computed from the atomic density approximation $\tilde{\rho}_x$ in (1). The Fourier dictionary can regress atomization energies to nearly 10 kcal/mol on average, which is a relatively large error. This indicates that the ground state energy is not well approximated in the functional space generated by the Fourier invariants. The main reason is that the Fourier representation $\Phi_F(\tilde{\rho}_x)$ is not stable to molecular deformations. If x and hence $\tilde{\rho}_x$ is deformed then high frequency terms will experience a large change in value. The same issue of instability to deformations appear for bispectrum representations. This instability can be reduced by replacing the Fourier transform by a windowed Fourier transform as in (Bartók et al., 2013)) for the regression of potential energy surfaces. However, such a representation then only captures localized interactions within the window size, which must not be too large to avoid deformation instabilities. Long range interactions, which represent a non-negligible part of the energy of many aromatic organic molecules are not captured by such representations. Stability to deformations and capturing short range and long range interactions requires a multiscale representation, which motivates the use of a wavelet representation.

4.2. Wavelet Invariants

A wavelet is a complex valued function ψ having zero average. We suppose, additionally, that ψ decays exponentially away from zero and that $\psi(-u) = \psi^*(u)$, where ψ^* denotes the complex conjugate. We utilize Morlet wavelets,

defined as:

$$\psi(u) = \exp\left(-\frac{u_1^2 + u_2^2/\zeta^2}{2}\right)(\exp(i\xi u_1) - C).$$

The slant $\zeta > 1$ yields an anisotropic Gaussian which controls the angular sensitivity of ψ. The Fourier transform of ψ is concentrated in a frequency domain centered at $(\xi, 0)$, while the constant C is set so that $\int \psi = 0$.

Normally, dyadic wavelets are dilated by scales 2^j for $j \in \mathbb{Z}$. We introduce a scale oversampling by a factor 2 and dilate the wavelet at scales $2^{j/2}$. In the following, we are concentrating on wavelets defined in two dimensions $u \in \mathbb{R}^2$. Two dimensional wavelets are rotated by r_θ for an angle $\theta \in [0, 2\pi)$. A dilated and rotated wavelet is then indexed as:

$$\psi_{j,\theta}(u) = 2^{-2j/2}\psi(2^{-j/2}r_{-\theta}u),$$

and the wavelet transform is defined by

$$\rho \mapsto \{\rho \star \psi_{j,\theta}(u)\}_{j,\theta,u}.$$

From the wavelet transform we derive isometry invariant functionals at different scales by averaging over translations $u \in \mathbb{R}^2$ and rotations $\theta \in [0, 2\pi)$:

$$\phi_{j,p}^p(\rho) = \int_{\mathbb{R}^2}\int_0^{2\pi} |\rho \star \psi_{j,\theta}(u)|^p\, d\theta\, du. \tag{5}$$

As with the Fourier dictionary, the wavelet dictionary is made finite by utilizing a finite range of scales and various L^1 and L^2 functionals:

$$\Phi_W(\rho) = \{\phi_0(\rho), \phi_{j,1}(\rho), \phi_{j,1}^2(\rho), \phi_{j,2}^2(\rho)\}_j.$$

The Fourier functionals (4) integrate the frequency energy of ρ over circles of radii γ. The wavelet functionals, however, take advantage of the multiscale structure of the energy and integrate the frequency energy of ρ over annuli of bandwidth $2^{j/2}$. Furthermore, as shown in (Mallat, 2012), the wavelet functionals linearize the action of diffeomorphisms on ρ. Numerical results from Table 1 indicate that the wavelet dictionary requires significantly fewer coefficients to achieve a comparable accuracy as the Fourier dictionary. However, the minimum average error remains similar, as the two dictionaries contain similar frequency information on the density ρ. Thus, while the wavelet dictionary satisfies all of the properties of Section 2.1, it is not a rich enough dictionary to model the energy E to state-of-the-art accuracy. This final limitation is resolved through the scattering transform. These results extend in $\mathbb{R}^3$ with a wavelet transform over $\mathbb{R}^3$.

4.3. Scattering

The scattering transform augments the wavelet dictionary by providing additional multiscale invariants through iterated wavelet transforms. This architecture is a type of deep convolutional network, with variations of it applied successfully in computer vision for texture classification (Sifre & Mallat, 2013; Bruna & Mallat, 2013) as well as object classification (Oyallon & Mallat, 2014).

Deep convolutional networks cascade linear and nonlinear operations through multilayer architectures (Bengio et al., 2013). In the first layer, features from a two dimensional function ρ are extracted via a collection of functionals $\{h_k\}_k$. These functionals apply a localized linear filter L_k across the function ρ via convolution, followed by a nonlinear function F that also downsamples the signal,

$$h_k(\rho) = F(\rho \star \mathrm{L}_k).$$

The linear filters L_k are learned from the training set via back-propagation. The nonlinear functions may be sigmoids, rectifiers or absolute values, followed by a pooling operator. Subsequent layers convolve linear filters both spatially and over the collection of functionals $\{h_k\}_k$ from the previous layer, thus combining information across filters.

Wavelets $\{\psi_{j_1,\theta_1}\}_{j_1,\theta_1}$ are predefined linear filters that capture information at scale $2^{j_1/2}$ and in the direction θ_1. The complex modulus is a pointwise nonlinear function which, when applied after the wavelet transform, yields functionals $\{|\rho \star \psi_{j_1,\theta_1}|\}_{j_1,\theta_1}$ analogous to $\{h_k\}_k$. These functionals are invariant to translation up to scale $2^{j_1/2}$. Since the energy is globally invariant to isometries, the wavelet invariant functionals $\{\phi_{j_1,p}\}_{j_1}$ of (5) are derived from global L^p averages (pooling) of $\{|\rho \star \psi_{j_1,\theta_1}(u)|\}_{j_1,\theta_1}$ over translations u and rotations θ_1. However, this integration removes the variation of $|\rho \star \psi_{j_1,\theta_1}(u)|$ along (u, θ_1), thus discarding a considerable amount of information. In order to recover some of this lost information we apply a second layer of wavelet transforms to the collection of functions $\{|\rho \star \psi_{j_1,\theta_1}(u)|\}_{j_1,\theta_1}$.

For fixed (j_1, θ_1), the function $u \mapsto |\rho \star \psi_{j_1,\theta_1}(u)|$ varies at scales bigger than $2^{j_1/2}$. Translation information at scale $2^{j_1/2}$ and angle θ_1 is recovered by applying a second spatial wavelet transform, with the same Morlet wavelet, for scales $j_2 > j_1$ and angles θ_2, yielding $\{|\rho \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2}\}_{j_2,\theta_2}$. Applying a second wavelet transform to each of the functions from the first layer wavelet transform gives the following collection of second layer spatial functions:

$$\{|\rho \star \psi_{j_1,\theta_1}| \star \psi_{j_2,\theta_2}(u)\}_{j_1,\theta_1,j_2,\theta_2,u}. \tag{6}$$

The collection (6), while recovering lost spatial information, does not recover the angular variability of the first

layer considered as a function $\theta_1 \mapsto |\rho \star \psi_{j_1,\theta_1}(u)|$ for fixed j_1 and u. For scales $2^{j_1/2}$ on the order of the distance between neighboring atoms, the behavior of these functions reflects the orientation of atomic bonds and hence bond angles. The variability over θ_1 for larger spatial scales indicates global geometric structure, such as the orientation of sub-molecules. This rotation information is recovered by applying a wavelet transform over the angles $[0, 2\pi)$ along the rotation variable θ_1. The wavelet transform is defined in terms of circular convolution:

$$g_1 \circledast g_2(\theta) = \int_0^{2\pi} g_1(\theta') g_2(\theta - \theta') \, d\theta'.$$

Periodic dilated wavelets are defined over $[0, 2\pi)$ by periodizing a one dimensional dilated Morlet wavelet $\psi^{1D} : \mathbb{R} \to \mathbb{C}$,

$$\overline{\psi}_{\ell_2}(\theta) = 2^{-\ell_2} \sum_{k \in \mathbb{Z}} \psi^{1D}(2^{-\ell_2}\theta - 2\pi k).$$

The resulting angular part of the second layer transform is then:

$$\{|\rho \star \psi_{j_1, \cdot}(u)| \circledast \overline{\psi}_{\ell_2}(\theta_1)\}_{j_1, \theta_1, j_2, \theta_2, u}. \tag{7}$$

Combining the second layer spatial transform (6) and the second layer angular transform (7), and applying the complex modulus, yields the following collection of functions:

$$\rho \mapsto \{||\rho \star \psi_{j_1, \cdot}| \star \psi_{j_2, \theta_2}(u) \circledast \overline{\psi}_{\ell_2}(\theta_1)|\}_{j_1, \theta_1, j_2, \theta_2, \ell_2, u}. \tag{8}$$

As in the wavelet dictionary, isometry invariant functionals are derived from these second layer wavelet transforms via integration over translations and rotations. Note, however, that a rotation of ρ propagates through the layers of the scattering transform. Indeed, if $\rho_\theta(u) = \rho(r_{-\theta}u)$ is a rotation of ρ by angle θ, then

$$||\rho_\theta \star \psi_{j_1, \cdot}| \star \psi_{j_2, \theta_2}(u) \circledast \overline{\psi}_{\ell_2}(\theta_1)| =$$
$$||\rho \star \psi_{j_1, \cdot}| \star \psi_{j_2, \theta_2 - \theta}(r_{-\theta}u) \circledast \overline{\psi}_{\ell_2}(\theta_1 - \theta)|$$

Thus both angular variables θ_1 and θ_2 are rotationally covariant. However, an orthogonal change of coordinates $(\theta_1, \theta_2) \mapsto (\alpha, \beta)$, with:

$$\alpha = \frac{\theta_1 - \theta_2}{2}, \qquad \beta = \frac{\theta_1 + \theta_2}{2},$$

yields one rotationally invariant variable α and one rotationally covariant variable β. Thus isometry invariant scattering functionals can be derived from (8) by integrating over $(u, \beta) \in \mathbb{R}^2 \times [0, 2\pi)$:

$$\phi^p_{j_1, \lambda_2, p}(\rho) =$$
$$\int_{\mathbb{R}^2} \int_0^{2\pi} ||\rho \star \psi_{j_1, \cdot}| \star \psi_{j_2, \beta - \alpha}(u) \circledast \overline{\psi}_{\ell_2}(\alpha + \beta)|^p \, d\beta \, du,$$

where $\lambda_2 = (j_2, \alpha, \ell_2)$ encodes the parameters of the second layer. A finite scattering dictionary is obtained by taking a finite number of scales j_1, j_2, ℓ_2 as well as a finite number of angles α, for a mixture of L^1 and L^2 functionals:

$$\Phi_S(\rho) = \{\phi_0(\rho), \phi_{j_1, 1}(\rho), \phi^2_{j_1, 1}(\rho), \phi^2_{j_1, 2}(\rho), \dots$$
$$\dots \phi_{j_1, \lambda_2, 1}(\rho), \phi^2_{j_1, \lambda_2, 1}(\rho), \phi^2_{j_1, \lambda_2, 2}(\rho)\}_{j_1, \lambda_2}.$$

The second layer of the scattering transform greatly expands the wavelet dictionary, but still satisfies the four properties of Section 2.1. Numerical experiments show (see Table 1) that the average regression error over the scattering dictionary is greatly reduced, to approximately 1.8 kcal/mol, thus indicating that the second layer functionals dramatically increase the ability of the dictionary to model the full variability of the energy. The extension of a scattering transform in $\mathbb{R}^3$ is done with three-dimensional spatial wavelet transforms, indexed by a direction $\vec{\theta}$ which belongs to the two-dimensional sphere in $\mathbb{R}^3$. The second order scattering coefficients are then calculated with a wavelet transform along $\vec{\theta}$, and thus on the two-dimensional sphere in $\mathbb{R}^3$ (Starck et al., 2006).

5. Numerical experiments

We compare the performance of Coulomb matrix representations with Fourier, wavelet and Scattering representations on two databases of *planar* organic molecules. Molecular atomization energies from these databases were computed using the PBE0 hybrid density functional (Adamo & Barone, 1999). The first database includes 454 nearly planar molecules among the 7165 molecules of the QM7 molecular database (Rupp et al., 2012). We also created a second database of 4357 strictly planar molecules, which we denote QM2D. We produced this new database with a similar procedure as the one outlined in (Rupp et al., 2012). Both databases consist of a set of organic molecules composed of hydrogen, carbon, nitrogen, oxygen, sulfur and additionally chlorine in the case of QM2D. The molecules featured in these sets cover a large spectrum of representative organic groups typically found in Chemical Compound Space (Rupp et al., 2012). Particular care was taken in producing well-balanced folds used in cross validation assessments. A proper partitioning of the data points among folds was indeed outlined in (Hansen et al., 2013) as critical to ensure low variance in test errors.

5.1. Coulomb matrix baseline and error metrics

In the case of the Coulomb matrix based regression we used the best performing representation assigning eight randomly sorted Coulomb matrices per molecule as described in (Hansen et al., 2013). The width σ of the Laplace

kernel and the ridge parameter λ were selected following the algorithm described in the same paper. The algorithm was validated by recovering the published accuracy over the full QM7 database which contains 7165 molecules. In our experiments, we restrict the database size to only 454 planar molecules so the regression error is larger. The same methodology is then followed to compute the optimal Coulomb matrix based regression on QM2D.

To evaluate the precision of each regression algorithm, each database is broken into five representative folds, and all tests are performed using five fold cross validation in which we reserve four folds for training, and the fifth fold for testing. This results in a regressed energy for each molecule in the database. We report the Mean Absolute Error (MAE or ℓ^1) over each database along with the Root Mean-Square Error (RMSE or ℓ^2), which is the square root of the average squared error.

5.2. Dictionary Implementations and Sizes

The number of elements in the Fourier, wavelet, and scattering dictionaries are very different and respectively equal to 1537, 61 and 11071 in numerical computations. This section explains the implementation of these dictionaries. Molecular configurations are centered at zero, and the two dimensional electron density $\tilde{\rho}_x$ is restricted to a box $[-a, a]^2$. The parameter a is chosen so that the density decays to nearly zero at the boundary (in our experiments, $a = 11$ angstroms). The box is then sampled with $2^J \times 2^J$ evenly spaced grid points, for some resolution J; in practice $J = 10$ for the Fourier and wavelet dictionaries, and $J = 9$ for the scattering dictionary. The grid naturally leads us to a discretized version of $\tilde{\rho}_x$.

For the Fourier representation, discrete Fourier transforms are computed with a two dimensional Fast Fourier Transform (FFT). Integration over circles of radii γ is approximated with finite sums over discretized circular contours, which are approximated using the original $2^J \times 2^J$ spatial sampling. This yields 2^{J-1} functionals for a fixed L^p average, resulting in a dictionary of size $1 + 3 \cdot 2^{J-1}$. For $J = 10$, we have 1537 dictionary elements.

The wavelet parameters are set according to the following specifications. The minimum scale is $j_{\min} = 0$ and the maximum scale is $j_{\max} = J - 1/2$, resulting in a total of $2J$ scales for the wavelet transform. L angles are evenly sampled from $[0, \pi)$ according to $\theta = k\pi/L$ for $k = 0, \ldots, L - 1$, which is equivalent to evenly sampling $2L$ angles over $[0, 2\pi)$ since the Morlet wavelet is symmetric relative to both the x and y axes. In practice we take $L = 8$. The slant is $\zeta = 1/2$, and the central frequency is fixed at $\xi = 3\pi/4$. Given a fixed slant ζ and central frequency ξ, increasing J and L should not decrease the accuracy of the regression. As a consequence, we fixed

the values for these parameters when the regression results appeared to plateau. Wavelet transform convolutions are computed as multiplications in frequency space by utilizing FFTs and inverse FFTs. The modulus of the output functions from these convolutions are then averaged over the discrete spatial grid and the discrete sampling of $[0, \pi)$ to obtain the wavelet dictionary functionals. This results in a wavelet dictionary of size $1 + 6J$. For $J = 10$, this yields 61 dictionary elements

For the second layer of the scattering transform, scales j_2 are computed for $j_2 > j_1$, resulting in $2J(2J - 1)/2$ pairs (j_1, j_2). Angles θ_2 are discretized with L samples, and the wavelet transform over $[0, \pi)$ is computed for dyadic scales 2^{ℓ_2} with $\ell_2 = 0, \ldots, \log_2 L$. The total size of the scattering dictionary is then $1 + 6J + 3(2J(2J - 1)L \log_2 L)/2$. For $J = 9$ and $L = 8$, the scattering dictionary contains 11071 functionals.

6. Results and discussion

The results from our numerical experiments are summarized in Table 1. The Scattering representation computed from the atomic density model offers state-of-the-art accuracy. Over the larger database, it has a 25% improvement in Mean Absolute Error over the best Coulomb matrix based regression technique. This improvement goes up to 53% in Root Mean Square Error. The larger improvement of the RMSE is due to the fact that a scattering regression has smaller error outliers.
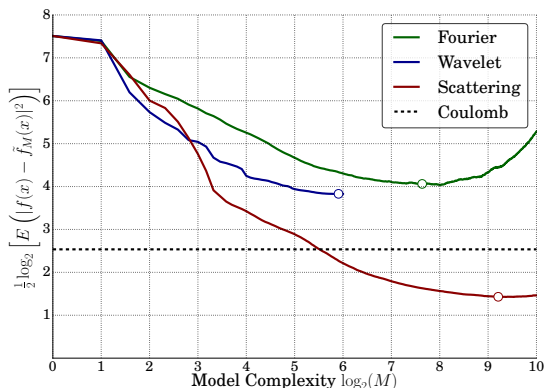
Figure 2. Decay of the log RMSE error $\frac{1}{2} \log_2 \left[E \left(|f(x) - \tilde{f}_M(x)|^2 \right) \right]$ over the larger database of 4357 molecules, as a function of $\log_2(M)$ in the Fourier (green), Wavelet (blue) and Scattering (red) regressions. The dotted line gives the Coulomb regression error for reference.

Table 1 shows that the error of Fourier and wavelet regression are of the same order although the Fourier dictionary

Table 1. Error in kcal/mol of regressed quantum molecular energies using different molecular representations (vertically) and different error measures (horizontally), on two databases of planar organic molecules: left and right parts of the table.

	454 2D molecules from QM7			4357 molecules in QM2D		
	M	ℓ^1: MAE	ℓ^2: RMSE	M	ℓ^1: MAE	ℓ^2: RMSE
Coulomb	N/A	7.0	20.5	N/A	2.4	5.8
Fourier	62	11.9	16.1	198	11.1	16.7
Wavelet	42	11.1	15.5	59	11.1	14.2
Scattering	74	**6.9**	**9.0**	591	**1.8**	**2.7**
Chemical Accuracy	1.0					

has 1537 elements and the wavelet dictionary has only 61. Figure 2 gives the decay of these errors as a function of M. This expected error is computed on testing molecules. The circles on the plot give the estimated value of M which yield a minimum regression error by cross-validation over the training set (reported in Table 1). Although the Fourier and wavelet regressions reach nearly the same minimum error, the decay is much faster for wavelets. When going from the smaller to the larger database, the minimum error of the Fourier and wavelet regressions remain nearly the same. This shows that the bias error due to the inability of these dictionaries to precisely regress $f(x)$ is dominating the variance error corresponding to errors on the regression coefficients. The Coulomb and Scattering representations on the other hand, achieve much smaller bias errors on the larger database.

The number of terms of the scattering regression is $M = 591$ on the larger database, although the dictionary size is 11071. A very small proportion of scattering invariants are therefore selected to perform this regression. The chosen scattering coefficients used for the regression are coefficients corresponding to scales which fall between the minimum and maximum pairwise distances between atoms in the molecular database. These selected coefficients are thus adapted to the molecular geometries.

7. Conclusion

This paper introduced a novel intermediate molecular representation through the use of a model electron density. The regression is performed on a scattering transform applied to a model density built from a linear superposition of atomic densities. This transform is well adapted to quantum energy regressions because it is invariant to the permutation of atom indices, to isometric transformations, it is stable to deformations, and it separates multiscale interactions. It is computed with a cascade of wavelet convolutions and modulus non-linearities, as a deep convolutional network. State-of-the-art regression accuracy is obtained over two databases of two-dimensional organic molecules, with a relatively small number of scattering vectors. Understanding the relation between the choice of scattering coefficients and the physical and chemical properties of these molecules is an important issue.

Numerical applications have been carried out over planar molecules, which allows one to restrict the electronic density to the molecular plane, and thus compute a two-dimensional scattering transform. A scattering transform is similarly defined in three dimensions, with the same invariance and stability properties. It involves computing a wavelet transform on the two-dimensional sphere S^2 in $\mathbb{R}^3$ (Starck et al., 2006) as opposed to the circle S^1. It entails no mathematical difficulty, but requires appropriate software implementations which are being carried out.

Energy regressions can also provide estimations of forces through differentiations with respect to atomic positions. Scattering functions are differentiable and their differential can be computed analytically. However, the precision of such estimations remain to be established.

References

Adamo, C. and Barone, V. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *The Journal of Chemical Physics*, 110(6158), 1999.

Bartók, A.P., Payne, M.C., Kondor, R., and Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403(4), 2010.

Bartók, A.P., Kondor, R., and Csányi, G. On representing chemical environments. *Physical Review B*, 87(18): 184115(16), 2013.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Blumensath, T. and Davies, M.E. On the difference between orthogonal matching pursuit and orthogonal least squares, 2007. URL http://eprints.soton.ac.uk/142469/1/BDOMPvsOLS07.pdf.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE Trans. on PAMI*, 35(8):1872–1866, August 2013.

Chen, S., Cowan, C.F.N., and Grant, P.M. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2): 302–309, 1991.

Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, O.A., Tkatchenko, A., and Müller, K.R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013. doi: 10.1021/ct400195d. URL http://dx.doi.org/10.1021/ct400195d.

Hohenberg, P. and Kohn, W. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, 1964. doi: 10.1103/PhysRev.136.B864. URL http://link.aps.org/doi/10.1103/PhysRev.136.B864.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Proc. of Neural Information Processing Systems*, 2012.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Oyallon, E. and Mallat, S. Deep roto-translation scattering for object classification. arXiv:1412.8659, 2014.

Rupp, M., Tkatchenko, A., Müller, K.R., and von Lilienfeld, O.A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, 2012. doi: 10.1103/PhysRevLett.108.058301. URL http://link.aps.org/doi/10.1103/PhysRevLett.108.058301.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013. URL http://arxiv.org/abs/1312.6229.

Sifre, L. and Mallat, S. Rotation, scaling and deformation invariant scattering for texture discrimination. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

Starck, J.-L, Moudden, Y., Abrial, P., and Nguyen, M. Wavelets, ridgelets and curvelets on the sphere. *Astronomy and Astrophysics*, 446:1191–1204, 2006.