

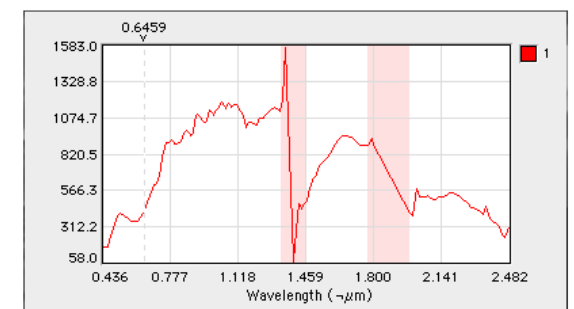
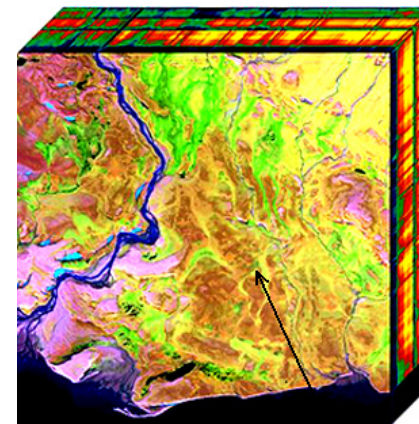
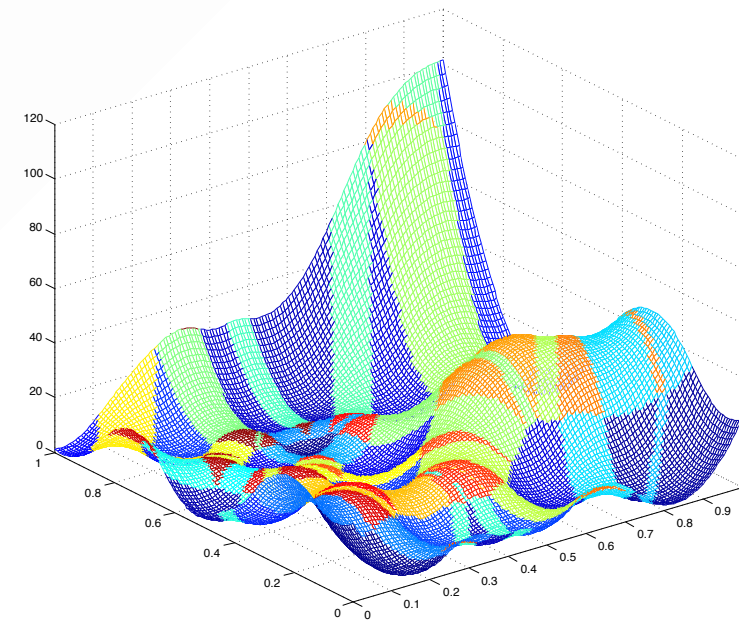
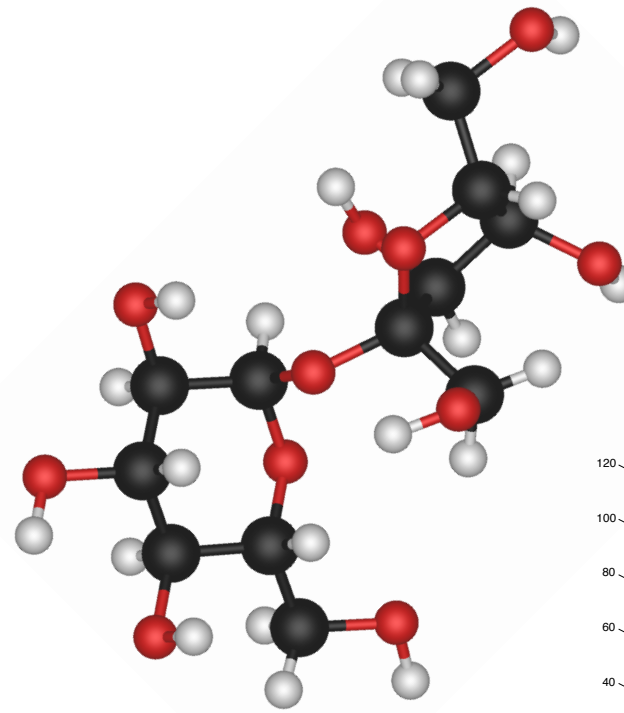
# Interpolation for Physical Big Data

Matthew Hirn  
École normale supérieure  
Département d'Informatique

City College of New York  
February 18, 2015

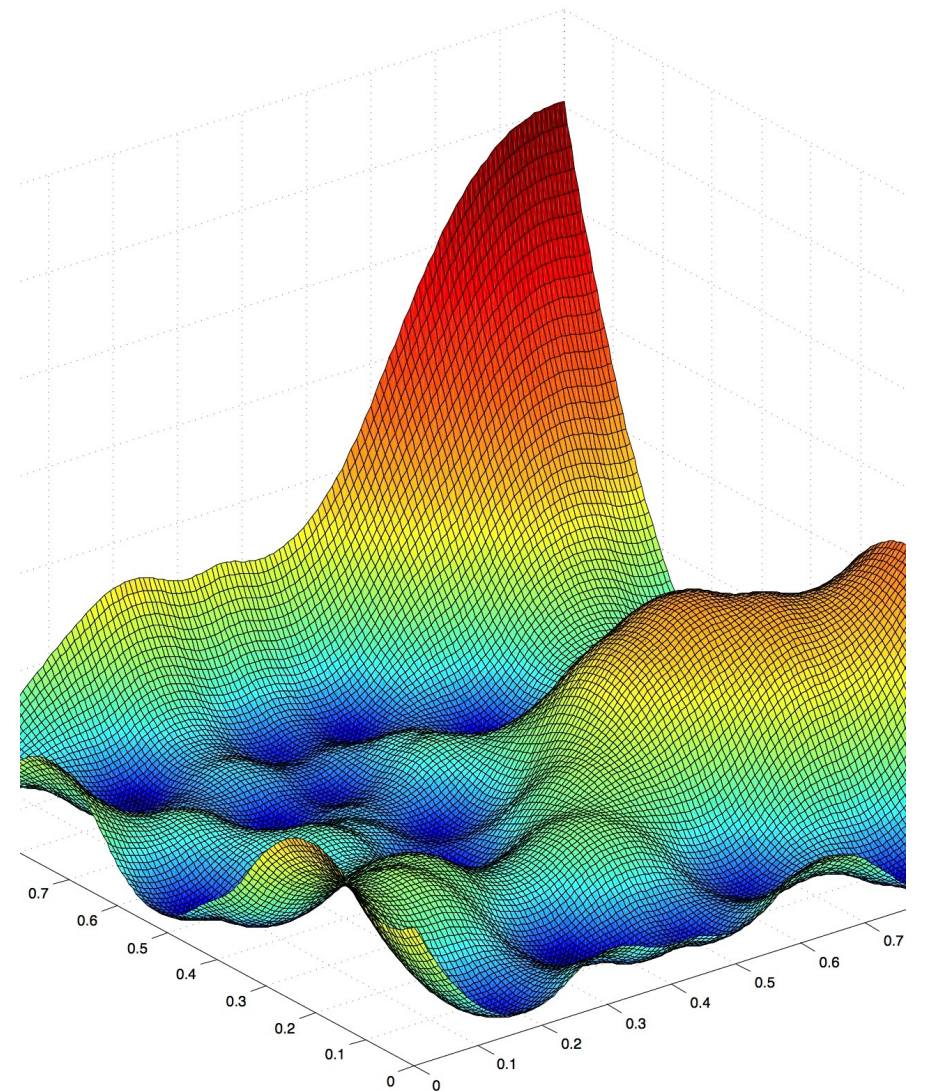
# Motivation

- Big Data: Massive amounts of high dimensional data
- Audio, medical, images, hyperspectral, video, dynamical systems, quantum chemistry
- Want to learn important features of new data fast
- Interplay between discrete and continuous at the interface of analysis, geometry, computer science, statistics, chemistry, physics



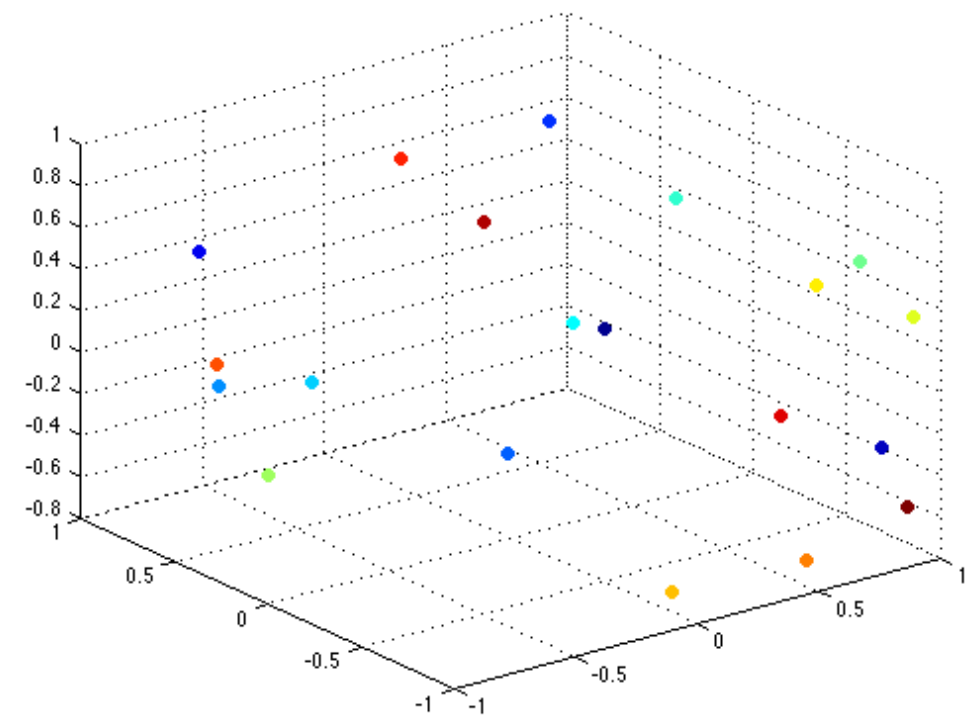
# Curse of Dimensionality

- High dimensional  $x \in X \subset \mathbb{R}^d$
- Approximate a function  $f(x)$  given  $n$  samples  $\{x_i, f(x_i)\}_i$
- $f(x)$  can be approximated from the samples by local interpolation if  $f$  is regular and there are close examples
- Need  $n = \epsilon^{-d}$  points to cover  $[0, 1]^d$  with an  $\epsilon$ -net



# Curse of Dimensionality

- High dimensional  $x \in X \subset \mathbb{R}^d$
- Approximate a function  $f(x)$  given  $n$  samples  $\{x_i, f(x_i)\}_i$
- $f(x)$  can be approximated from the samples by local interpolation if  $f$  is regular and there are close examples
- Need  $n = \epsilon^{-d}$  points to cover  $[0, 1]^d$  with an  $\epsilon$ -net





# Regularity Priors

- Geometric Regularity

**Manifold Learning:**  $X$  lies on a low dimensional smooth Riemannian manifold  $\mathcal{M}$

$$f : \mathcal{M} \rightarrow \mathbb{R}, \quad \dim(\mathcal{M}) \ll d$$

- Function Regularity

**Minimal Interpolants:** Specify a functional space  $\mathcal{F}(X)$  and look for the function  $\tilde{f} \in \mathcal{F}(X)$  that best interpolates  $\{x_i, f(x_i)\}_i$

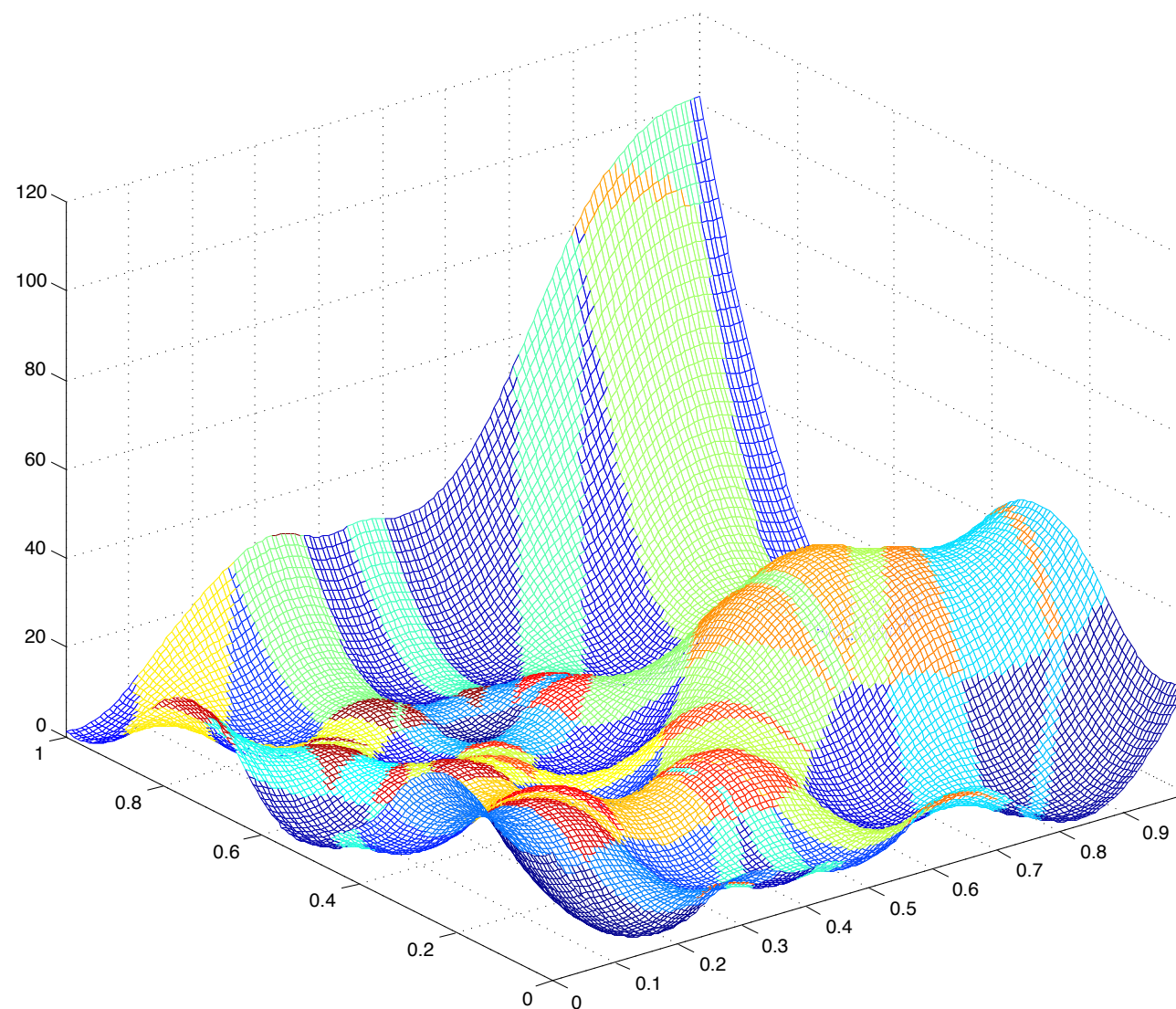
$$\tilde{f} = \arg \min \{ \|g\|_{\mathcal{F}} : g(x_i) = f(x_i) \quad \forall i \}$$

- Geometric and Function Regularity

**Sparse Regression:** Sparse linear expansion over a family of functions  $\Phi = \{\phi_j\}_j$  that share the regularity properties of  $f$

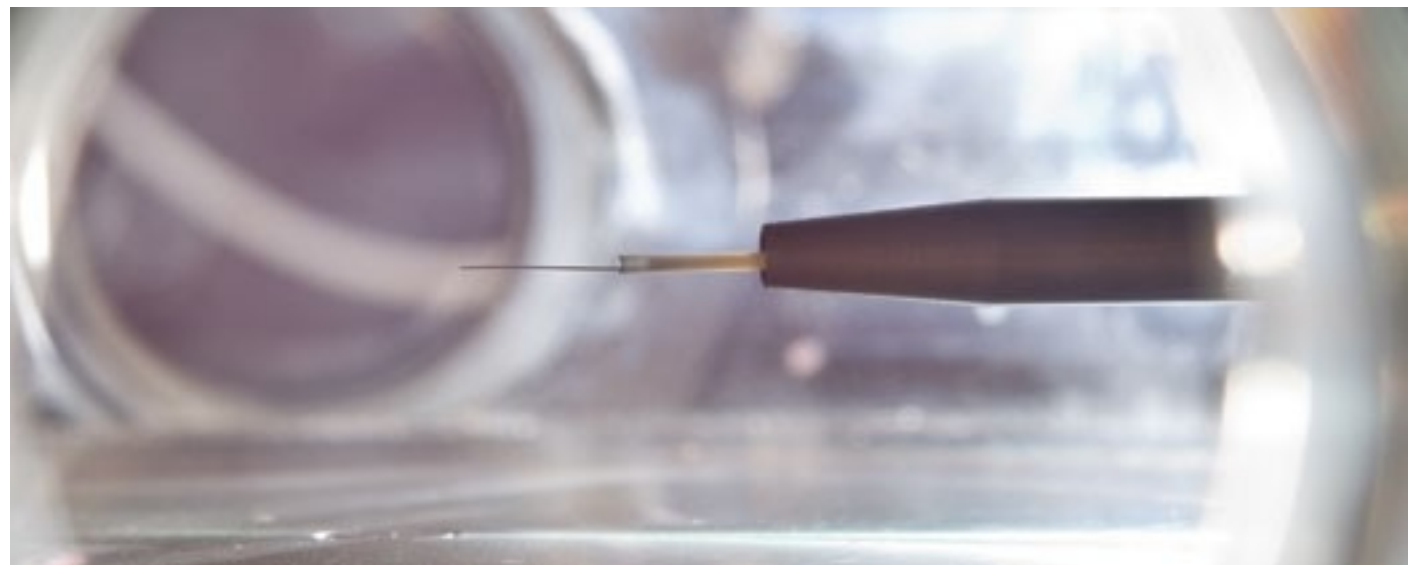
$$f(x) \approx \tilde{f}(x) = \sum_j \alpha_j \phi_j(x)$$

# Minimal Interpolants (Whitney Extensions)



# Possible Application in Physics

Suppose a scientist wants to conduct a costly experiment in which he must deposit a thin film of  $\text{SiO}_2$  in a special tool that has plasma in it. The success of this experiment depends on several factors, such as the pressure in the chamber, the temperature of the substrate, the voltage of the plasma, and the ratios of the gases involved. He wants to find the optimal conditions for performing the experiment. He knows that the voltage is a smooth function of the other parameters, but it is difficult to measure. Consequently, he can only measure it for a few different combinations of initial conditions. He varies each parameter slightly while holding the others constant to find the rate of change of the voltage with respect to that parameter. Now he has data points (configurations of the parameters), function values (measured voltages), and partial derivatives. Using a Whitney type interpolation algorithm, it would be possible to compute a good estimate of the voltage for any configuration of the parameters and thereby determine the optimal conditions for the experiment.



# Whitney's Extension Theorem 1934

- Suppose  $m \in \mathbb{N}$  and  $E \subseteq \mathbb{R}^d$  is closed. Assume for each  $x \in E$  there exists a polynomial

$$P_x(y) = \sum_{|\beta| \leq m} a_\beta y^\beta, \quad a_\beta \in \mathbb{R}$$

so that

$$\lim_{|x-y| \rightarrow 0} |\partial^\alpha P_y(y) - \partial^\alpha P_x(y)| |x-y|^{-(m-|\alpha|)} = 0$$

uniformly on compact subsets of  $E$  for each multi-index  $\alpha$  with  $|\alpha| \leq m$ . Then there exists a function  $\tilde{f} \in C^m(\mathbb{R}^d)$  such that

$$\partial^\alpha \tilde{f}(x) = \partial^\alpha P_x(x), \quad \forall \quad |\alpha| \leq m, \quad x \in E$$

# Minimal Lipschitz Extensions

- Smoothness prior:  $f \in C^{m-1,1}(\mathbb{R}^d)$   
Want the “best” (minimal) interpolant for  $\{x_i, f(x_i)\}_i$
- $C^{0,1}(\mathbb{R}^d)$  = Lipschitz functions:  $\text{Lip}(f) < \infty$
- Whitney, McShane 1934: Simple construction of interpolant:

$$\tilde{f}(x) = \inf_i (f(x_i) + M|x - x_i|), \quad \text{Lip}(\tilde{f}) = M$$

- Minimum Lipschitz constant is

$$\text{Lip}(\tilde{f}) = \max_{i \neq j} \frac{|f(x_i) - f(x_j)|}{|x_i - x_j|}$$

Isometric Extension  
(Lipschitz constant  
is preserved)

- Aronsson 1967: Absolutely Minimal Lipschitz Extensions (AMLE)
  - Applications to inpainting, PDEs, computer science



# Absolutely Minimal Lipschitz Extensions

- The minimal Lipschitz extension is not unique! The Absolutely Minimal Lipschitz Extension (AMLE) is the locally best Lipschitz extension

- Aronsson Definition 1967: Let  $f : E \rightarrow \mathbb{R}$  be Lipschitz, and let  $\tilde{f} : \mathbb{R}^d \rightarrow \mathbb{R}$  be an isometric extension, i.e.,  $\text{Lip}(\tilde{f}) = \text{Lip}(f)$ . The function  $\tilde{f}$  is an AMLE if for every open subset  $V \subseteq \mathbb{R}^d \setminus E$ ,

$$\text{Lip}(\tilde{f}|_V) = \text{Lip}(\tilde{f}|_{\partial V})$$

- AMLEs are very interesting functions:
  - Infinity harmonic functions (Aronsson 1967; Jensen 1993)
  - Stochastic games of chance (Tug-of-war: Peres, Schramm, Sheffield, Wilson 2009)
  - Applications to inpainting (Caselles, Haro, Sapiro, Verdera 2006) and sandpile shapes (Aronsson, Evans, Wu 1996)

# Minimal Interpolants in $C^{1,1}(\mathbb{R}^d)$

- $C^{1,1}(\mathbb{R}^d)$  = Derivative is Lipschitz:  $\text{Lip}(\nabla f) < \infty$
- Significantly more complicated than Lipschitz extensions

- Specify function values  $f(x_i)$  and (optionally) gradients  $y_i \in \mathbb{R}^d$   
Interpolation means:

$$\tilde{f}(x_i) = f(x_i) \text{ and } \nabla \tilde{f}(x_i) = y_i$$

- Wells 1973: Construction of interpolant  $\tilde{f}$  with specified  $\text{Lip}(\nabla \tilde{f}) = M$
- Le Gruyer 2009: Minimum value of  $\text{Lip}(\nabla \tilde{f})$
- H. and Le Gruyer 2014: AMLEs for  $C^{1,1}(\mathbb{R}^d)$
- Herbert-Voss, H., McCollum 2015: Efficient, practical algorithms

# The Minimal Value of $\text{Lip}(\text{grad } F)$

- Le Gruyer 2009: The minimum value of  $\text{Lip}(\nabla \tilde{f})$  is:

$$\inf_{\substack{\tilde{f}(x_i)=f(x_i) \\ \nabla \tilde{f}(x_i)=y_i}} \text{Lip}(\nabla \tilde{f}) = \Gamma^1(P) = 2 \sup_{z \in \mathbb{R}^d} \max_{i \neq j} \frac{|P_i(z) - P_j(z)|}{|z - x_i|^2 + |z - x_j|^2}$$

$$P_i(z) = f(x_i) + y_i \cdot (z - x_i)$$

- The  $\Gamma^1$  functional can be interpreted as the Lipschitz constant for 1-fields. Indeed, the 1-field

$$x \mapsto J_x \tilde{f}(z) = \tilde{f}(x) + \nabla \tilde{f}(x) \cdot (z - x)$$

extends the given 1-field  $x_i \mapsto P_i$  while preserving  $\Gamma^1$ , i.e.,

$$\Gamma^1(J.\tilde{f}) = \Gamma^1(P)$$

# Non-Scalar Valued AMLEs

- Non-scalar valued AMLE results are difficult...
  - Tree valued functions (Naor and Sheffield 2012)
  - Vector valued functions (Sheffield and Smart 2012)
- **Theorem** (H. and Le Gruyer 2014): Suppose that  $X$  is a complete, proper, midpoint convex, and distance convex metric space,  $Z$  is a complete metric space,  $E \subset X$ , and  $f : E \rightarrow Z$ . Let  $\Phi$  be a functional for which an isometric extension always exists. Then, for each  $\epsilon > 0$ , there exists a function  $\tilde{f} : X \rightarrow Z$  such that  $\tilde{f}|_E = f$ ,  $\Phi(\tilde{f}) = \Phi(f)$ , and
$$|\Phi(\tilde{f}|_V) - \Phi(\tilde{f}|_{\partial V})| < \epsilon, \quad \forall \text{ open } V \subset X \setminus E$$
- Examples:
  - $X$  and  $Z$  are Hilbert spaces,  $\Phi = \text{Lip}$
  - $Z$  is metrically convex with the binary intersection property,  $\Phi = \text{Lip}$
  - 1-fields:  $X = \mathbb{R}^d$ ,  $Z = \mathcal{P}^1$ , and  $\Phi = \Gamma^1$
  - $C^{m,1}(\mathbb{R}^d)$ ,  $m > 1$ ?

# Computation of Whitney Extensions

- Given  $\{x_i, f(x_i)\}_i$ , efficiently compute an interpolant  $\tilde{f} \in C^m(\mathbb{R}^d)$  such that  $\tilde{f}(x_i) = f(x_i)$
- If  $M^*$  is the best possible norm for an interpolant of  $\{x_i, f(x_i)\}_i$ , want
$$c(m, d)M^* \leq \|\tilde{f}\| \leq C(m, d)M^*$$
- Fefferman and Klartag 2009: Algorithm to compute such an  $\tilde{f}$ :
  - $O(n)$  storage
  - $O(n \log n)$  time to return  $\|\tilde{f}\|$  and lay the groundwork for  $\tilde{f}$
  - $O(\log n)$  time to evaluate  $\tilde{f}(x)$  for any  $x \in \mathbb{R}^d$
  - Super-exponential constants! (beautiful, but not practical)



# Computing $\text{Lip}(\nabla \tilde{f})$

- Le Gruyer 2009: The minimum value of  $\text{Lip}(\nabla \tilde{f})$  is:

$$\inf_{\tilde{f}} \text{Lip}(\nabla \tilde{f}) = \Gamma^1 = 2 \sup_{x \in \mathbb{R}^d} \max_{i \neq j} \frac{|P_i(x) - P_j(x)|}{|x - x_i|^2 + |x - x_j|^2}$$

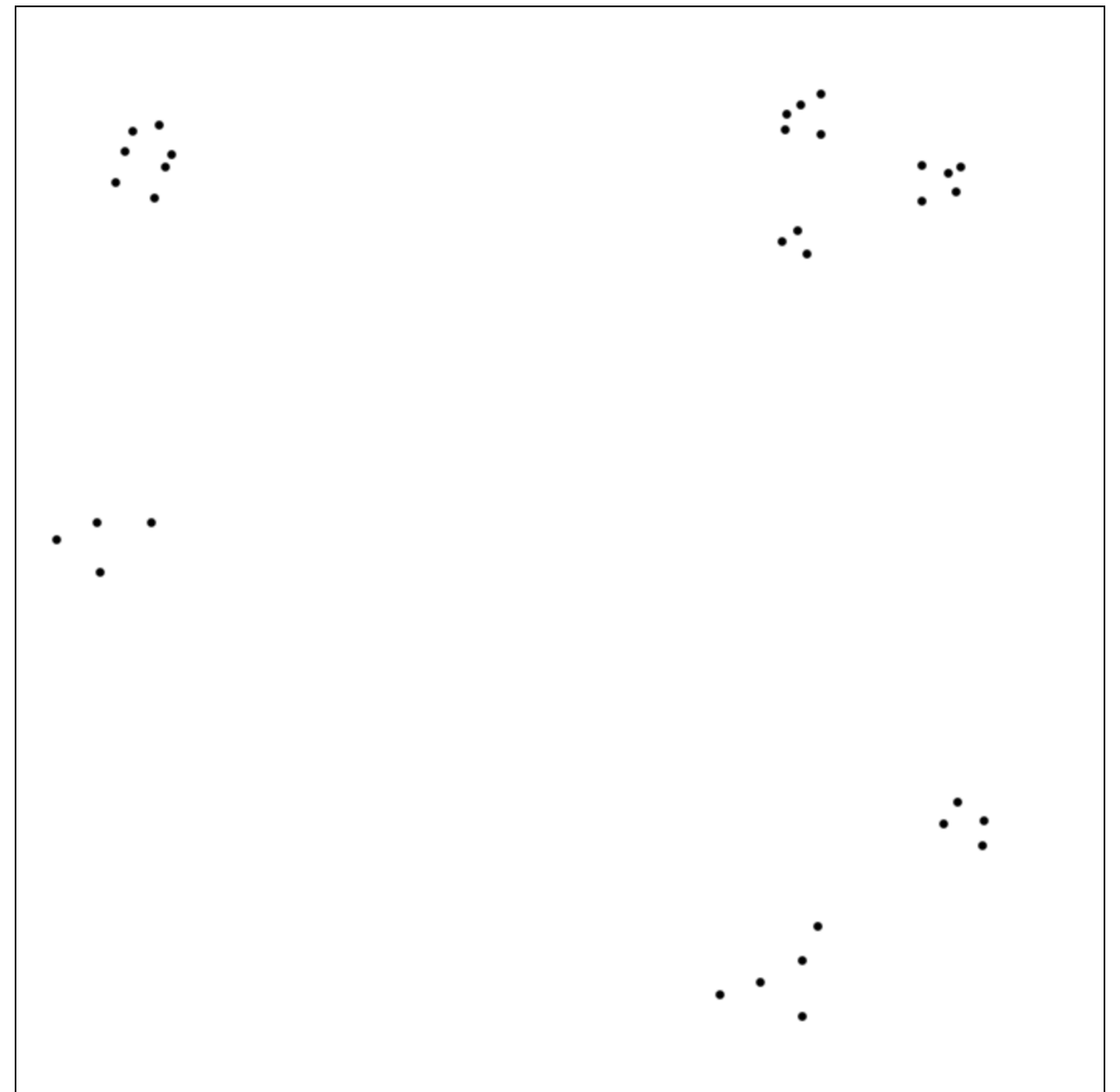
$$P_i(x) = f(x_i) + y_i \cdot (x - x_i)$$

- **Theorem** (Herbert-Voss, H., McCollum 2015): There is an algorithm that computes a value  $M$  such that

$$M \leq \Gamma^1 \leq CM, \quad C \leq 20$$

The run time is  $O(n \log n)$  and the storage is  $O(n)$ .

# Well Separated Pairs Decomposition



# Well Separated Pairs Decomposition

- For Lip, want to compute:

$$\max_{i \in A, j \in B} \frac{|f(x_i) - f(x_j)|}{|x_i - x_j|}$$

- Direct:  $O(|A||B|)$  computations
- Using WSPD:

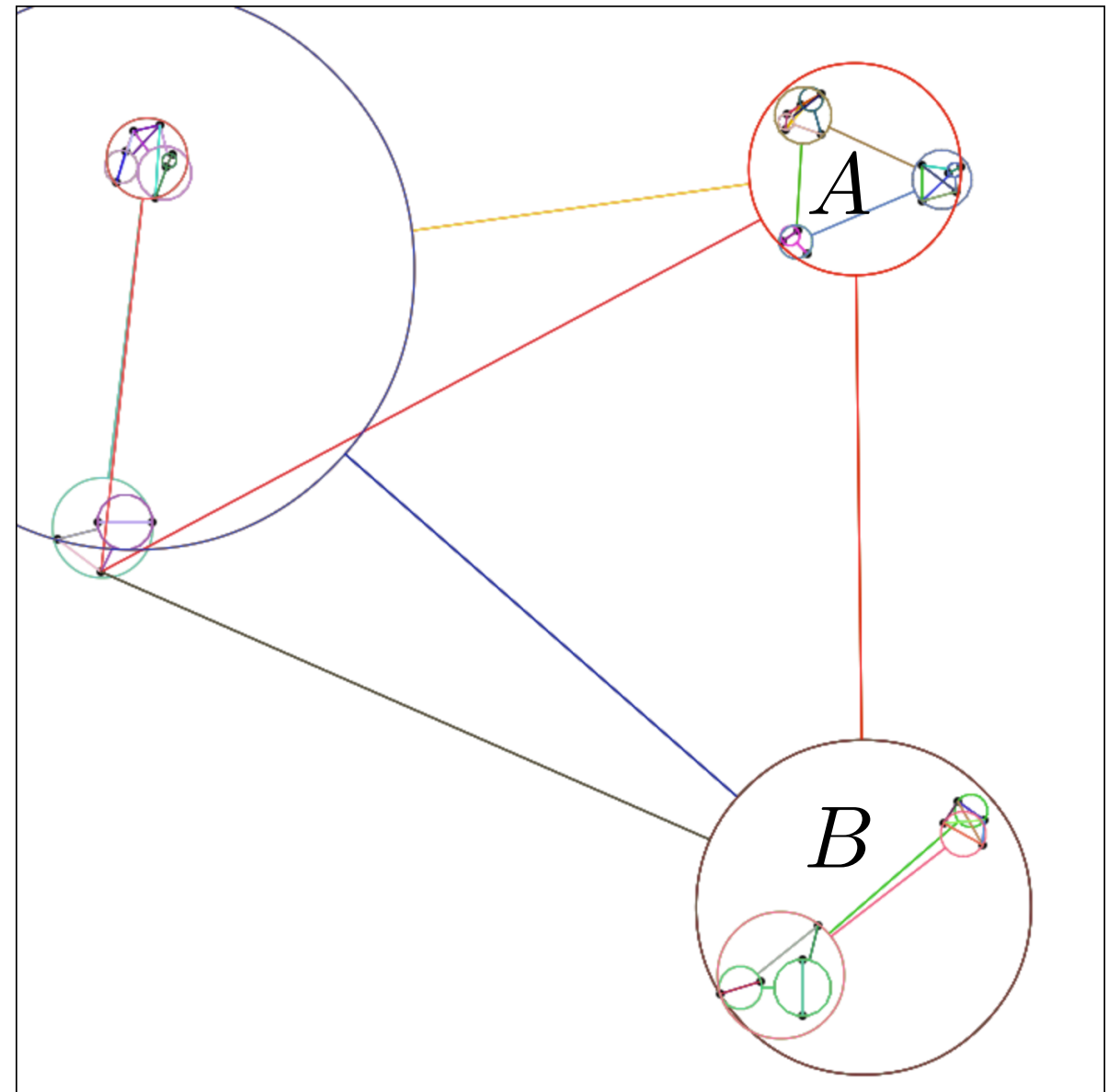
$$|x_i - x_j| \approx C \text{ for } i \in A, j \in B$$

So just need to compute:

$$\max_{i \in A, j \in B} |f(x_i) - f(x_j)|$$

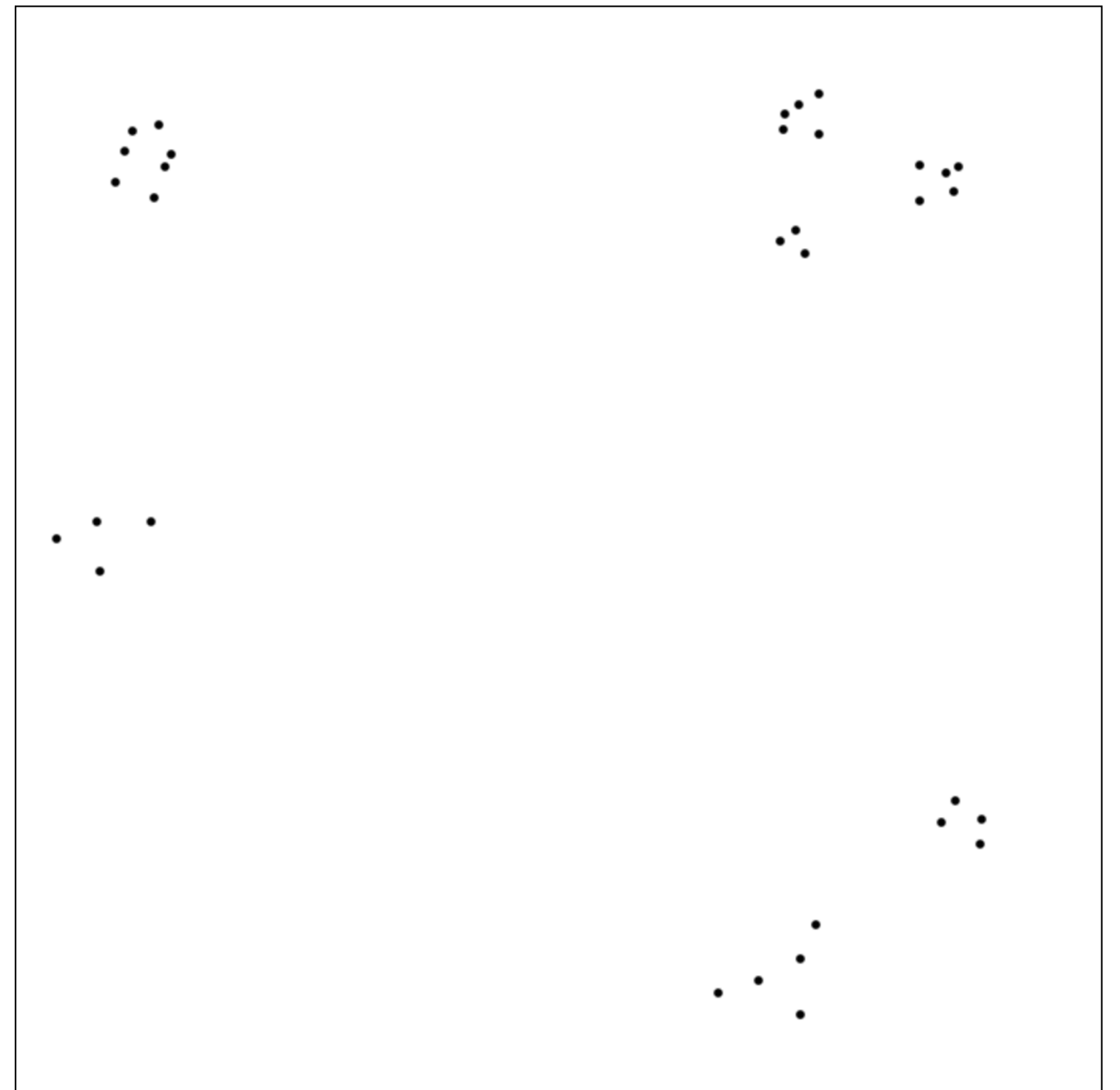
$$\Rightarrow O(|A| + |B|) \text{ computations}$$

- $\Gamma^1$  algorithm is in the same spirit



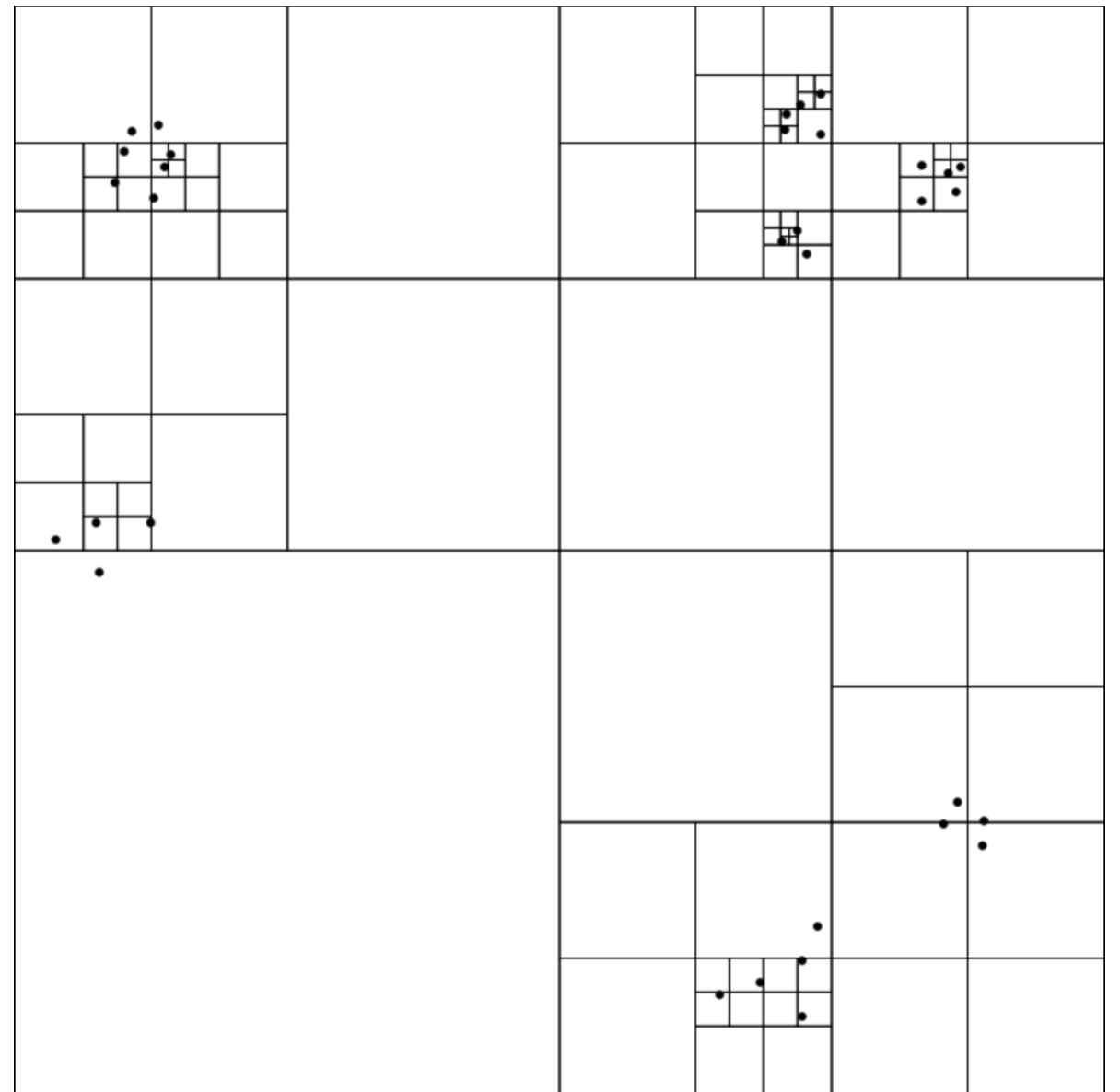
$$\max\{\text{diam}(A), \text{diam}(B)\} < \epsilon \cdot \text{dist}(A, B)$$

# “Standard” Way of Computing Interpolants in $C^m(\mathbb{R}^d)$



# “Standard” Way of Computing Interpolants in $C^m(\mathbb{R}^d)$

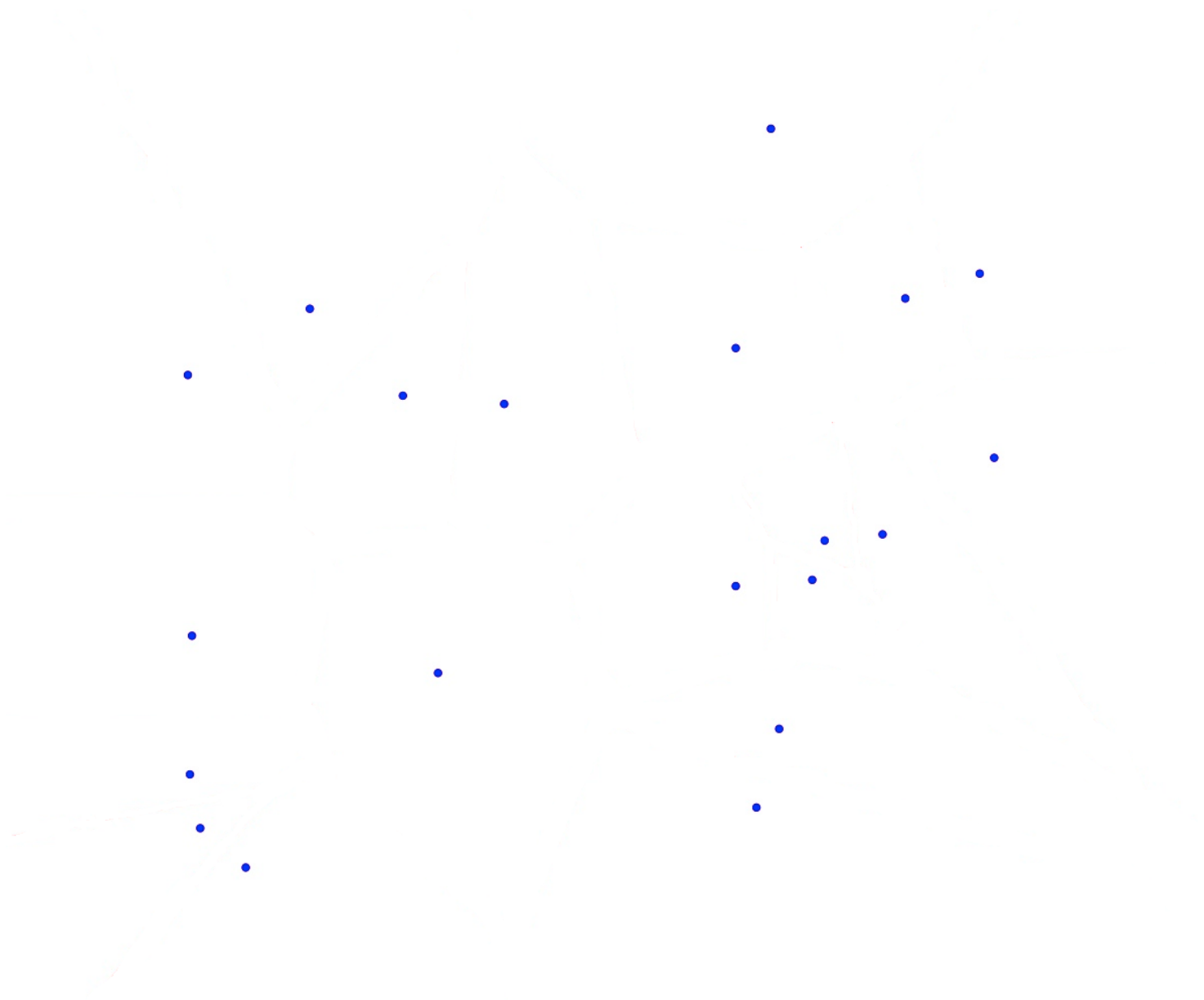
- Euclidean space is partitioned into cubes with a Calderon-Zygmund decomposition adapted to the points  $\{x_i\}_i$
- On each cube, a local interpolant is defined
- The local interpolants are patched together with a partition of unity





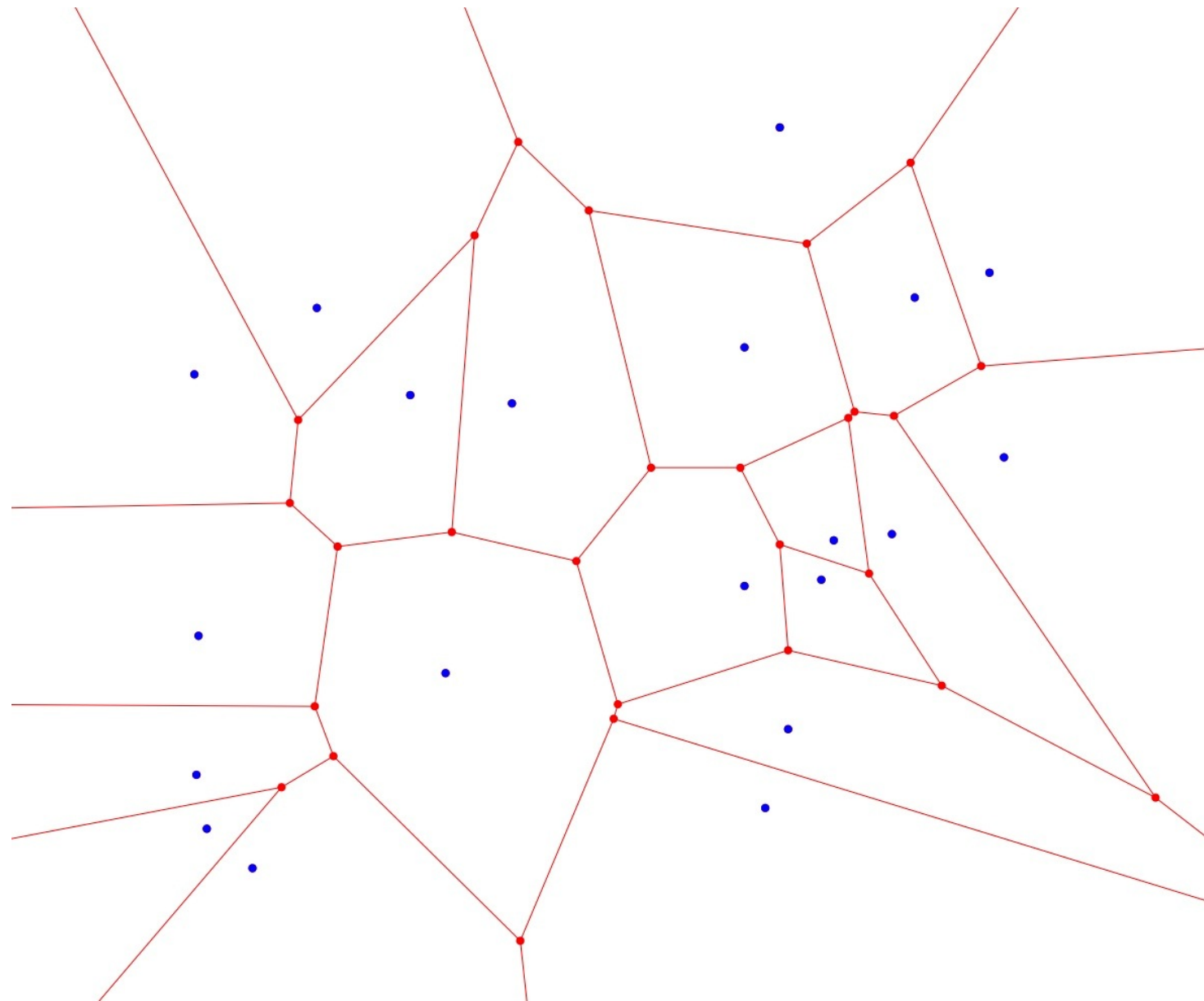
# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- Initial data  $\{x_i, f(x_i), y_i\}_i$



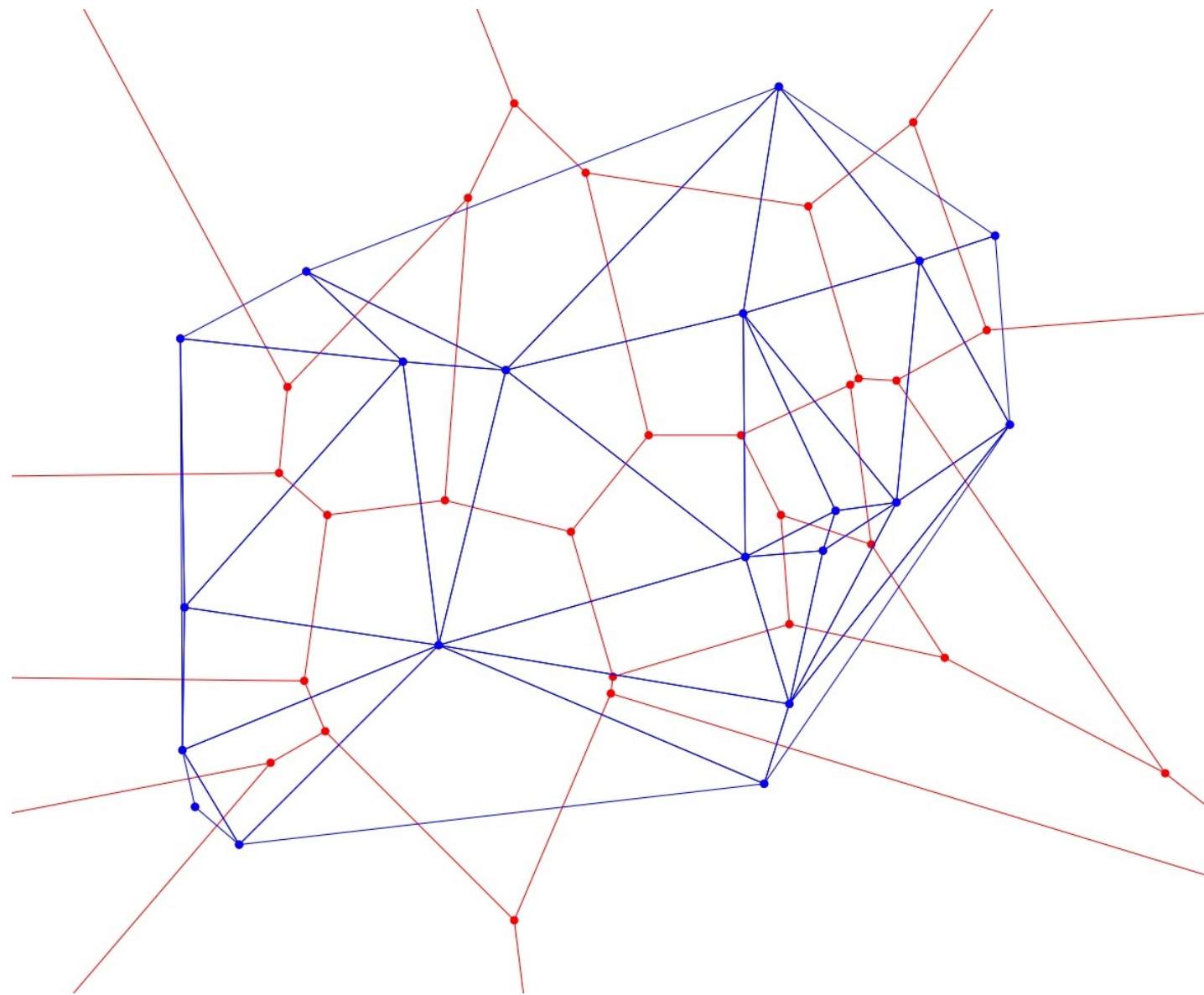
# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- Initial data  $\{x_i, f(x_i), y_i\}_i$
- Weighted Voronoi decomposition based on the initial data



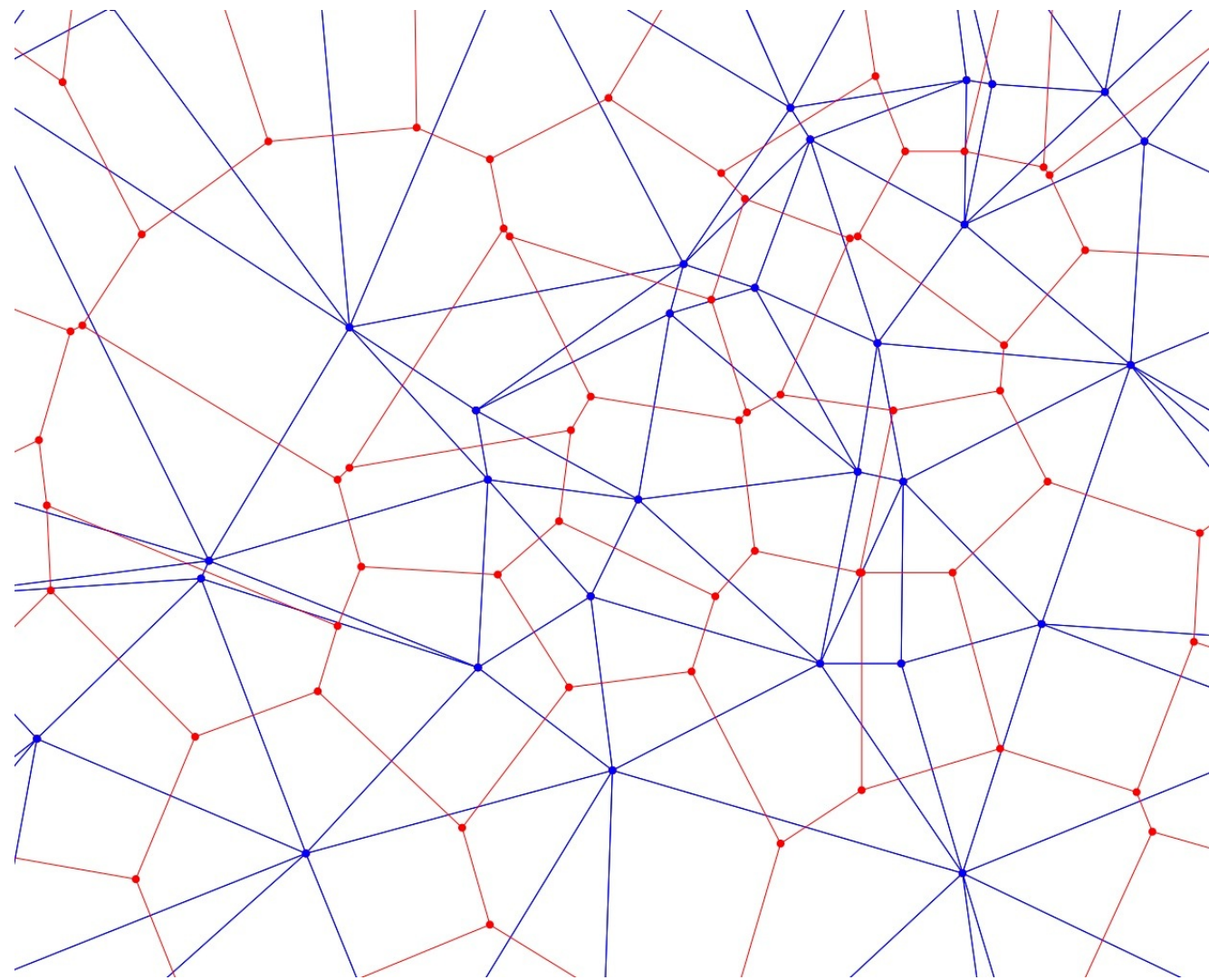
# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- Initial data  $\{x_i, f(x_i), y_i\}_i$
- Weighted Voronoi decomposition based on the initial data
- Dual triangulation



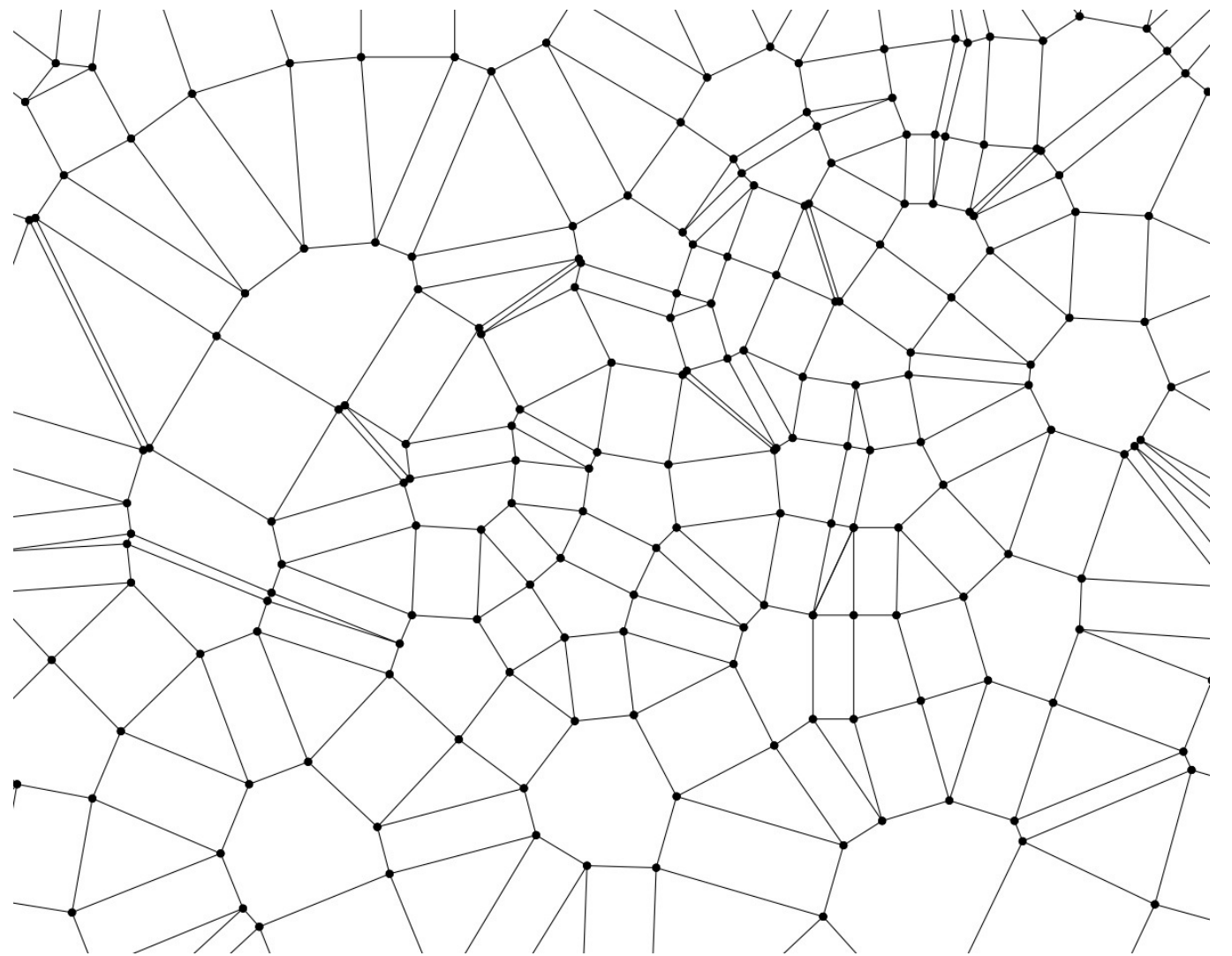
# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- Initial data  $\{x_i, f(x_i), y_i\}_i$
- Weighted Voronoi decomposition based on the initial data
- Dual triangulation



# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- Initial data  $\{x_i, f(x_i), y_i\}_i$
- Weighted Voronoi decomposition based on the initial data
- Dual triangulation
- Merge the Voronoi diagram and the dual triangulation to form the final partition of Euclidean space





# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

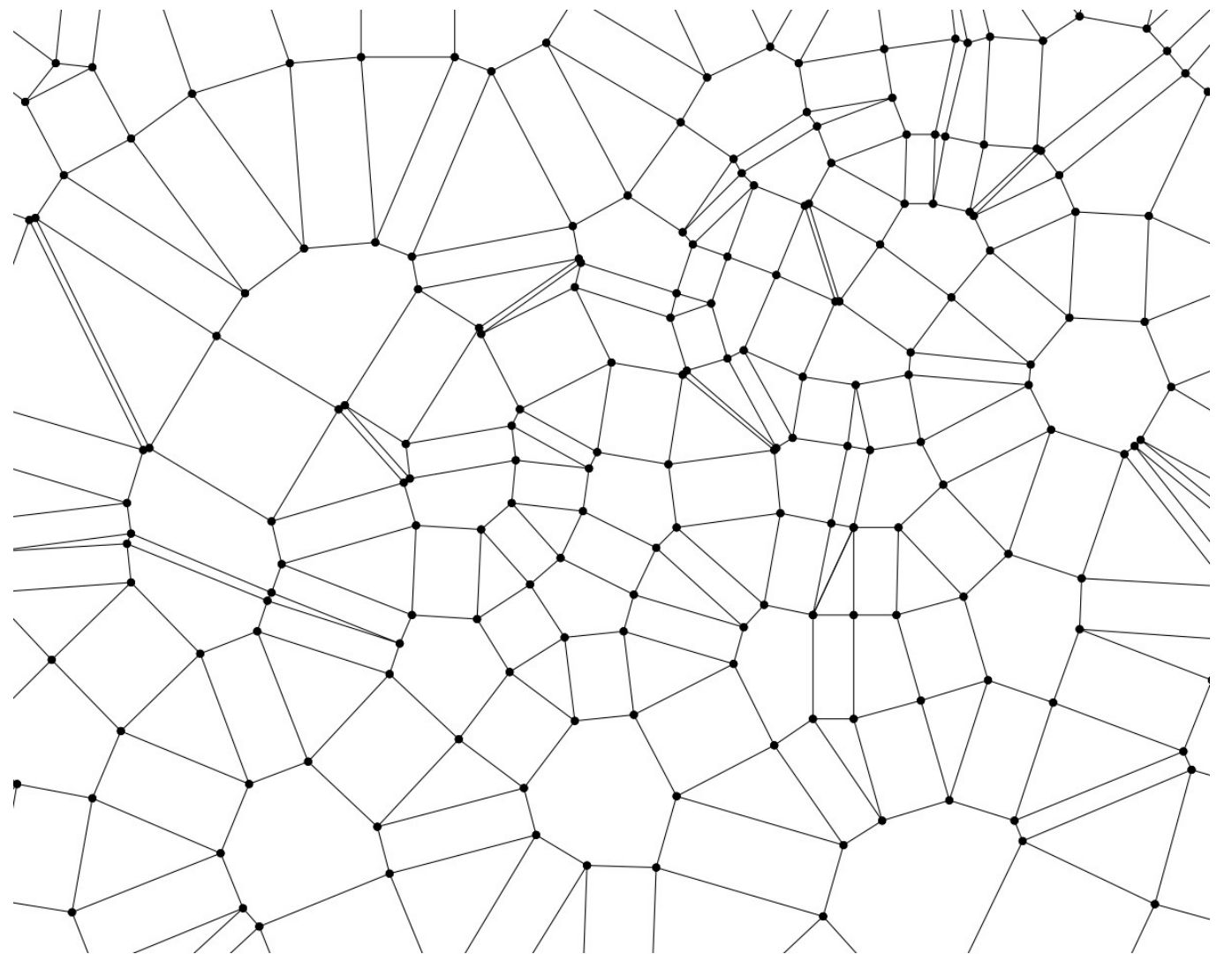
- A local interpolant is defined on each cell
- The local interpolants fit together *perfectly* to form a global interpolant

- Time to compute the cellular decomposition:

$$O(n^{d/2} + n \log n)$$

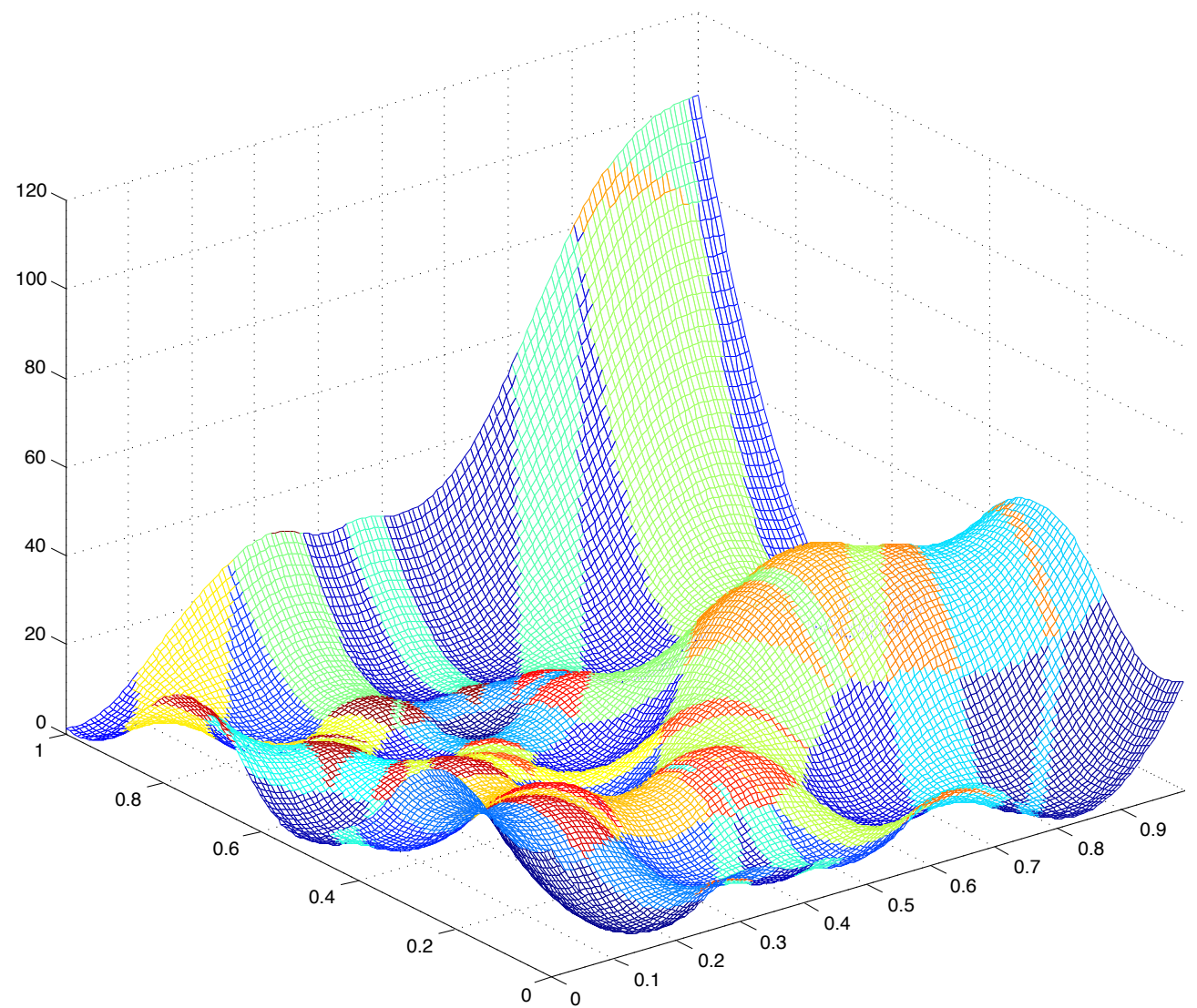
- Time to evaluate global interpolant at a new point:

$$O(\log n)$$



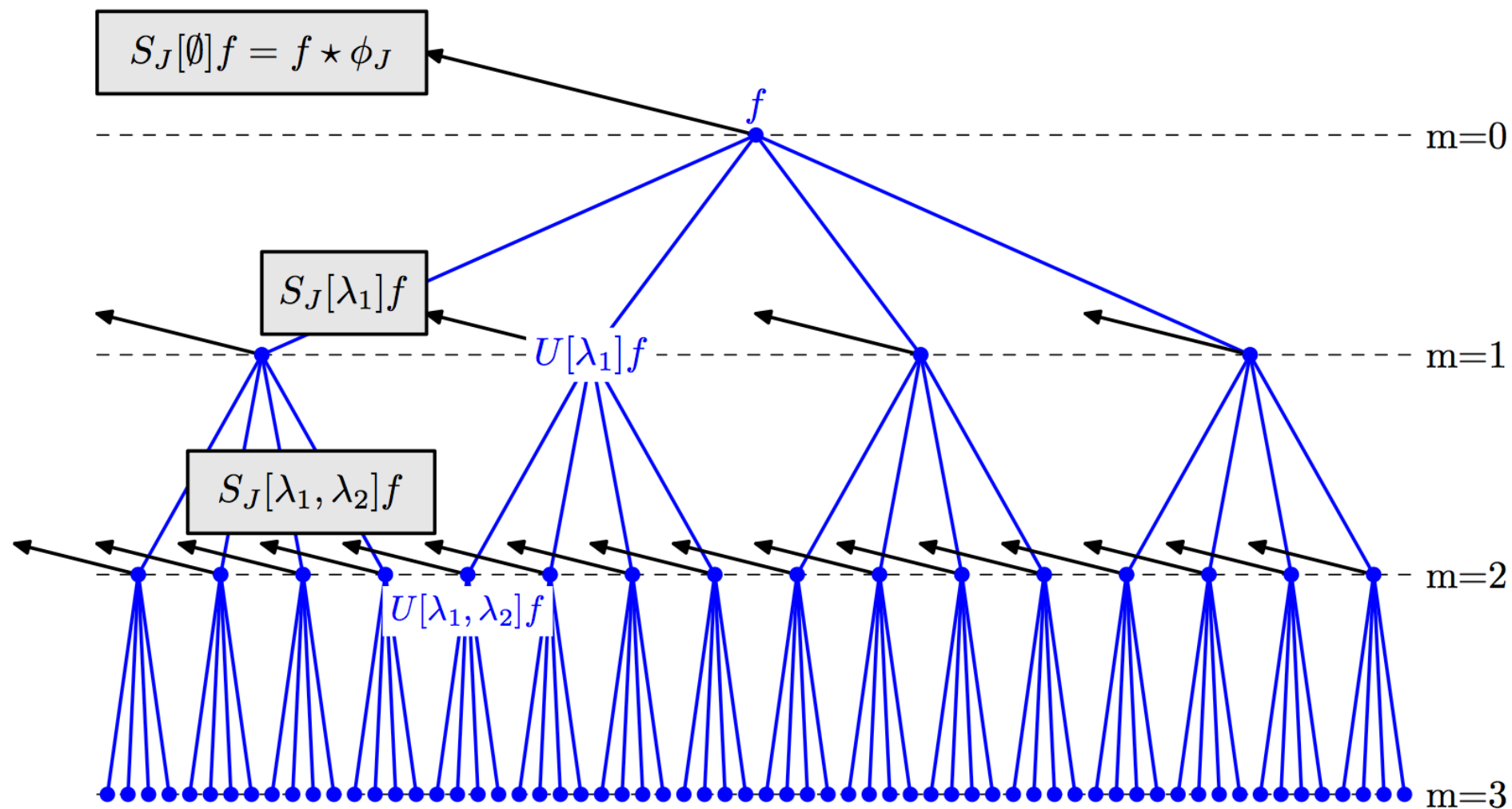
# Computing Interpolants in $C^{1,1}(\mathbb{R}^d)$

- First practical Whitney extension algorithm
- Works for large numbers of training data in small dimensions (two or three)
- Also works for small numbers of training data in dimensions up to ten



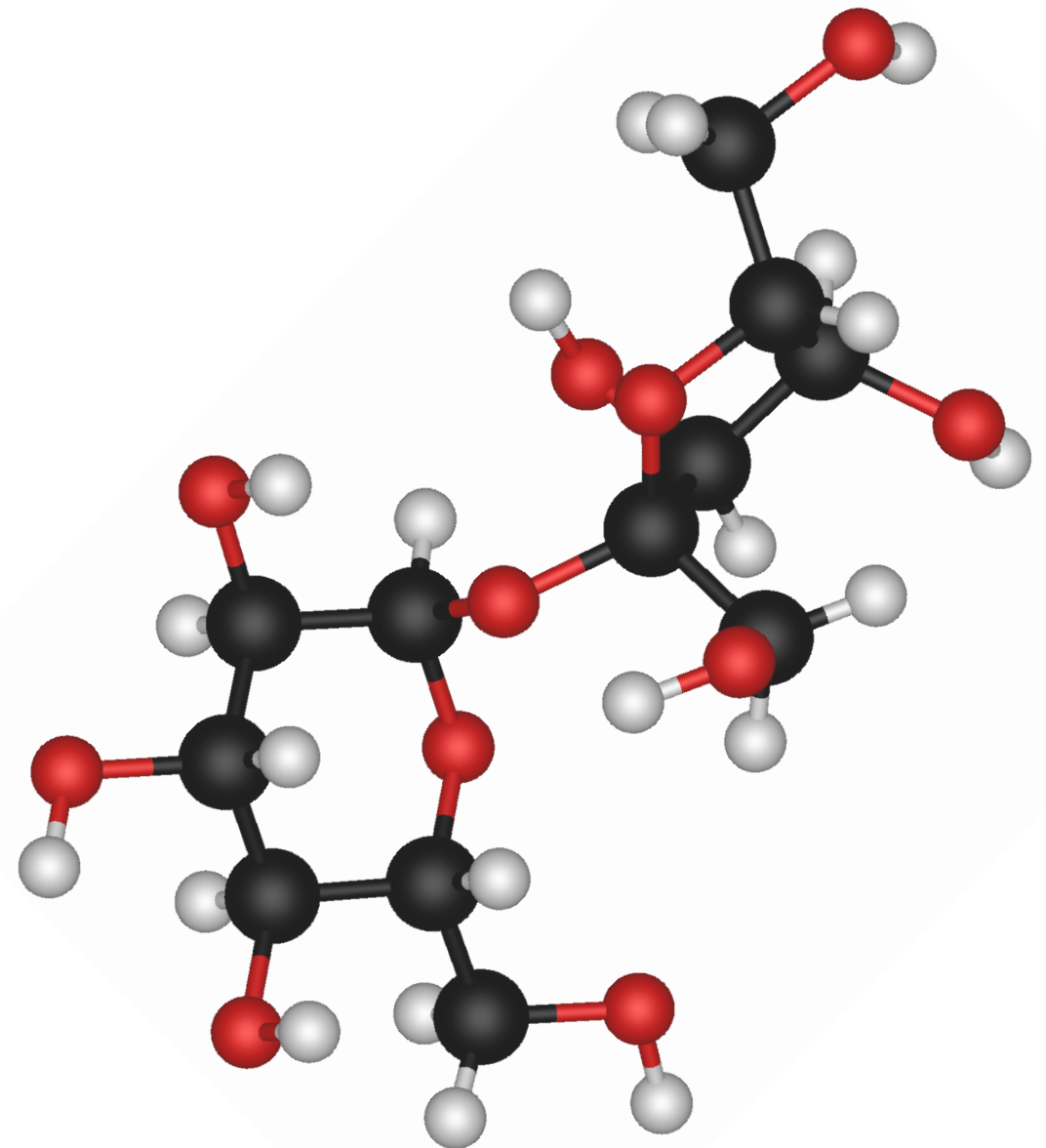
# Sparse Regression

## (Deep Learning and Quantum Chemistry)



# Quantum Chemistry Motivation

- Chemists want to build “Google of molecules”
- Applications in pharmaceutical industry, materials science, among others
- Need to compute potential energy of each molecule
- Billions of molecules
- Complex, time consuming computation



# Energy Computation

- Exact:

Schrödinger's Equation:  $\hat{H}\Psi = E\Psi$

Extremely high dimensional eigenvalue problem

Example: Alcohol  $\text{C}_2\text{H}_6\text{O}$  is  $\sim 2^{300}$  dimensional!

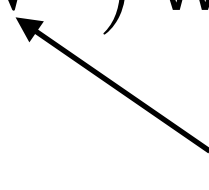
- Approximate:

ab-initio methods:

- Coupled cluster methods

- Density functional theory

Scales as  $O(N^\alpha)$  where  $4 \leq \alpha \leq 7$

 Number of electrons

# Regression

- High dimensional  $x \in \mathbb{R}^d$
- Approximate a functional  $f(x)$  given  $n$  sample values  $\{x_i, f(x_i)\}_i$

- Many body problems:

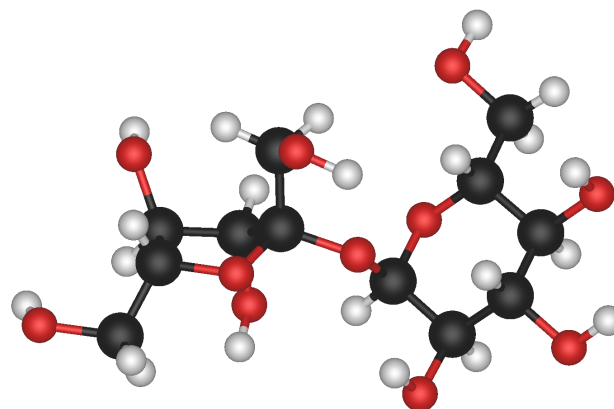
Energy  $f(x)$  of a state  $x = \{(p_k, q_k)\}_k$

Position

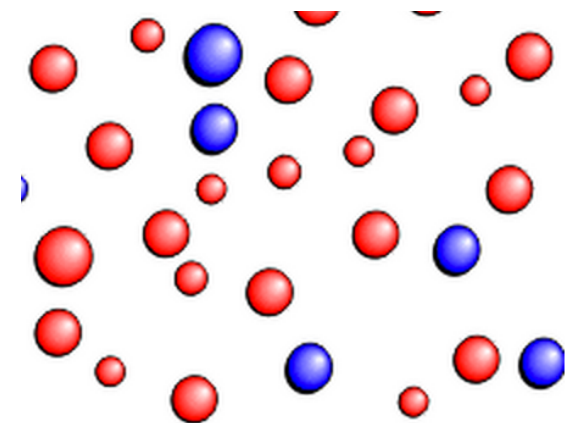
Celestial mechanics: mass of body  
Classical electrostatics: charge of particle  
Quantum chemistry: total protonic charge of atom



Astronomy



Quantum  
Chemistry



Classical  
Electrostatics

# Sparse Linear Regression

- Representation of  $x$  :  $\Phi(x) = \{\phi_j(x)\}_j$
- Regression  $\tilde{f}(x)$  of  $f(x)$  linear in  $\Phi(x)$ :
$$\tilde{f}(x) = \sum_j \alpha_j \phi_j(x)$$
- Few samples  $\{x_i, f(x_i)\}_i$  so want a low dimensional approximation of  $f$  to avoid curse of dimensionality
- Find regression functions  $\{\phi_j\}_j$  with similar properties as  $f$  to allow us to compute a sparse regression



# Energy Properties

- State:  $x = \{(p_k, q_k)\}_k$  positions of atoms and number of protons
- Energy:  $f(x)$

## 1. **Permutation Invariance:**

Invariant to permutations of the indexation of the atoms in each molecule

## 2. **Isometry Invariance:**

Invariant to actions of the isometry group  $E(3) = \mathbb{R}^3 \rtimes O(3)$  on the molecular state

## 3. **Deformation Stability:**

The energy is differentiable with respect to the distances between atoms

## 4. **Multiscale Interactions:**

- Highly energetic covalent bonds between neighboring atoms
- Weaker energetic exchanges at larger distances (Van der Waals interactions)

- Want a representation  $\Phi(x)$  that satisfies these properties



# Current State of the Art

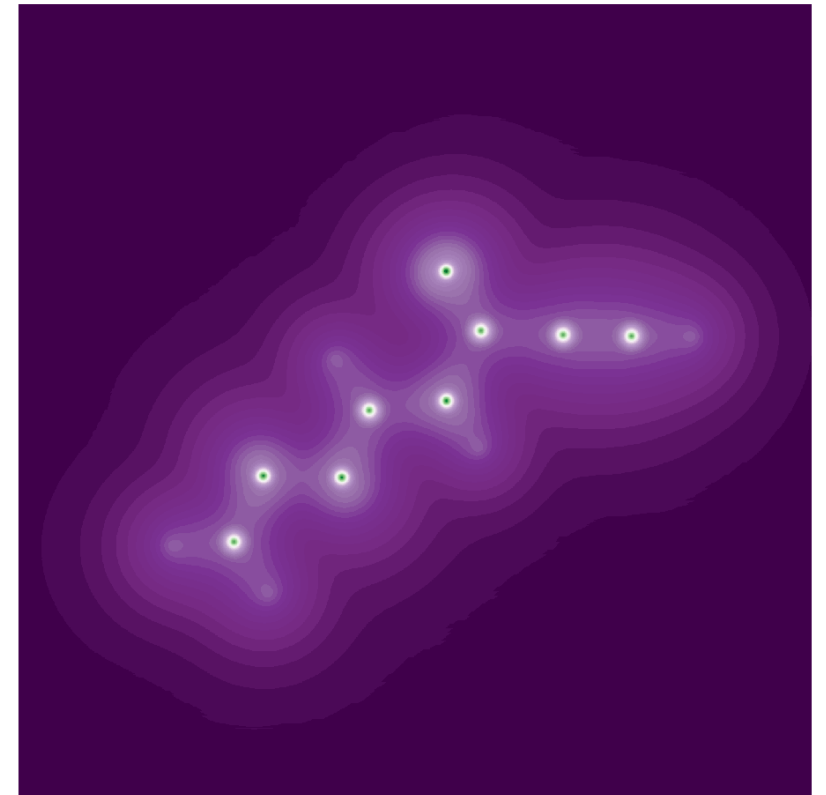
- The set of pairwise distances between atoms defines a set of isometry invariant descriptors that is stable to deformations
- Coulomb matrices [Rupp, et al 2012-2014] refine this idea:

$$C_{k,l}(x) = \begin{cases} \frac{1}{2} q_k^{2.4} & k = l \\ \frac{q_k q_l}{|p_k - p_l|} & k \neq l \end{cases}$$

- Issues:
  - Not permutation invariant (sorted random matrices)
  - Different size matrices (zero padding)
  - All length scales are treated equally (nonlinear kernel)

# Density Functional Theory

- State:  $x = \{(p_k, q_k)\}_k$
- Energy:  $f(x)$
- Electronic density:  $x \mapsto \rho_x(u)$
- Hohenberg and Kohn 1964:



$$\rho_x = \arg \min_{\rho} E(\rho) \text{ and } f(x) = E(\rho_x)$$

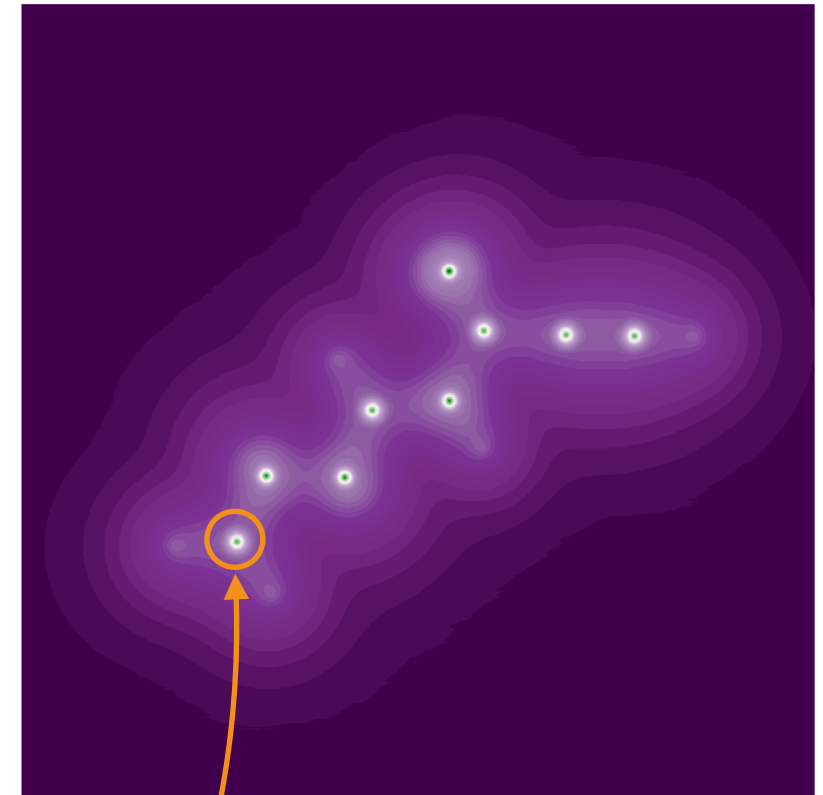
$$E(\rho) = \underbrace{T(\rho)}_{\text{Kinetic energy}} + \underbrace{\int_{\mathbb{R}^3} \rho(u) V_e(u) du}_{\text{External energy (electron-nuclei attraction)}} + \underbrace{\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(u)\rho(v)}{|u-v|} du dv}_{\text{Coulomb energy (electron-electron repulsion)}} + \underbrace{E_{xc}(\rho)}_{\text{Exchange correlation energy}}$$

# Density Functional Theoretic Learning rather than Computing

- Construct a representation  $\Phi(\rho) = \{\phi_j(\rho)\}_j$  and compute a linear regression:

$$\tilde{f}(x) = \tilde{E}(\tilde{\rho}_x) = \sum_j \alpha_j \phi_j(\tilde{\rho}_x)$$

- To avoid computing  $\rho_x$ , we take  $\tilde{\rho}_x$  to be an approximation of  $\rho_x$
- Local behavior near the nucleus of an atom is the same as the isolated electronic density of that atom



$\rho_x(u)$

# Density Functional Theoretic Learning rather than Computing

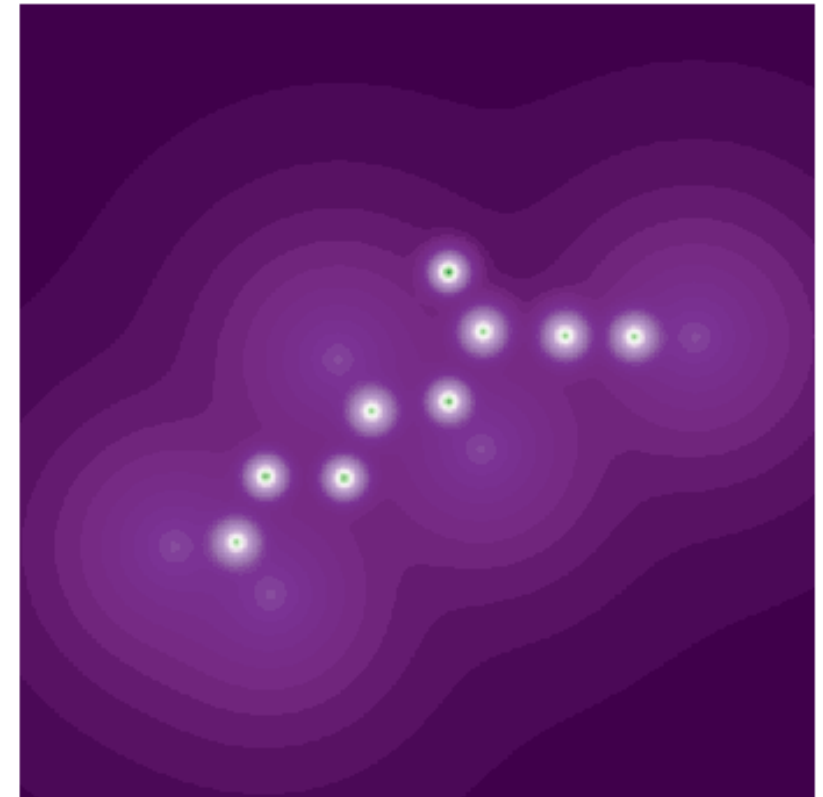
- Construct a representation  $\Phi(\rho) = \{\phi_j(\rho)\}_j$  and compute a linear regression:

$$\tilde{f}(x) = \tilde{E}(\tilde{\rho}_x) = \sum_j \alpha_j \phi_j(\tilde{\rho}_x)$$

- To avoid computing  $\rho_x$ , we take  $\tilde{\rho}_x$  to be an approximation of  $\rho_x$
- Crude electronic density approximation:

$$\tilde{\rho}_x(u) = \sum_k \rho_{a(k)}(u - p_k)$$

$\rho_a$  = density of atom  $a$  centered at zero



$\tilde{\rho}_x(u)$

# Stability to Deformations

- Deformation operator:  $\rho_x = D\rho$
- Then  $f(x) = E(\rho_x) = ED(\rho)$
- Want to linearly expand  $ED(\rho)$  in  $\Phi(\rho)$
- Diffeomorphism model:  $\rho_x(u) = D_\tau \rho(u) = \rho(u - \tau(u))$
- Want  $\Phi$  to be Lipschitz continuous to diffeomorphisms:

$$\|\Phi(\rho) - \Phi(D_\tau \rho)\| \leq C \cdot \sup_{u \in \mathbb{R}^3} \|\nabla \tau(u)\| \cdot \|\rho\|_2$$

- If  $E(\rho)$  is well approximated by a linear regression in  $\Phi(\rho)$ , and  $\Phi(\rho)$  is Lipschitz continuous over  $D_\tau$ , then we can still linearly expand  $ED_\tau(\rho)$  in  $\Phi(\rho)$  with small error

# Coulomb Potential Energy

- Coulomb Potential Energy:

$$U(\rho) = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \rho(u) \rho(v) V(u - v) \, du \, dv, \quad V(u) = |u|^{-1}$$

- Convolutional formula for Coulomb energy:

$$U(\rho) = \frac{1}{2} \int_{\mathbb{R}^3} \rho * \bar{\rho}(u) V(u) \, du, \quad \bar{\rho}(u) = \rho(-u)$$

- Fourier transform:

$$\hat{\rho}(\omega) = \int_{\mathbb{R}^3} \rho(u) e^{-iu \cdot \omega} \, du$$

- Coulomb energy in Fourier:

$$U(\rho) = \frac{1}{2(2\pi)^3} \int_{\mathbb{R}^3} |\hat{\rho}(\omega)|^2 \hat{V}(\omega) \, d\omega$$

# Fourier Regression of Coulomb Potential Energy

- Coulomb energy in Fourier:

$$U(\rho) = \frac{1}{2(2\pi)^3} \int_{\mathbb{R}^3} |\hat{\rho}(\omega)|^2 \hat{V}(\omega) d\omega$$

- Isometry invariant Fourier:

In polar coordinates  $\omega = \gamma\eta$  with  $\gamma = |\omega|$  and  $\eta \in S^2$ ,  $\hat{V}(\omega) = \hat{V}(\gamma)$  so

$$U(\rho) = \frac{1}{2(2\pi)^3} \int_{\mathbb{R}} \hat{V}(\gamma) \phi_{\gamma}^2(\rho) d\gamma, \quad \phi_{\gamma}^2(\rho) = \int_{|\omega|=\gamma} |\hat{\rho}(\omega)|^2 d\omega$$

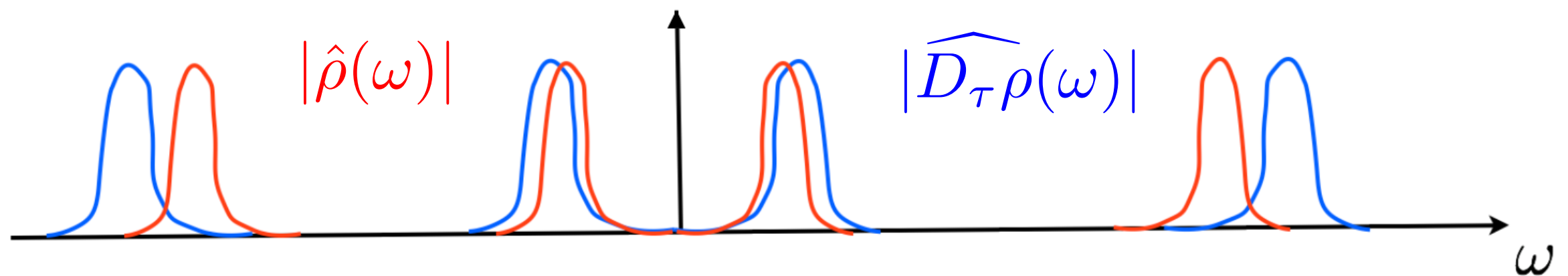
- To learn discrete weights, approximate with Riemann sum:

$$\tilde{U}(\rho) = \frac{\Delta}{2(2\pi)^3} \sum_{m=1}^M \hat{V}(m\Delta) \phi_{m\Delta}^2(\rho)$$

- To get  $|U(\rho) - \tilde{U}(\rho)| < \epsilon$  need  $\Delta \sim \epsilon$  and so  $M = O(\epsilon^{-1})$

# Fourier Limitations

- The Fourier representation does not take advantage of the regularity of  $\widehat{V}(\omega)$  away from  $\omega = 0$ . Therefore it needs  $O(\epsilon^{-1})$  terms to achieve precision  $\epsilon$ .
- Fourier is not stable to small diffeomorphisms at the high frequencies.



$$\| |\hat{\rho}| - |\widehat{D_\tau \rho}| \|_2 \gg \sup_{u \in \mathbb{R}^3} \|\nabla \tau(u)\| \cdot \|\rho\|_2$$



# Wavelets

- Complex valued Morlet wavelet:

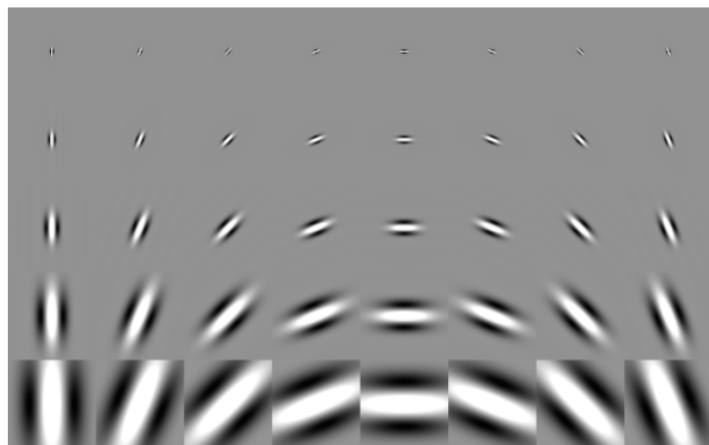
$$\psi(u) = g(u)(e^{i\eta_0 \cdot u} - C), \quad \int_{\mathbb{R}^3} \psi(u) du = 0$$

- Wavelet transform dilates and rotates the wavelet:

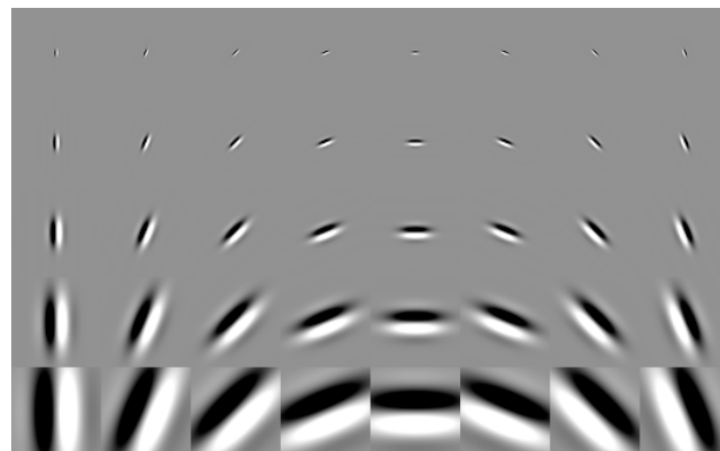
$$\psi_{j,r}(u) = 2^{-3\frac{j}{Q}} \psi(2^{-\frac{j}{Q}} r^{-1} u), \quad (j, r) \in \mathbb{Z} \times \mathbf{O}(3)$$

$Q \in \mathbb{N}$  : Scale oversampling factor

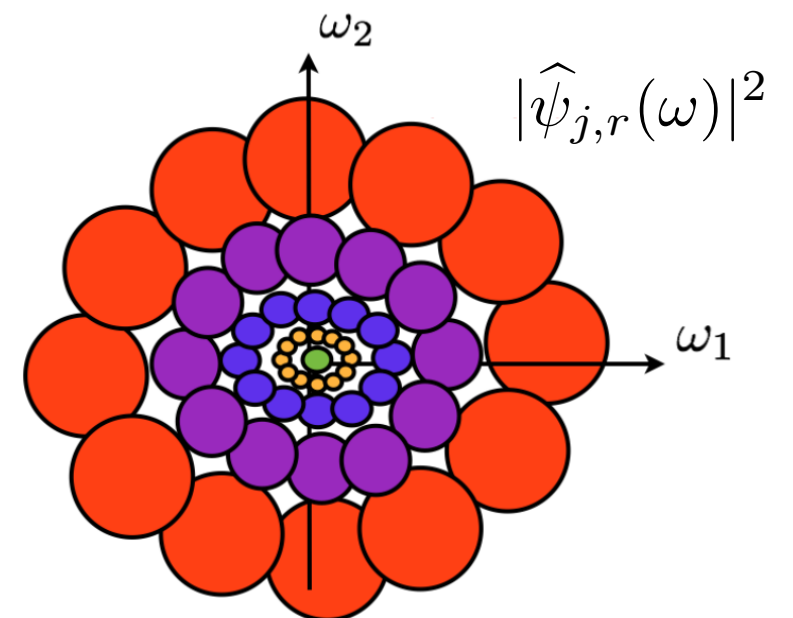
$$W[j, r]\rho(u) = \{\rho * \psi_{j,r}(u)\}_{j \in \mathbb{Z}, r \in \mathbf{O}(3), u \in \mathbb{R}^3}$$



Real parts

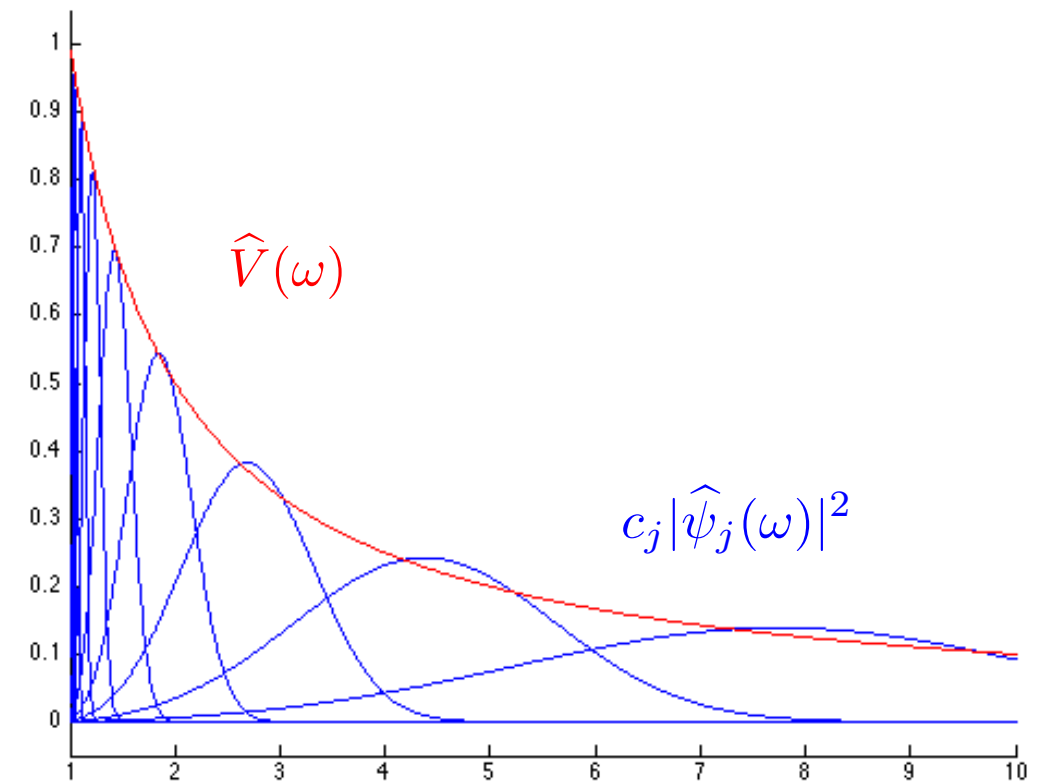


Imaginary parts



# Fourier vs Wavelets

- Wavelets separate scales logarithmically and can thus take advantage of the multiscale structure of the energy. For the Coulomb potential energy, wavelets take advantage of the regularity of  $\hat{V}(\omega)$  away from  $\omega = 0$ .



- Mallat 2012: Wavelets are Lipschitz continuous to the action of diffeomorphisms:

$$\|[W, D_\tau]\| = \|W D_\tau - D_\tau W\| \leq C \cdot \sup_{u \in \mathbb{R}^3} \|\nabla \tau(u)\|$$

# Wavelet Regression of Coulomb Potential Energy

- Regression with wavelet energy functionals:

$$\tilde{U}(\rho) = \sum_{j=j_{\min}}^{j=j_{\max}} \alpha_j \phi_j^2(\rho), \quad \phi_j^2(\rho) = \int_{\mathbb{R}^3} \int_{O(3)} |\rho * \psi_{j,r}(u)|^2 dr du$$

- **Theorem** (H., Mallat, Poilvert 2015): For all  $\epsilon > 0$  there exists a scale oversampling factor  $Q \in \mathbb{N}$  such that

$$|U(\rho) - \tilde{U}(\rho)| < \epsilon \cdot \max(\|\rho\|_1^2, \|\rho\|_2^2)$$

with  $|j_{\min} - j_{\max}| = O(|\log \epsilon|)$ .

# Quantum Wavelet and Fourier Dictionaries

- Full quantum energy is not quadratic. Need linear and quadratic terms.
- Covalent bonds between atoms dominate the energy. These involve two electrons each. Thus the majority of the energy is proportional to the sum of the charges:

$$\phi_0(\rho) = \int_{\mathbb{R}^3} \rho(u) du = \sum_k q_k$$

- We complement Fourier and Wavelet dictionaries by incorporating this linear term with  $\mathbf{L}^1$  and  $\mathbf{L}^2$  terms.

# Quantum Wavelet and Fourier Dictionaries

- Fourier  $\mathbf{L}^p$  terms and dictionary:

$$\phi_{\gamma,p}(\rho) = \left( \int_{|\omega|=\gamma} |\hat{\rho}(\omega)|^p d\omega \right)^{1/p}$$

$$\Phi_F(\rho) = \{\phi_0(\rho), \phi_{m\Delta,1}(\rho), \phi_{m\Delta,1}^2(\rho), \phi_{m\Delta,2}^2(\rho)\}_{m \in \mathbb{N}}$$

- Wavelet  $\mathbf{L}^p$  terms and dictionary:

$$\phi_{j,p}(\rho) = \left( \int_{\mathbb{R}^3} \int_{\mathbf{O}(3)} |\rho * \psi_{j,r}(u)|^p dr du \right)^{1/p}$$

$$\Phi_W(\rho) = \{\phi_0(\rho), \phi_{j,1}(\rho), \phi_{j,1}^2(\rho), \phi_{j,2}^2(\rho)\}_{j \in \mathbb{Z}}$$

# Learning the Weights

- Training set:  $\{(x_i, f(x_i))\}_i \mapsto \{(\tilde{\rho}_{x_i}, E(\rho_{x_i}))\}_i$
- Greedy algorithm to compute M-term sparse regression:

$$\tilde{f}_M(x) = \tilde{E}_M(\tilde{\rho}_x) = \sum_{k=1}^M \alpha_k \phi_{j_k}(\tilde{\rho}_x)$$

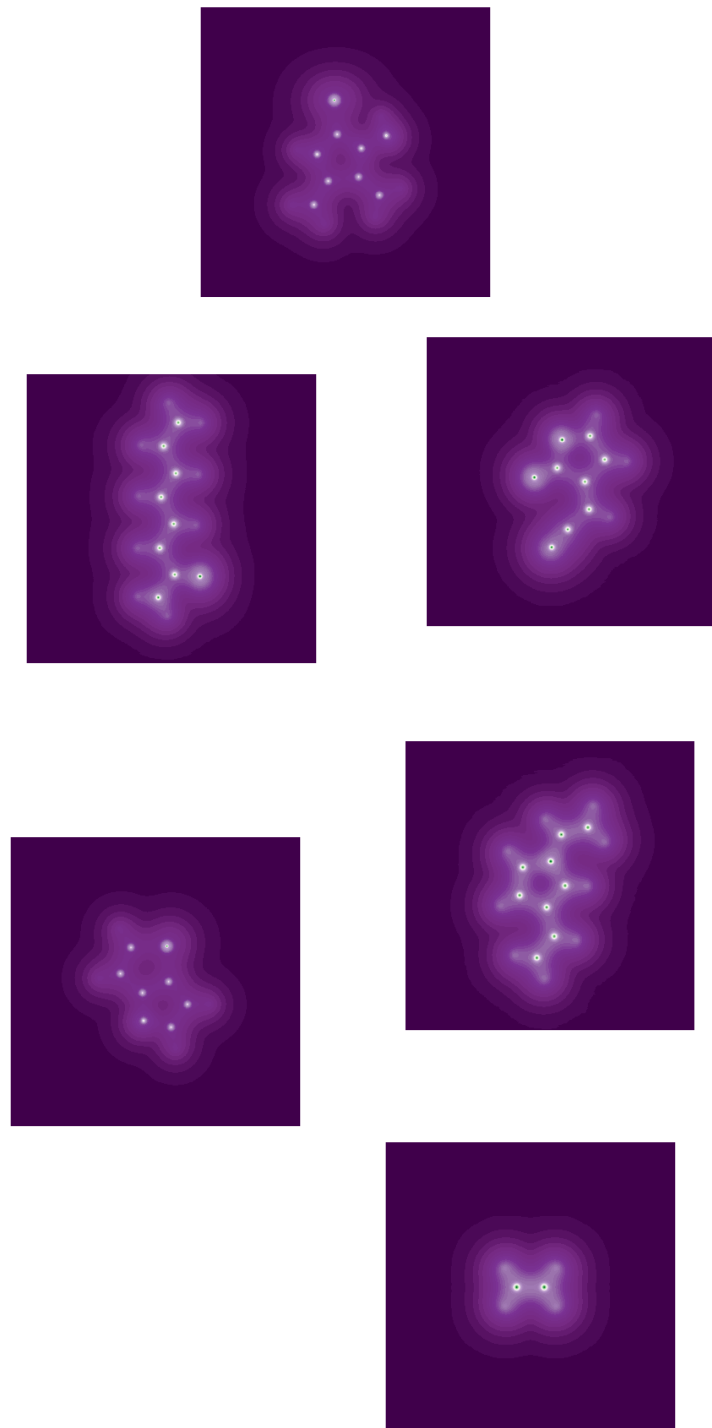
- The algorithm selects the functions  $\{\phi_{j_k}\}_k$  and learns the weights  $\{\alpha_k\}_k$  by minimizing

$$\sum_i |E(\rho_{x_i}) - \tilde{E}_m(\tilde{\rho}_{x_i})|^2$$

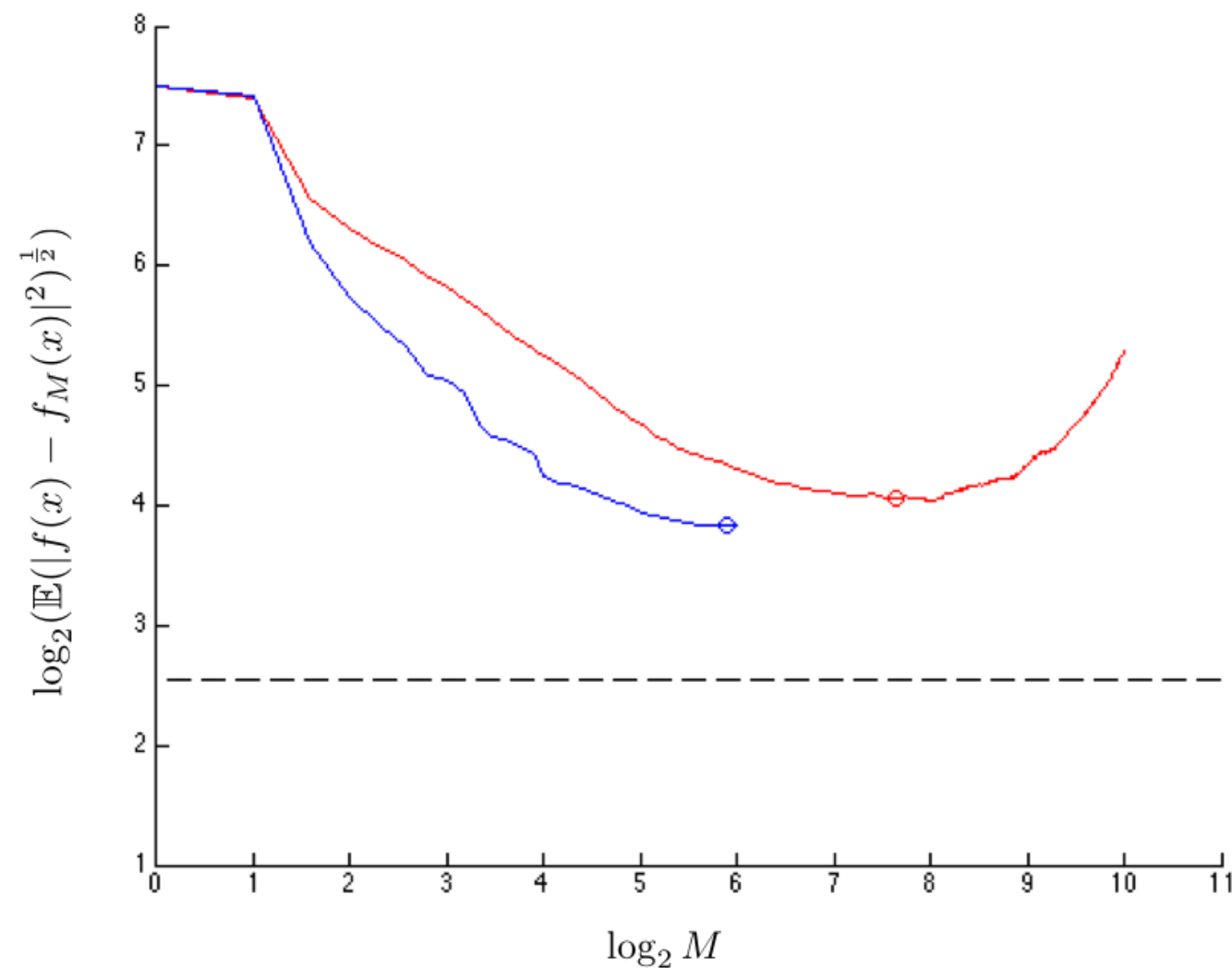
at each iteration  $m = 1, \dots, M$

# Data Set

- Data set  $\{x_i, f(x_i)\}_i$  consisting of over 4000 planar organic molecules made up of hydrogen, carbon, nitrogen, oxygen, sulfur, and chlorine.
- Molecules have between 6 and 20 atoms
- Each molecule  $x_i$  is unique and in its ground state configuration (configuration that minimizes energy)
- $f(x_i)$  is the atomization energy of the molecule (energy necessary to break atomic bonds)



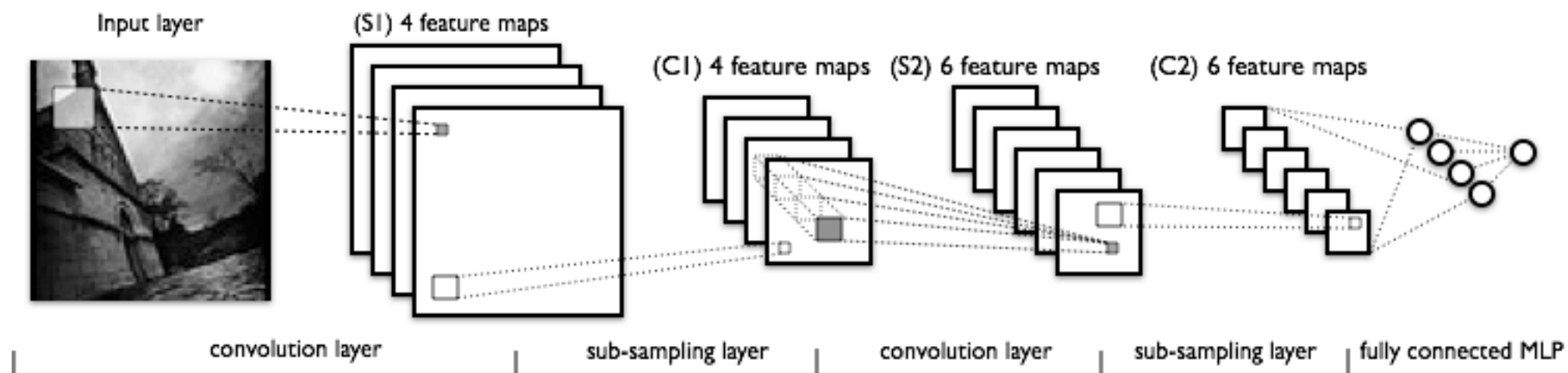
# Fourier and Wavelet M-term Regression Error



Key: Fourier, Wavelets, Coulomb matrices (dashed line)



# Deep Convolutional Networks



- Convolution layer:  $h(u) = \tanh(g * L_k(u) + b_k)$
- Sub-sampling layer (nonlinear): Max pooling
- Linear filters  $L_k$  and weights  $b_k$  are learned from training data via back-propagation

# Scattering Dictionary

Layer 0

$\rho$

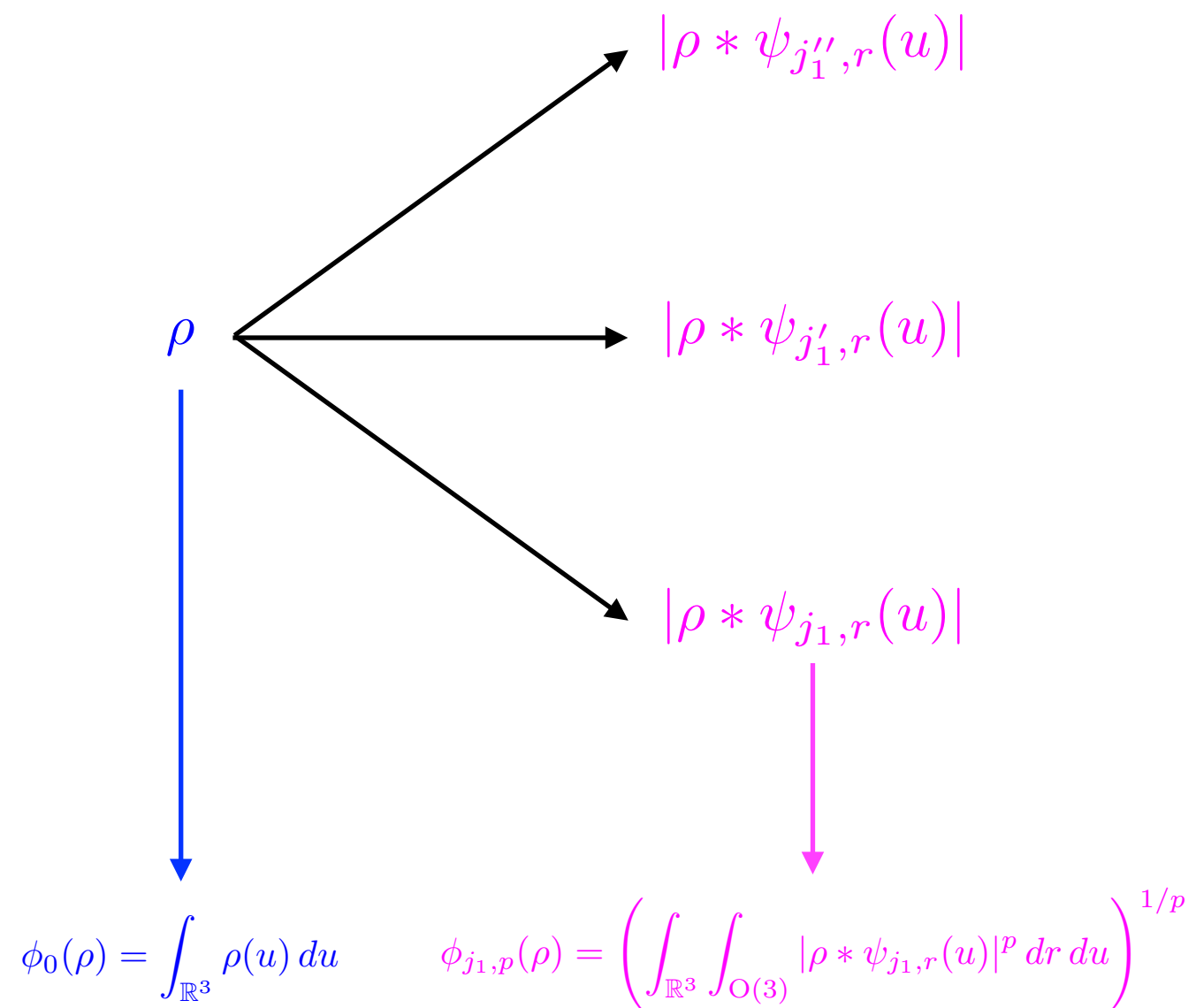


$$\phi_0(\rho) = \int_{\mathbb{R}^3} \rho(u) du$$

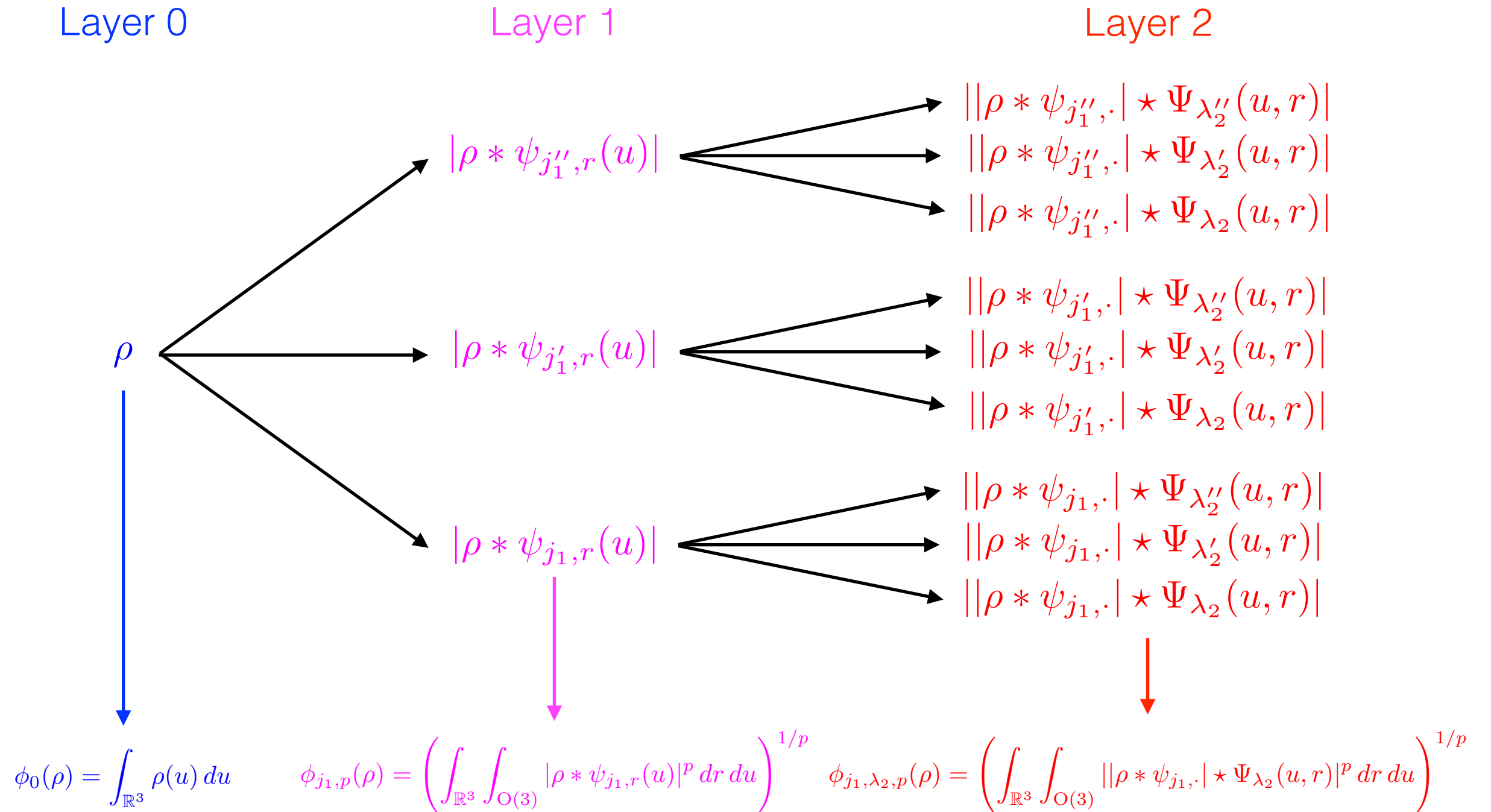
# Scattering Dictionary

Layer 0

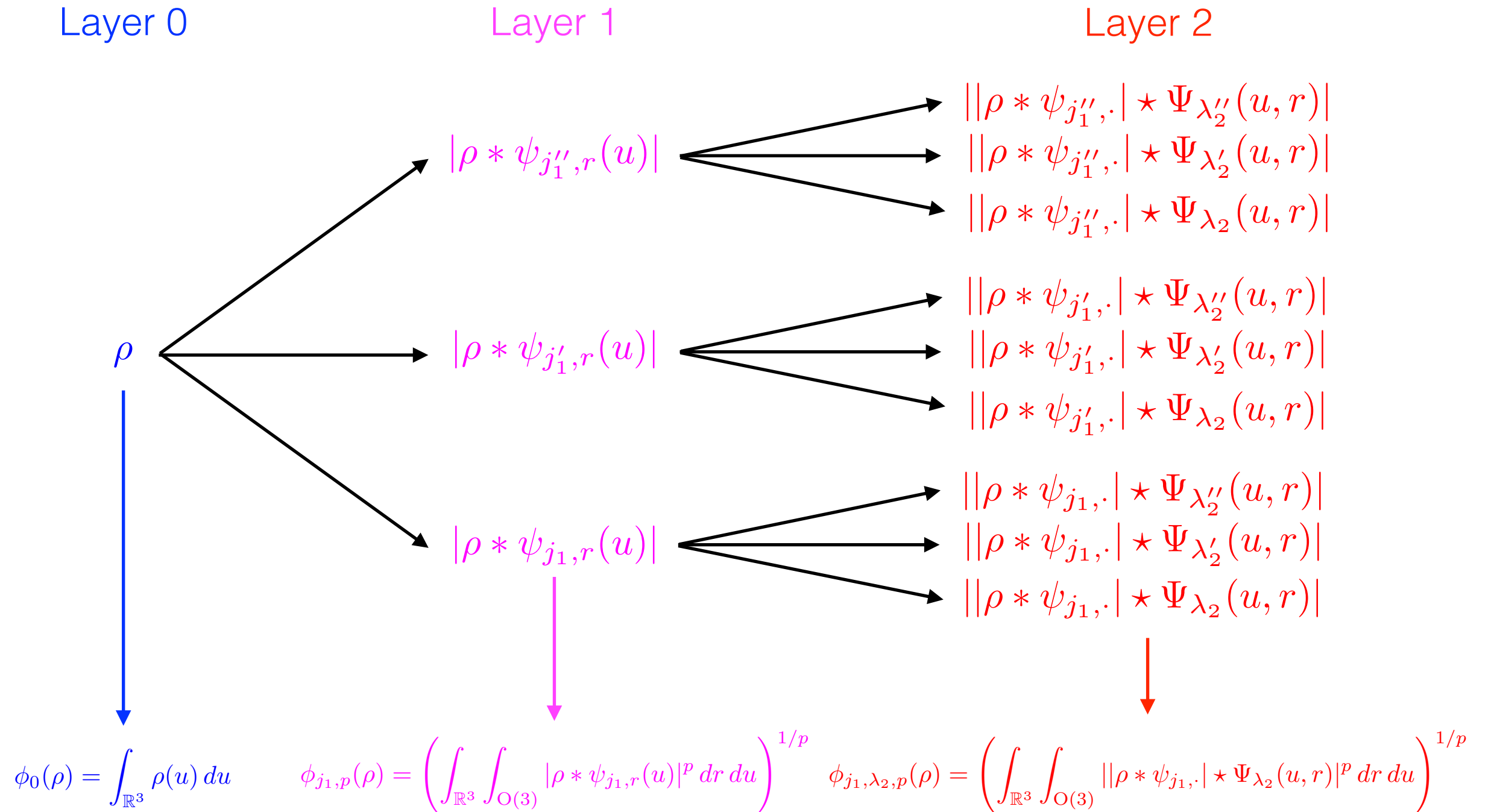
Layer 1



# Scattering Dictionary

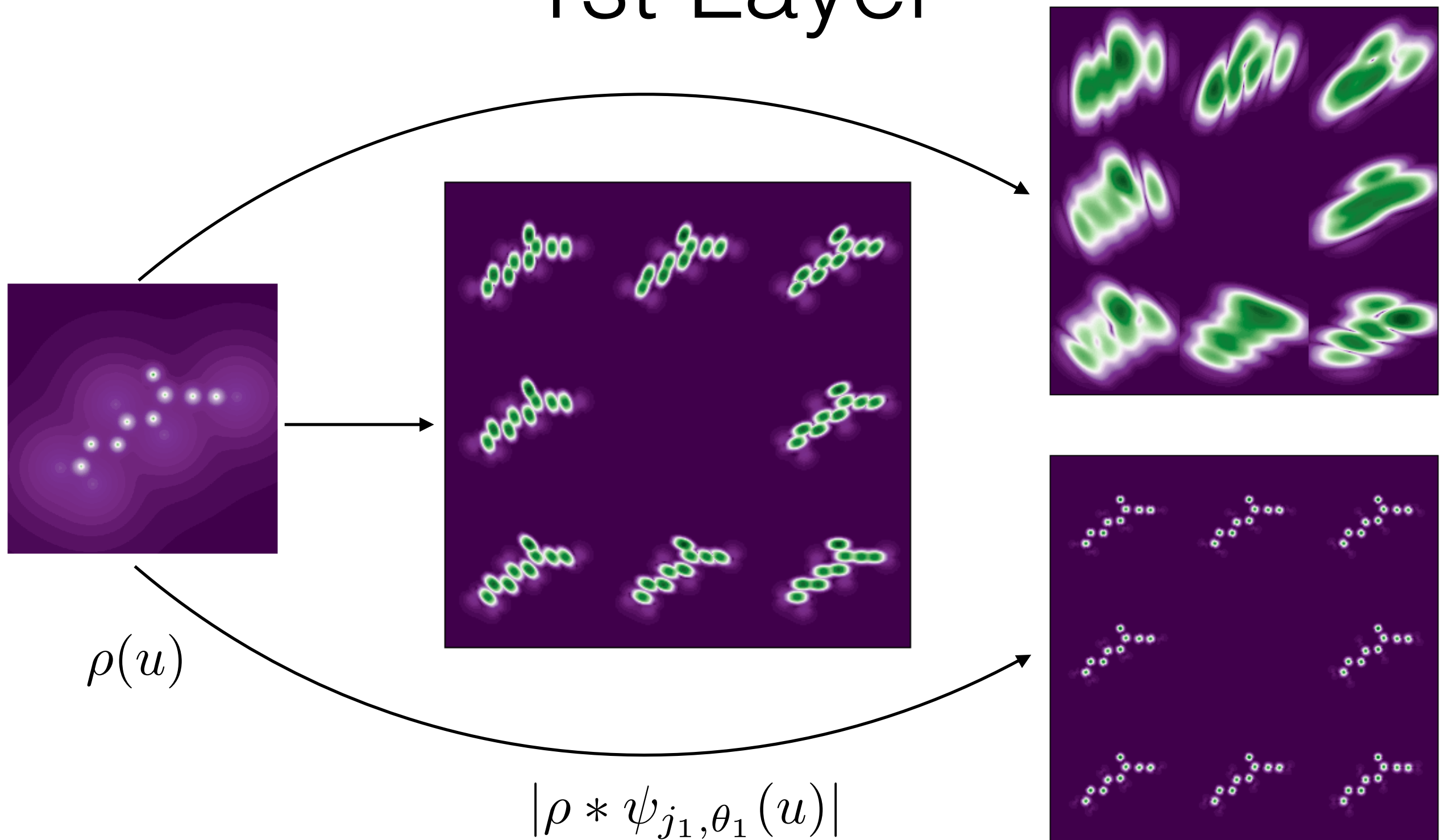


# Scattering Dictionary

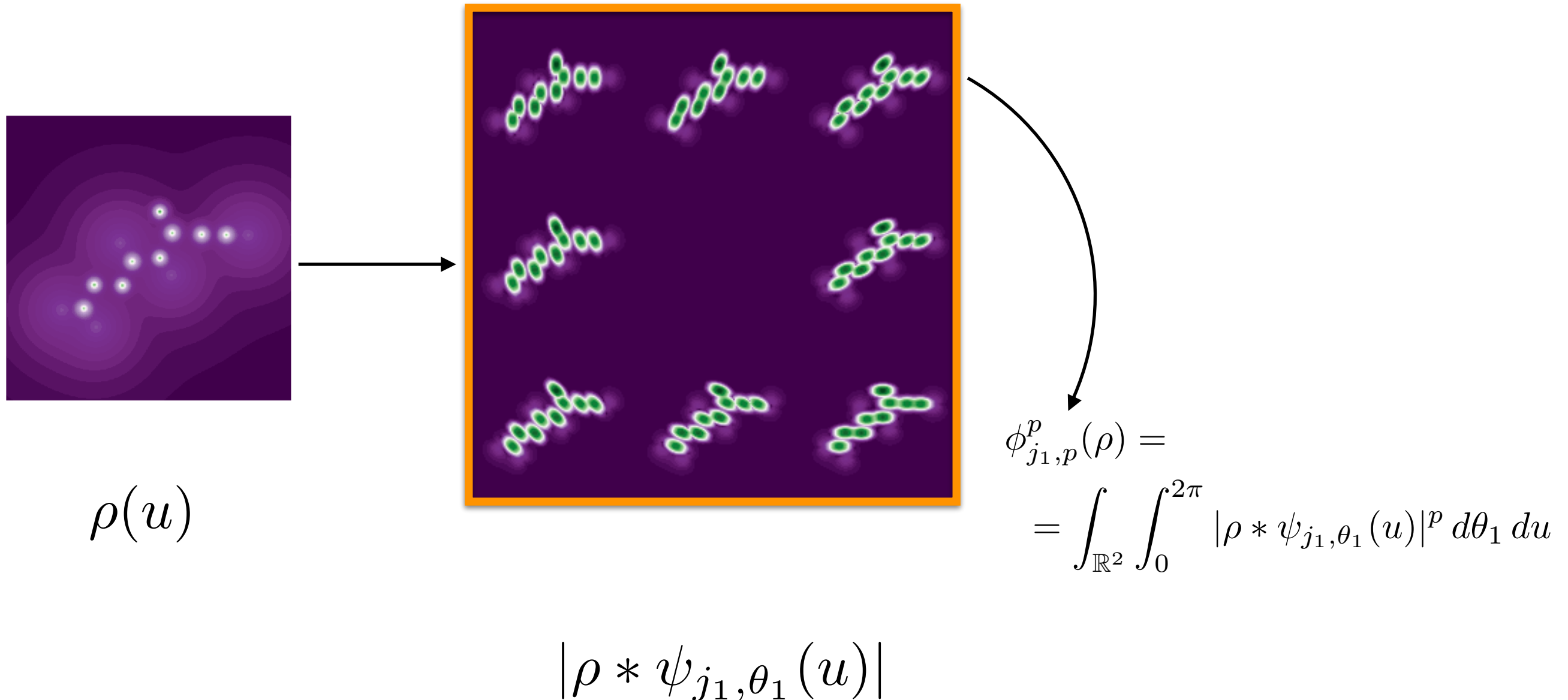


$$\Phi_S(\rho) = \{\phi_0(\rho), \phi_{j_1, 1}(\rho), \phi_{j_1, 1}^2(\rho), \phi_{j_1, 2}^2(\rho), \phi_{j_1, \lambda_2, 1}(\rho), \phi_{j_1, \lambda_2, 1}^2(\rho), \phi_{j_1, \lambda_2, 2}^2(\rho)\}_{j_1, \lambda_2}$$

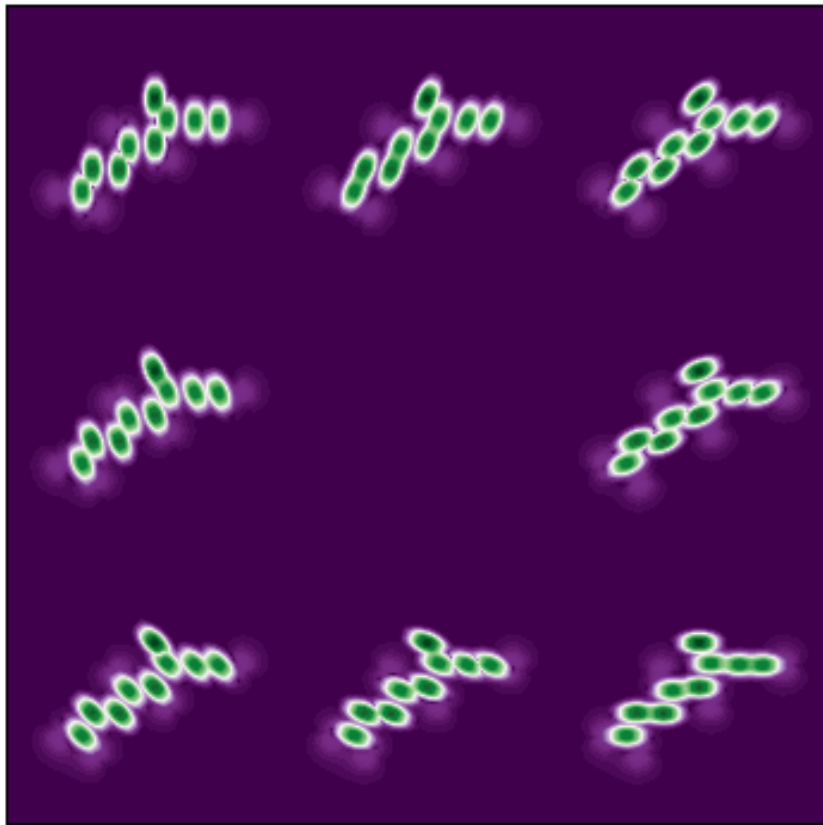
# Scattering in 2D: 1st Layer



# Scattering in 2D: 1st Layer



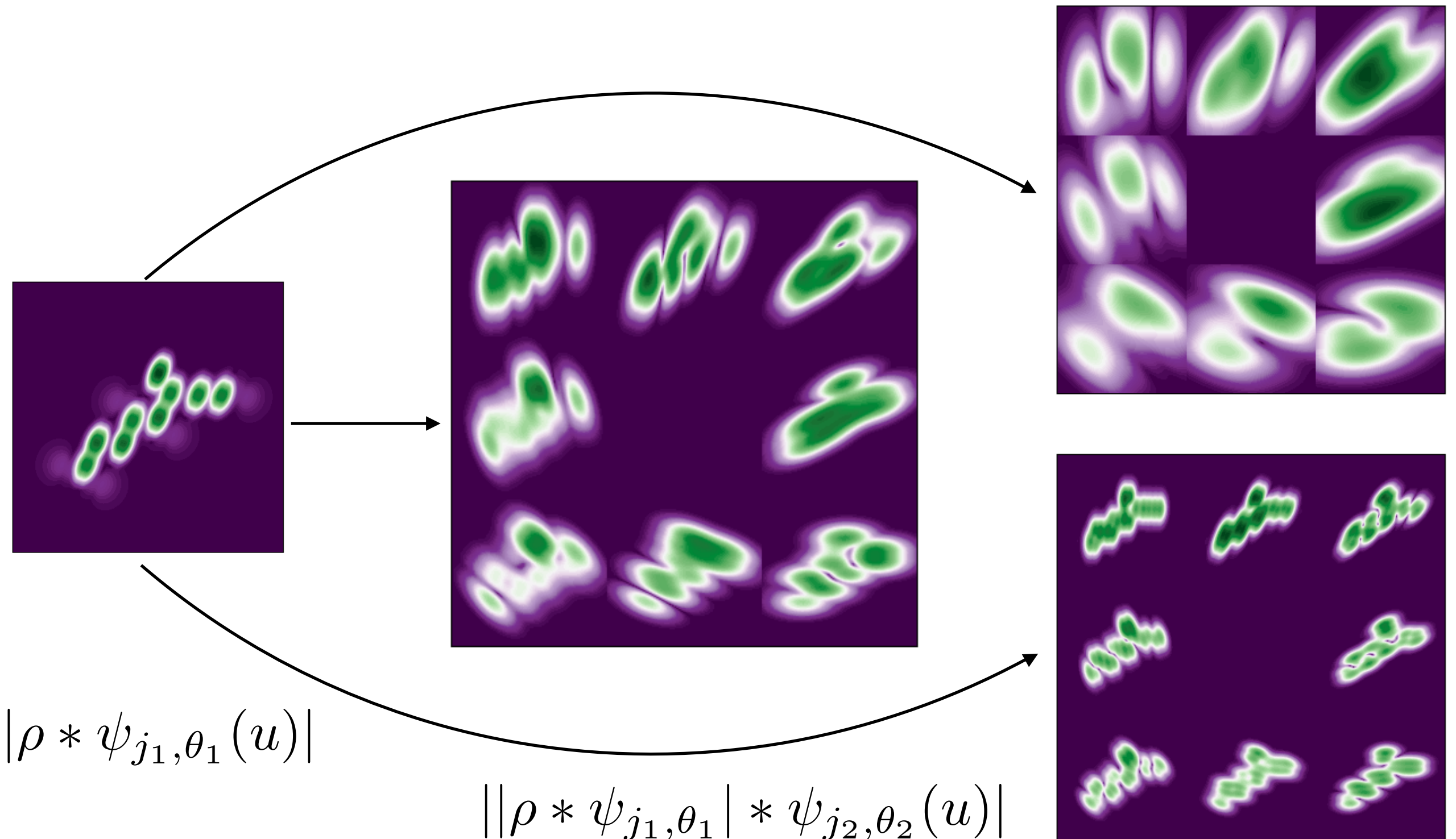
# Scattering in 2D: 2nd Layer Translation Variability



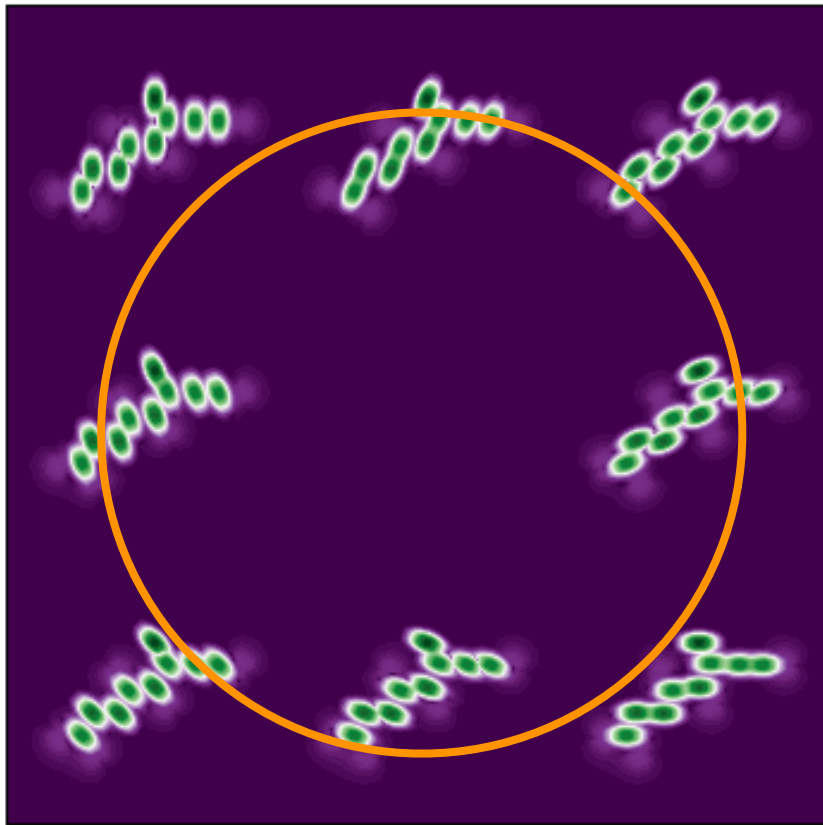
$$|\rho * \psi_{j_1, \theta_1}(u)|$$



# Scattering in 2D: 2nd Layer Translation Variability

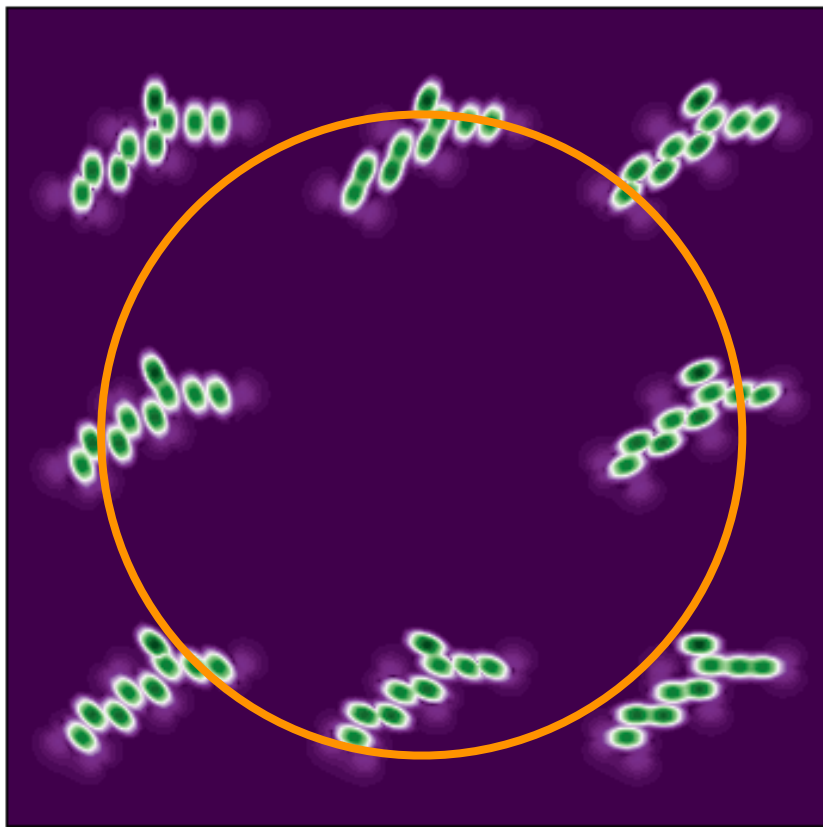


# Scattering in 2D: 2nd Layer Rotation Variability



$$|\rho * \psi_{j_1, \theta_1}(u)|$$

# Scattering in 2D: 2nd Layer Rotation Variability



- 1D wavelet  $\psi^{1D}$  periodized over  $[0, 2\pi)$ :

$$\bar{\psi}_l(\theta) = \sum_{k \in \mathbb{Z}} \psi_l^{1D}(\theta - 2\pi k)$$

- 2nd layer wavelet transform over angles defined in terms of circular convolution:

$$|\rho * \psi_{j_1, \theta_1}(u)|$$

$$|\rho * \psi_{j_1, \cdot}(u)| \circledast \bar{\psi}_{l_2}(\theta_1)$$

# Scattering in 2D:

## Roto-Translation 2nd Layer

- Spatial 2D convolution to recover translation variability:

$$|\rho * \psi_{j_1, \theta_1}| * \psi_{j_2, \theta_2}(u)$$

- Circular 1D convolution to recover rotation variability:

$$|\rho * \psi_{j_1, \cdot}(u)| \circledast \overline{\psi}_{l_2}(\theta_1)$$

- Combining yields a 3D convolution to recover roto-translation variability:

$$||\rho * \psi_{j_1, \cdot}(u)| \star \Psi_{j_2, \theta_2, l_2}(u, \theta_1)| = ||\rho * \psi_{j_1, \cdot}| * \psi_{j_2, \theta_2}(u) \circledast \overline{\psi}_{l_2}(\theta_1)|$$

where:

$$\Psi_{j_2, \theta_2, l_2}(u, \theta) = \psi_{j_2, \theta_2}(u) \overline{\psi}_{l_2}(\theta)$$

$$\star = (*, \circledast)$$

# Scattering in 3D:

## 1st Layer

- $E(3) = \mathbb{R}^3 \rtimes O(3)$  and  $O(3) = S^2 \rtimes O(2)$
- If we use a wavelet  $\psi$  that is radially symmetric about an axis  $\eta_0$ , then we can ignore the  $O(2)$  component since  $\psi$  will not vary over  $O(2)$

if  $r\eta_0 = \eta_0$  then  $\psi(ru) = \psi(u)$ ,  $r \in O(3)$

$$\psi(u) = g(u)(e^{i\eta_0 \cdot u} - C)$$

- For the first layer wavelet transform, this means we can index the rotation by  $\eta \in S^2$ :

$$\psi_{j,r}(u) = \psi_{j,\eta}(u) = 2^{-3\frac{j}{Q}} \psi(2^{-\frac{j}{Q}} r^{-1}u), \quad \eta = r\eta_0 \in S^2, \quad j \in \mathbb{Z}$$

$$\rho(u) \mapsto |\rho * \psi_{j,\eta}(u)|$$

$$\phi_{j,p}(\rho) = \left( \int_{\mathbb{R}^3} \int_{S^2} |\rho * \psi_{j,\eta}(u)|^p d\eta du \right)^{1/p}$$

# Scattering in 3D:

## 2nd Layer

- The second layer can be computed as two separable wavelet transforms, one over translations ( $\mathbb{R}^3$ ) and one over rotations ( $S^2$ ).

- Isotropic wavelet over  $S^2$ :

$$\bar{\psi}_{l,\nu} : S^2 \rightarrow \mathbb{R}, \text{ scale } 2^l \text{ and translation } \nu \in S^2$$

- Wavelet transform over  $\mathbb{R}^3$  with the same Morlet wavelet:

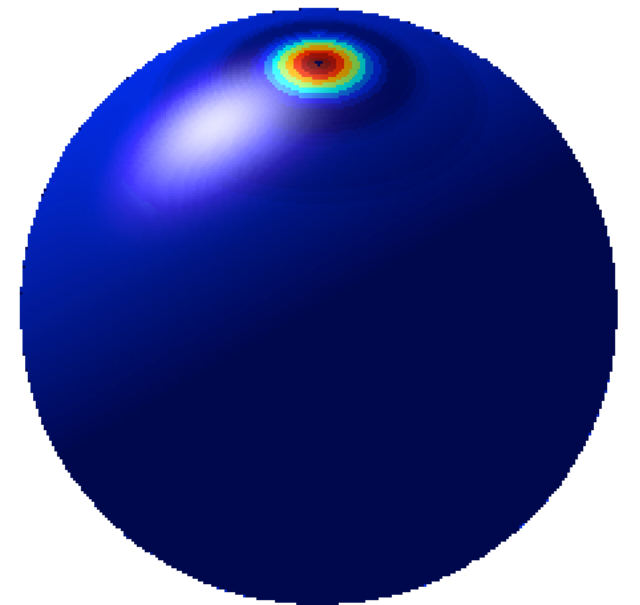
$$|\rho * \psi_{j_1,\eta}| * \psi_{j_2,\eta_2}(u)$$

- Followed by the wavelet transform over  $S^2$ :

$$\int_{S^2} |\rho * \psi_{j_1,\eta}| * \psi_{j_2,\eta_2}(u) \bar{\psi}_{l_2,\nu}(\eta) d\eta$$

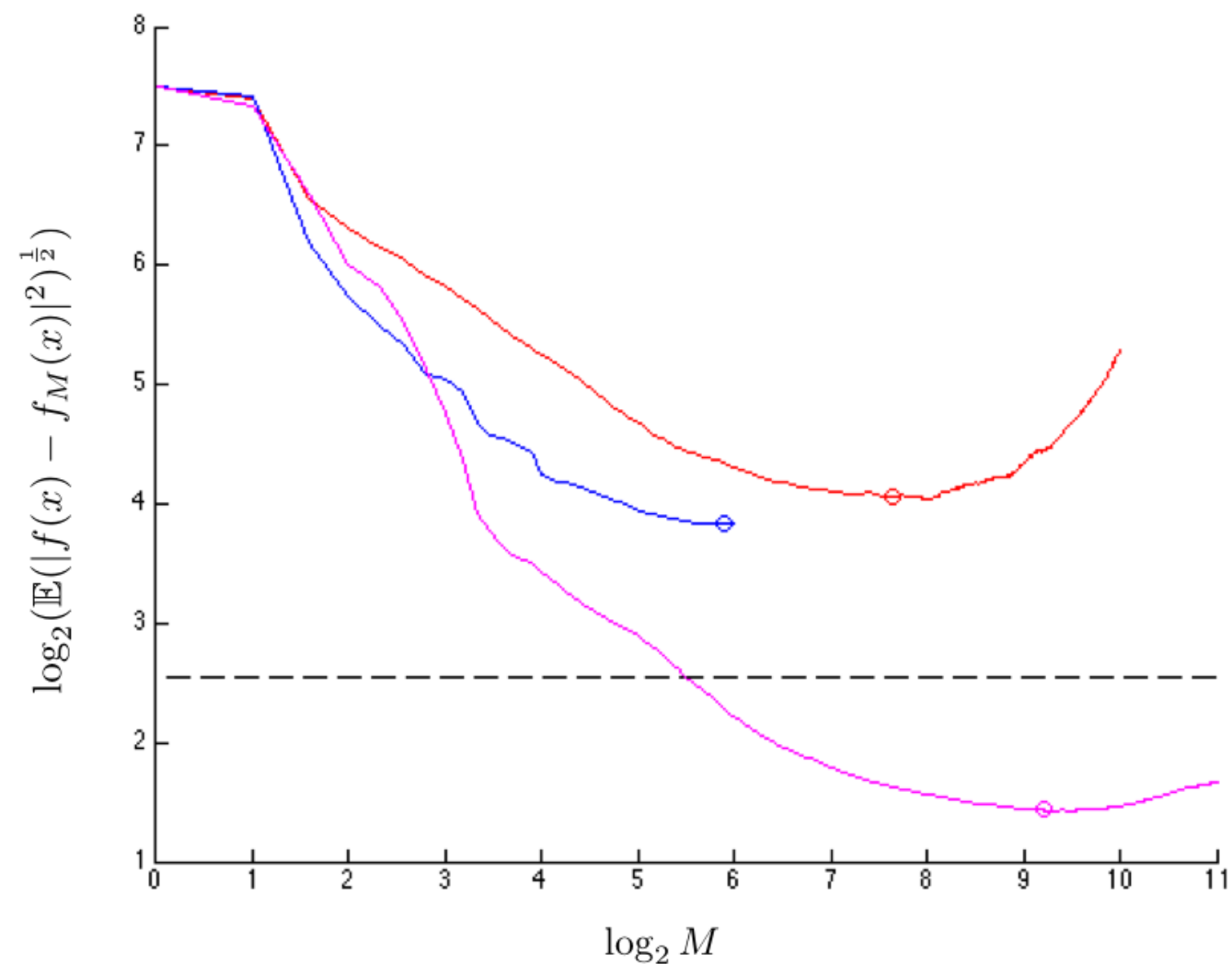
- Second layer functionals:

$$\phi_{j_1,j_2,\eta_2,l_2,p}(\rho) = \left( \int_{\mathbb{R}^3} \int_{S^2} \left| \int_{S^2} |\rho * \psi_{j_1,\eta}| * \psi_{j_2,\eta_2}(u) \bar{\psi}_{l_2,\nu}(\eta) d\eta \right|^p d\nu du \right)^{1/p}$$



# Scattering

## M-term Regression Error



Key: Fourier, Wavelets, Scattering, Coulomb (dashed line)

# Numerical Results

	Coulomb	Fourier	Wavelet	Scattering	Chemical Accuracy
$\ell^1$ : MAE	2.4	11	11	<b>1.8</b>	1.0
$\ell^2$ : RMSE	5.8	17	14	<b>2.7</b>	
$\ell^\infty$ : Max	224	272	97	<b>42</b>	

Error in kcal/mol

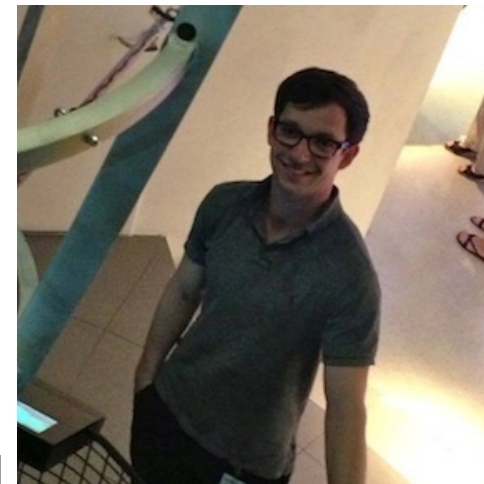
- Scattering terms:

- First term is total charge:  $\phi_0(\rho) = \int_{\mathbb{R}^3} \rho(u) du = \sum_k q_k$
- Other selected terms correspond to important geometric scales that range over the distance between two neighbouring atoms and the diameter of the molecule



# Collaborators

- Ronald Coifman
- Ariel Herbert-Voss
- Roy Lederman
- Erwan Le Gruyer
- Stéphane Mallat
- Nicholas Marshall
- Frederick McCollum
- Nicolas Poilvert
- Christian Smith



# Conclusion

- No “one size fits all” approach to Big Data. Different types of data require different mathematical tools.
- However, there are recurring themes: regularity, geometry, stability, local to global tools, multiscale structure.
- For Whitney interpolants in dimension greater than one, we are moving from pure analysis, to theoretical computer science, and now to applied CS/math with the first practical algorithms.
- With the scattering transform, we can learn physics through data and compute fast.

<http://www.di.ens.fr/~hiron/>