The Scattering Transform for Data with Geometric Structure

Michael Perlmutter

Department of Mathematics
University of California, Los Angeles

Outline

- The Euclidean Scattering Transform
- Graph and Manifold Scattering
- Incorporating Learning

The (Euclidean) Scattering Transform - S. Mallat (2012)

Overview:

- Model of Convolutional Neural Networks.
- Predefined (wavelet) filters.

Advantages:

- Provable stability and invariance properties.
- Very good numerical results in certain situations.
- Needs less training data.

Example Task: Image Classification

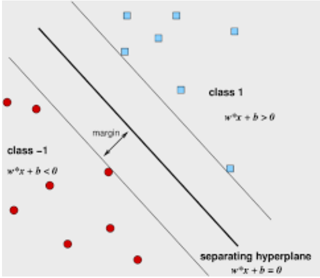
- You have a data set of many photos of cats and dogs.
- How do you decide if a new image is a cat or a dog?

Scattering is an Embedding

- Deep Neural Networks consist of an embedding an a classifier
- An **embedding** (front end) creates a hidden representation of each input in some high-dimensional vector space

$$\mathbf{x} \mapsto h(x) = (h_i(\mathbf{x}))_{i=1}^{H}$$

- The **classifier** (back end) then makes the final prediction



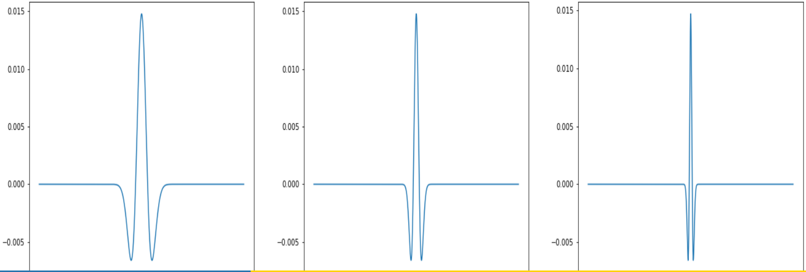
The Wavelet Transform

Definition:

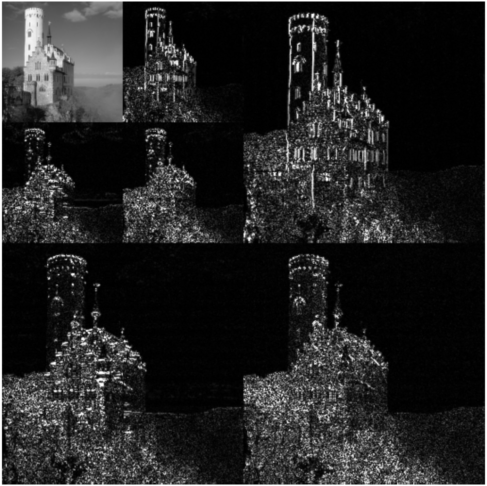
- $W_j f(x) = (\psi_j \star f)(x)$,
- $\psi_j(x) = \frac{1}{2^j}\psi\left(\frac{x}{2^j}\right)$ for some mean zero "mother wavelet" ψ.

Properties

- Collects information at different scales of resolution or frequency bands
- Heuristic: $\text{supp}(\hat{\psi}_j) \approx [2^{-j}a, 2^{-j}b]$



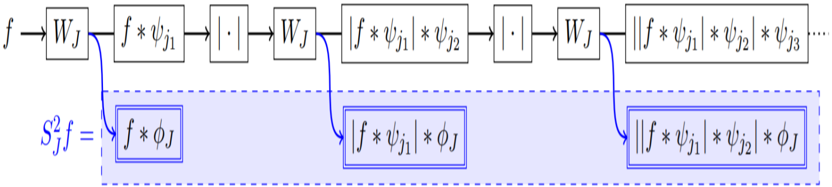
Wavelets Sparsify Natural Images



The Scattering Transform

The Scattering Transform:

- Multilayered cascade of nonlinear measurements.
- Each "layer" uses a wavelet transform W_J and a nonlinearity,
- $U_j f(x) = \sigma((\psi_j \star f)(x)),\ j \leq J, \quad \sigma(x) = M(x) = |x|.$
- $U_{j_1,j_2} f(x) = U_{j_2} U_{j_1} f(x)$
- $U_{j_1,\ldots,j_m} f(x) = U_{j_m} \ldots U_{j_1} f(x)$
- $S_{j_1,\ldots j_m} f(x) = \phi_J \star U_{j_1,\ldots,j_m} f(x), \quad \phi_J(x) = \frac{1}{2^J}\phi\left(\frac{x}{2^J}\right),$ or,
- $\bar{S}_{j_1,\ldots j_m} f = \|U_{j_1,\ldots,j_m} f\|_1.$



Why a Nonlinear Structure?

A good representation should be:

- Stable on $\mathbf{L^2}$
- Invariant to translations (or rotations etc.)
- Sufficiently descriptive

The limits of linearity:

A linear network can be invariant or descriptive, but not both.

- $\widehat{f}(0) = \int_{\mathbb{R}^d} f(x)dx$ is invariant, but throws away all high-frequency information.
- Filters which focus in on high-frequency information are unstable to translations.

The wavelet transform captures high-frequency information, and the modulus pushes this information down to lower frequencies.

Theorem (Mallat 2012)

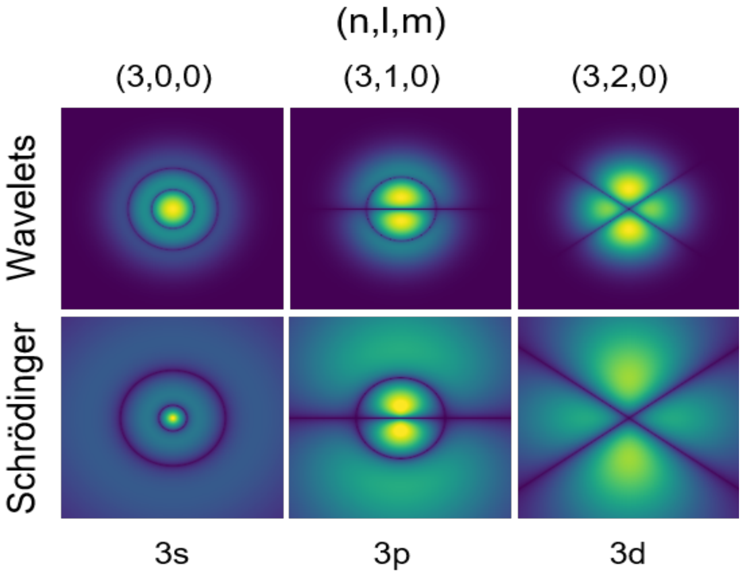
Scattering is stable on $\mathbf{L}^2$ and invariant to translations.

Limited Data Environment - Scattering for Stylometry

Which one is a Van Gogh?

- *Scattering Transform and Sparse Linear Classifiers for Art Authentication* (Leonarduzzi, Liu, and Wang)
- Dataset of 64 real Van Gogh's and 15 fakes.
- Scattering achieves state-of-the-art (96%) accuracy.

Scattering for Quantum Chemistry



Same Power Spectrum, Different Scattering

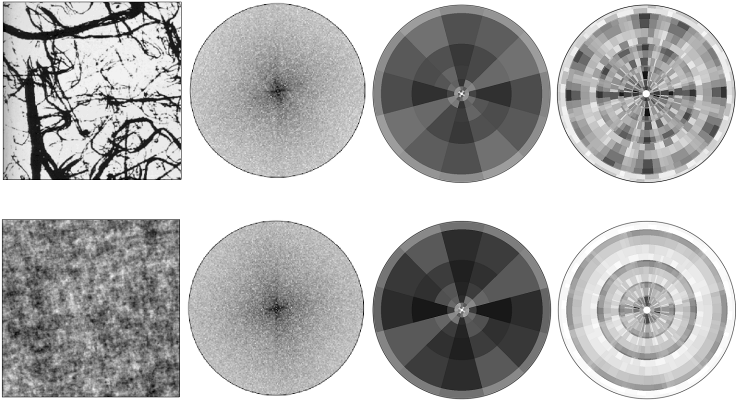
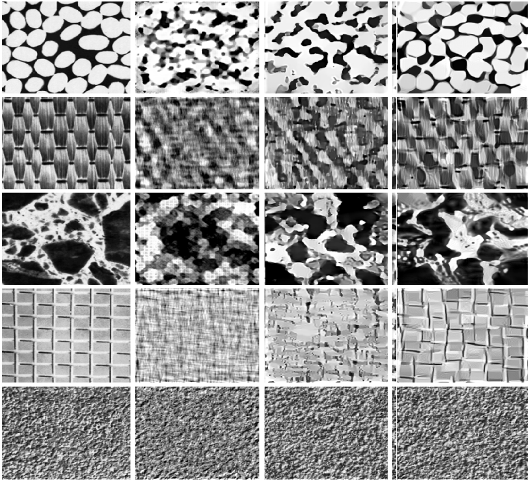


Figure 9: Two different textures having the same Fourier power spectrum. (a) Textures $X(u)$. Top: Brodatz texture. Bottom: Gaussian process. (b) Same estimated power spectrum $\widehat{R}X(\omega)$. (c) Nearly same scattering coefficients $S_J[p]X$ for $m = 1$ and 2^J equal to the image width. (d) Different scattering coefficients $S_J[p]X$ for $m = 2$.

Synthesis of random textures



(a) (b) (c) (d)

(a): Original texture. (b): texture synthesized with wavelet $\mathbf{l^2}$ norms. (c): synthesized with wavelet $\mathbf{l^1}$ norms. (d): synthesized with scattering coefficients.

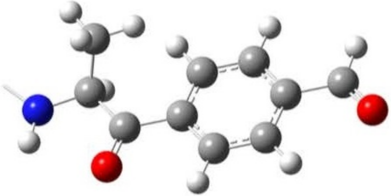
Geometric Scattering on Graphs and Manifolds

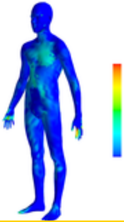
Geometric Wavelets

- Key challenge is defining wavelets.
- Once wavelets are defined, scattering is then an alternating cascade of wavelets and non-linearities.

Different Version of the Graph Scattering Transform

- Dongmian Zou and Gilad Lerman
- Fernado Gama, Alejandro Ribeiro, and Joan Bruna
- Gao, Wolf, and Hirn





Spectral Convolution

Generalized Fourier Multiplication

Let L be the Laplace-Beltrami operator or graph Laplacian with eigenbasis $\{\varphi_k\}$, $L\varphi_k = \lambda_k\varphi_k$. A spectral convolution operator has the form

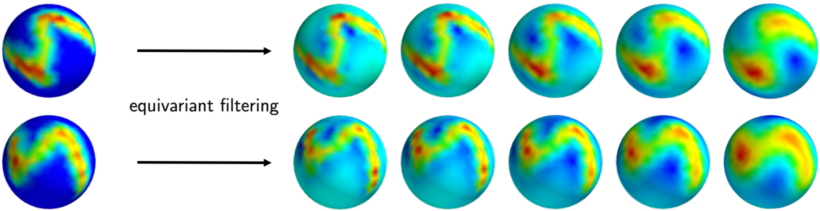
$$Tf = \sum_{k=0}^{\infty} h_k \langle f, \varphi_k \rangle \varphi_k.$$

This notion of convolution is used in many popular Graph Neural Networks such as ChebNet (Defferrard et al. 2016) or CayleyNet (Levie et al. 2017)

Spectral filters

T is called a spectral filter if $h_k = h(\lambda_k)$

Equivariant Filters



Theorem: (P., Gao, W., Hirn)

Spectral filters commute with isometries on a manifold or permutations of a graph.

Heat Flow

Heat Semigroup

$\{P_t\}_{t\geq 0}$ family of operators such that $u(x,t) = P_t f(x)$ solves

$$L_x u = \partial_t u, \quad u(x,0) = f(x).$$

Spectral Representation

$$P_t f(x) = \sum_{k=0}^{\infty} g(\lambda_k)^t \langle f, \varphi_k \rangle \varphi_k, \quad g(\lambda) = e^{-\lambda}$$

Geometric Descriptor

- Heat diffuses differently on manifolds of different shapes.

Probablistic Interpretation

- $P_t f(x) = \mathbb{E}(X_t | X_0 = x)$, where $(X_t)_{t\geq 0}$ is a Brownian Motion.

Spectral Wavelets

Definition

$$\mathcal{W}_J^{(1)} f(x) = \{\Psi_j^{(1)} f(x), \Phi_J^{(1)} f(x)\}_{0 \leq j \leq J},$$

where $\Phi_J^{(1)} = P_{2^J}$, $g(\lambda) = e^{-\lambda}$ and

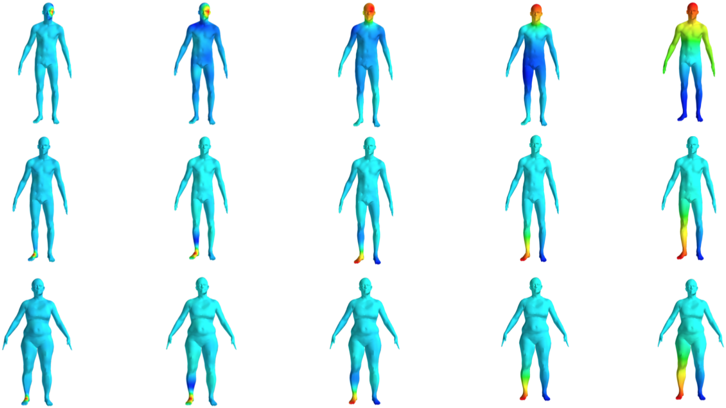
$$\Psi_j^{(1)} f = (P_{2^{j+1}} - P_{2^j})^{1/2} f = \sum_{k=0}^{\infty} [g(\lambda_k)^{2^{j+1}} - g(\lambda_k)^{2^j}]^{1/2} \langle f, \varphi_k \rangle \varphi_k.$$

Theorem: P., Gao, Wolf, Hirn

$\mathcal{W}_J^{(1)}$ is an isometry, i.e.,

$$\sum_j \|\Psi_j^{(1)} f\|^2 + \|\Phi_J^{(1)} f\|^2 = \|f\|^2.$$

Wavelets on the Faust Dataset



Diffusion Wavelets on Manifolds

Definition

$$\mathcal{W}_J^{(2)} f(x) = \{\Psi_j^{(2)} f(x), \Phi_J^{(2)} f(x)\}_{0 \le j \le J},$$

where $\Phi_J^{(2)} = P_{2^{J+1}}$, $g(\lambda) = e^{-\lambda}$ and

$$\Psi_j^{(2)} f = (P_{2^{j+1}} - P_{2^j}) f = \sum_{k=0}^{\infty} [g(\lambda_k)^{2^{j+1}} - g(\lambda_k)^{2^j}]^1 \langle f, \varphi_k \rangle \varphi_k.$$

Theorem: P., Gao, Wolf, Hirn

$\mathcal{W}_J^{(2)}$ is a non-expansive frame on a suitable weighted space, i.e.,

$$c\|f\|^2 \le \sum_j \|\Psi_j^{(2)} f\|^2 + \|\Phi_J^{(2)} f\|^2 \le \|f\|^2.$$

Diffusion Wavelets on Graphs

Probabilistic interpretation

- On a manifold, the heat-semigroup describes the transistion probabilities of a Brownian motion.
- Natural Analog on graphs is a (lazy) random-walk.

Definition

Let G be a graph and let P be a lazy random walk matrix. For $0 \leq j \leq J$, let

$$\Psi_j^{(2)} = P^{2^{j+1}} - P^{2^j}, \quad \Phi_J^{(2)} = P^{2^{J+1}}.$$

LEGS - Learnable Scales

Subsequent work with Tong et. al showed that dyadic scales are unnecessary and the same result holds with *any* sequence of increasing scales. Moreover, one may learn the scales through data.

Theoretical Guarantees Manifold Scattering

Theorem (P. Gao, Wolf, Hirn)

$$\|Sf_1 - Sf_2\| \leq \|f_1 - f_2\|, \quad \forall f_1, f_2 \in \mathbf{L}^2(\mathcal{M}).$$

Theorem (P. Gao, Wolf, Hirn)

Let ζ be an isometry, $V_\zeta f(x) = f(\zeta^{-1}(x))$.

$$\|Sf - SV_\zeta f\| = \mathcal{O}\left(2^{-dJ}\right) \quad \forall f \in \mathbf{L}^2(\mathcal{M}) \ .$$

Theorem (P. Gao, Wolf, Hirn)

Let ζ be an diffeomorphism, and assume f is bandlimited (finitely many non-zero Fourier coefficients). Then

$$\|Sf - SV_\zeta f\| = \mathcal{O}\left(2^{-dJ}\right) + \mathcal{O}\left(\lambda_{\max}^d d(\zeta, Isom)\right).$$

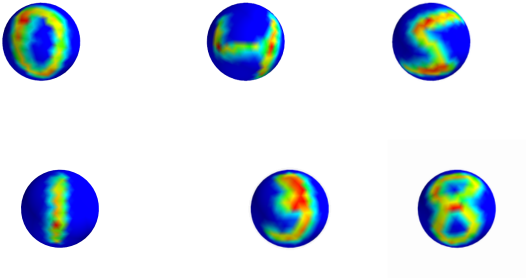
Theorem (P., Gao, Wolf, Hirn)

Similar results hold for graph scattering.

Manifold Scattering Results

Example (Spherical MNIST)

MNIST digits projected on the sphere:

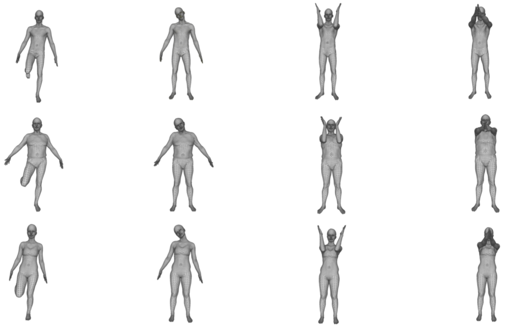


- Single manifold, multiple signals
- 95% classification accuracy from scattering features

Manifold Scattering Results

Example (FAUST dataset)

Ten people in ten different poses:



- Mesh grids & Shot features (Tombari et al., 2010; Prakya et al.,2015)
- Accuracy: 81% person recognition, 95% pose classification

Scattering on Point Clouds

Motivation (The Manifold Hypothesis)

- In many real-world applications you don't know the manifold
- Instead, you have a high-dimensional point cloud which you model as lying upon an unknown manifold

ICML Workshop (TAGs ML)- Joint work with Chew, Steach, Viswani, Needell, Krishnaswamy, Hirn, and Wu

- Diffusion maps style algorithm for implementing manifold scattering on point clouds
- Recover mesh-based results on spherical MNIST
- Apply method to single-cell data
- Convergence guarantees coming soon

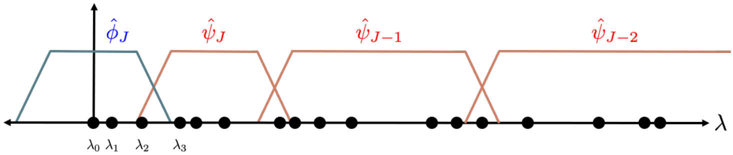
Geometric Wavelets vs GCN style filters

GCN Style Filters

- Take averages over local neighborhoods - promote smoothness
- Low-pass filter

Wavelets

- Detects changes at different scales
  - How is my four-step neighborhood different than my two-step neighborhood?
- Band-pass filter
- Capture long range interactions



Discriminative Power

When can a network tell two nodes apart?

- Necessary condition: The network learns different representations of the two nodes
- Lots of work on the analogous problem for graph classification
  - GCN $\lesssim$ Weisfeiler-Lehman Kernel
- Little work for node classification
- Do GCNs rely on informative features? Or can they learn from the geometry of the graph?

Theorem (Wenkel, Min, Hirn, P., and Wolf (2022))

- *There are situations where GCN provably not discriminate two nodes if their local neighborhoods have the same structure*
- *Graph Scattering can discriminate some of those nodes*
- *Thus GCN-Scattering Hybrid networks have more discriminative power than pure GCN networks.*

Incorporating Learning

- Scattering helps us understand GNNs and a theoretical level
- Let's use this understanding to build (trained) GNNs incorporating the principals of scattering

Scattering GCN Hybrid

Scattering Channels

Layer-wise update rule:

$$X_{sct}^{\ell} := \sigma\left((P^{2^{J+1}} - P^{2^J})X^{\ell-1}\Theta\right).$$

Hybrid Network

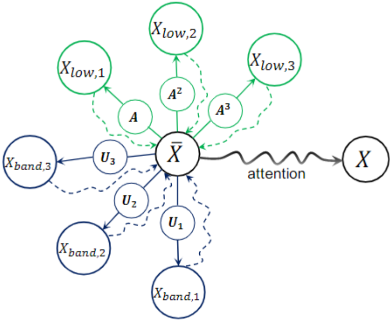
- Wenkel, Min, Hirn, P., and Wolf (2022) use both GCN channels and Scattering channels of each layer.
- GCN channels focus on low-frequency information.
- Scattering Channels retain high-frequency information.
- Can use an attention mechanism to balance channel ratios.

Scattering Attention Network

Attention Mechanism

$$\mathbf{X}^{\ell} = C^{-1}\tilde{\sigma}\left(\sum_{j=1}^{C_{\text{low}}} \alpha^{\ell}_{\text{low},j} \odot \bar{\mathbf{X}}^{\ell}_{\text{low},j} + \sum_{j=1}^{C_{\text{band}}} \alpha^{\ell}_{\text{band},j} \odot \bar{\mathbf{X}}^{\ell}_{\text{band},j}\right)$$

$$C = C_{\text{low}} + C_{\text{high}}, \quad \alpha \odot \mathbf{X} = \text{diag}(\alpha)\mathbf{X}$$



Attention Network Results

Dataset	Classes	Nodes	Edges	Homophily	GCN	GAT	Sc-GCN	GSAN
Texas	5	183	295	0.11	59.5	58.4	60.3	60.5
Chameleon	5	2,277	31,421	0.23	28.2	42.9	51.2	61.2
CoraFull	70	19,793	63,421	0.57	62.2	51.9	62.5	64.5
Wiki-CS	10	11,701	216,123	0.65	77.2	77.7	78.1	78.6
Citeseer	6	3,327	4,676	0.74	70.3	72.5	71.7	71.3
Pubmed	3	19,717	44,327	0.80	79.0	79.0	79.4	79.8
Cora	7	2,708	5,276	0.81	81.5	83.0	84.2	84.0
DBLP	4	17,716	52,867	0.83	59.3	66.1	81.5	84.3

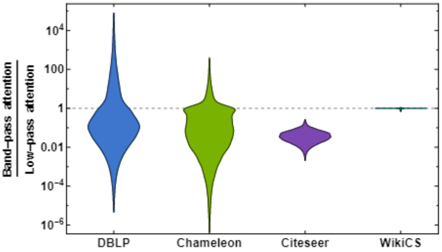
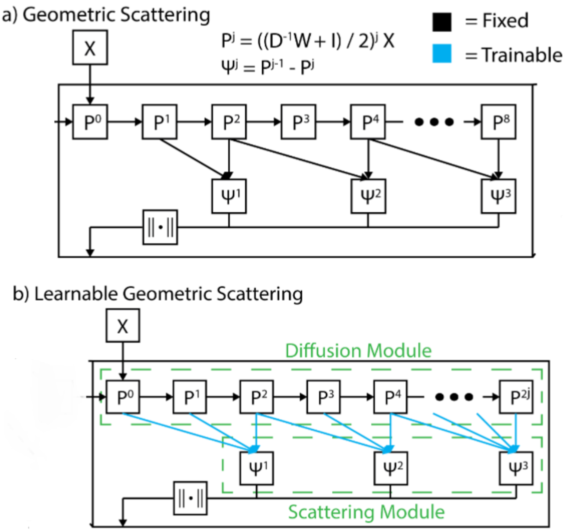


Fig. 6. Distribution of attention ratios per node between band-pass (scattering) and low-pass (GCN) channels across all heads for DBLP, Chameleon, Citeseer, and WikiCS.

LEGS - Learning the Scales



Molecular Graph Generation via Geometric Scattering (GRASSY) - Bhaskar, Grady, P., Krishnaswamy

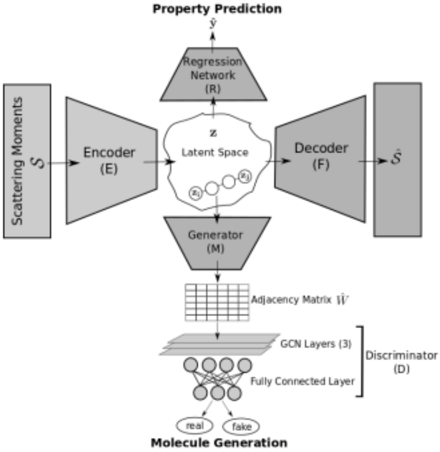


Figure: GRaph Scattering SYnthesis network

Conclusion

- The Euclidean scattering transform is a model of CNNs.
  - Provable Stability / Invariance Guarantees
  - Designed filters - useful for low-data environments
- Geometric Versions for Graphs and Manifolds
  - Similar theoretical guarantees to the Euclidean scattering transform
  - Wavelets can be constructed either spatially or spectrally
  - Can be incorporated in hybrid Scattering - GCN networks

THANK YOU!