PhD Defense:
On the efficiency of local methods in image classification and energy regression in physics.

Supervised by Prof. Stéphane Mallat.

Louis Thiry,
DATA Team, Computer Science Department, ENS, PSL.

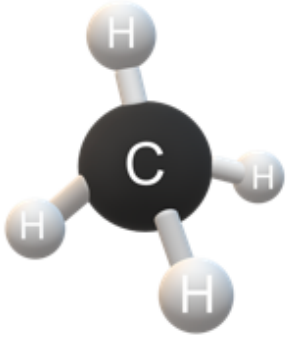
Image classification and Energy regression

High-dimensional learning problems

- Classification function F

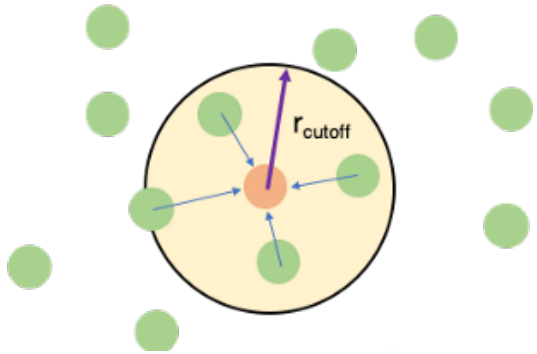
$$F : x = \text{  } \mapsto \text{"cat"}$$

- Energy function E (potential)

$$E : x = \text{  } \mapsto -417.96 \text{ kcal/mol}$$

Local methods

- Energy regression: Separation into atomic neighborhoods $\mathcal{N}_i$



Energy sum of local contributions

$$E(x) = \sum_i E\left(\mathcal{N}_i\right)$$

- Image Classification: separation of the image into patches

Sum over patch evidences

$$F(x) = \sum_{p \in x} f(p)$$

Efficiency of local methods

Spectacular progress in the last 10 years

- ImageNet classification: 70 % local methods $\rightarrow$ 98 % with Convolutional Neural Networks
- Energy regression: empirical potentials' "poor" accuracy $\rightarrow$ machine learning potential close to DFT accuracy.

- Do the local methods performs significantly worse than non-local methods on image classification and energy regression?
- How can we capture non-local components of the function we are trying to approximate?
- What are the benefits of using local separation for the predictions' interpretability and the functions' mathematical analysis?

Multi-scale and Invariance properties

Multi-scale

- Physics : small-scale ionic and covalent bonds, medium-scale Van-der-Waals interactions, large-scale Coulomb interactions.
- Image: small-scale texture information, medium-scale pattern information, large-scale shape information.

Invariance

- Energy invariant to atoms rotations and translations
- Image class invariant to scale, lightening and translations

$\rightarrow$ incorporate these a priori information to learn the classification and energy functions

Plan

1. Multi-scale invariant representation for energy regression
2. Hybrid Local Convolutional Neural Network for Image Classification
3. Patch K-nearest-neighbor classifier
4. Human-machine interactive creation

Multi-scale descriptor for energy regression

Energy is **invariant** to translation and rotations, and results from **multi-scale** interactions.

- How can we build an invariant multi-scale description of the systems ?
- Do we need a multi-scale description of the system to regress the energy of usual physical systems ?
- Shall we treat differently the different scales in such a description ?

Solid Harmonic Scattering Transform by Eickenberg, Exarchakis, Hirn, Mallat, and Thiry (2018).

Solid Harmonic Scattering Transform

1. **Density:** sum of Gaussians g_i centered at the atom positions

$$\rho(r) = \sum_i g_i(r).$$

2. **Solid Harmonic Wavelets**

- Spherical harmonics: $\mathbb{S}^2$ Fourier modes

$$Y_l^m : (\theta, \phi) \in \mathbb{S}^2 \to Y_l^m(\theta, \phi) \in \mathbb{C}, \ l \geq 0, -l \leq m \leq l$$

- Spherical harmonic mother-wavelet

$$\psi_{l,m}(u) = e^{-|u|^2/2} |u|^l Y_l^m (\theta_u, \phi_u)$$

- Dilation of the mother wavelet at the scale 2^j

$$\psi_{l,m,j}(u) = 2^{-3j} \psi_{l,m}(2^{-j} u).$$

Solid Harmonic Scattering Transform

3. "Convolution and modulus": translation and rotation **equivariant**

$$|\rho * \psi_{l,j}|(r) \triangleq \left(\sum_{m=-l}^{l} |\rho * \psi_{l,m,j}|^2(r) \right)^{1/2}$$

4. Multi-scale coefficients: translation and rotation **invariant**

- Scale coefficients

$$S^1_{l,j,q} = \| |\rho * \psi_{l,j}| \|_q = \int |\rho * \psi_{l,j}|^q, \ q = 1, 2$$

- Scale interaction coefficients

$$S^2_{l,j,l',j',q} = \left\| \left| |\rho * \psi_{l,j}| * \psi_{l',j'} \right| \right\|_q, \ q = 1, 2$$

- Same frequencies l for all the scales

Spatial location, aliasing, and grid size

1. **Spatial location**: set Gaussians' width σ to keep atom location.
   - Minimal interatomic distance, d_{min}, Gaussian overlap amplitude α

$$\sigma = \frac{d_{\text{min}}}{\sqrt{-8\log(\alpha)}}$$

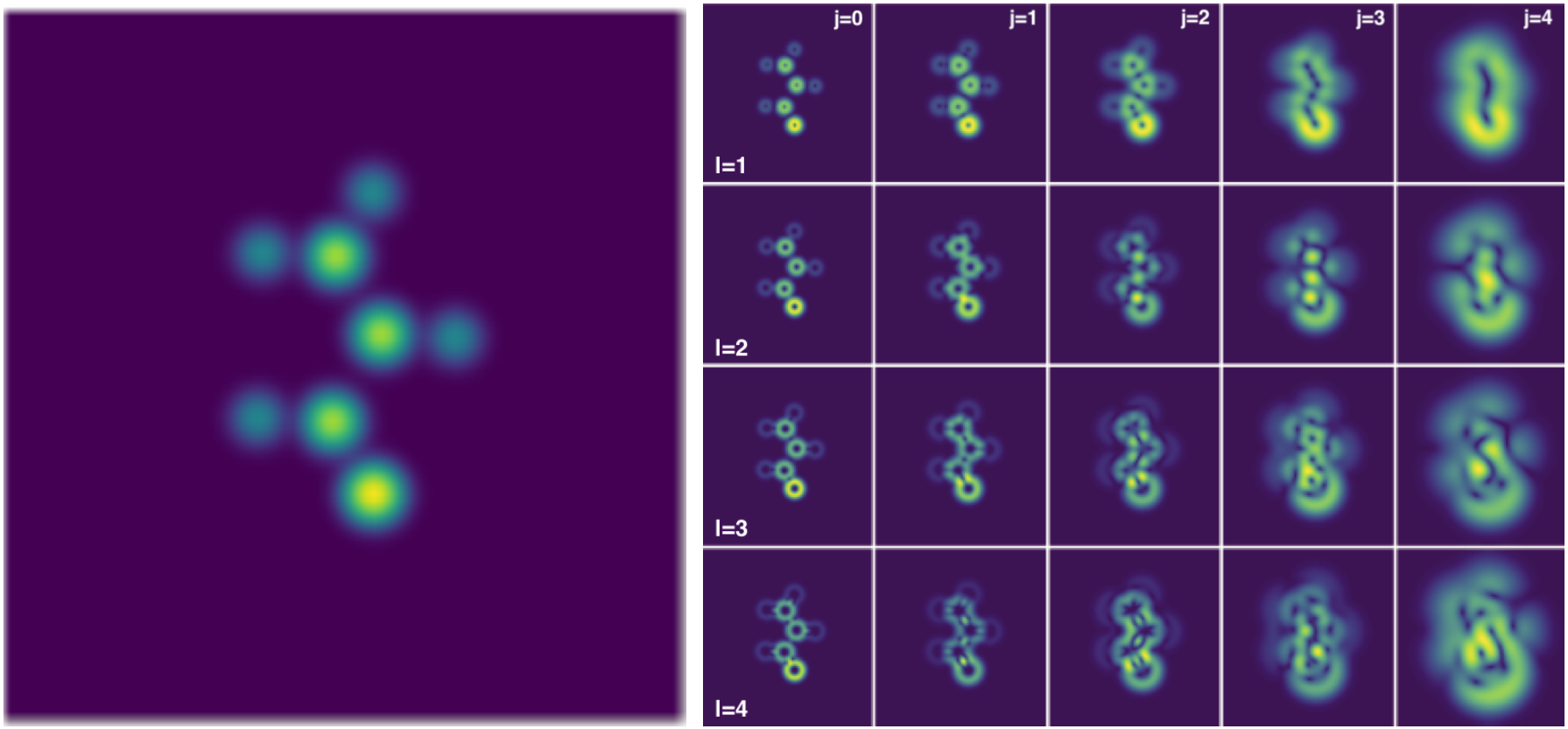
2. **Aliasing** : Density ρ sampling errors discards roto-translation invariance $\rightarrow$ control the sampling step δ limit aliasing
   - Aliasing tolerance ϵ

$$\delta = -\frac{\sigma^2}{\pi^2 \log(\epsilon)}$$

3. **Size of the grid** N: L molecule or solid maximal length

$$N = \frac{L}{\delta}$$

Density and convolutions



(left) C_3H_4O molecule density and (right) Convolution and modulus with solid harmonics wavelets $\psi_{j,l}$

Organic molecules energy regression

Is Solid Harmonic Scattering Transform a suitable descriptor for molecules energy regression ?

QM9 database (Ramakrishnan et al., 2014) :

- atomization energies of $130,000$ molecules ~ -1000 kcal/mol
- computed with quantum mechanics (Density Functional Theory)
- Up to 9 non-hydrogen atoms per molecule, length up to $30\dot{A}$

QM9 atomization energy regression

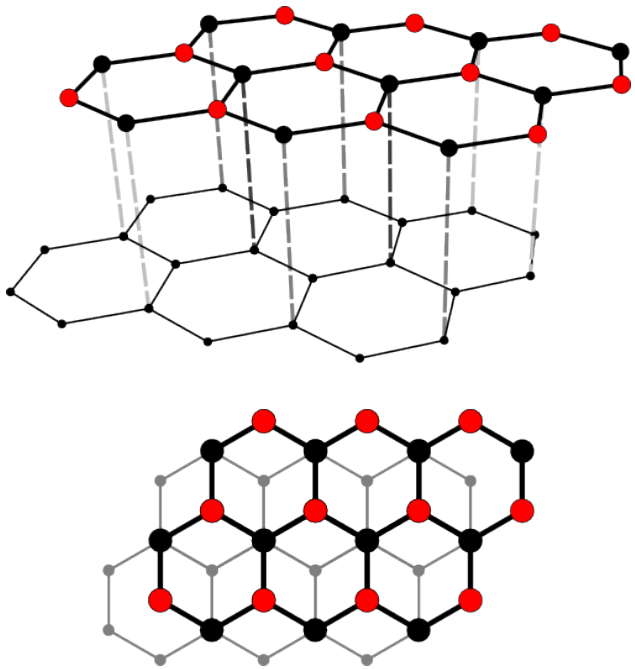
Test Mean Absolute Error on QM9 atomization energies.

method regression	Scatt. Linear	Scatt. Tri-linear	CM KRR	SchNet CNN	HDAD KRR	SOAP KRR
local	×	×	×	×	×	✓
MAE (kcal/mol)	**1.89**	**0.56**	**2.95**	**0.34**	**0.58**	**0.41**

- Efficient local method with neighborhood radius of $3\dot{A}$ (Willatt et al., 2018).
- Emprirical evidence of locality of small organic molecules' energy
- Energy is concetrated in small scales $\rightarrow$ no need for large scale and scale interaction description

Long-range energies regression

Collaboration with University of Luxemburg (Pr. A. Tkatchenko) and Cambridge University (Pr. G. Csanyi).



- Crystal of carbon atoms
- Stack of graphene layers (hexagonal structure)
- **Long-range Van-der-Waals interactions**

Long-range energies regression

Graphite database:

- 3D cubic periodic cells
- $\sim$ 500 carbon atoms
- Volker Deringer: 2500 configurations generated.
- Martin Stoehr: Many Body Dispersion (Tkatchenko et al., 2012)
  Energies computed, $\sim$ -50 eV = -1150 kcal/mol.

Multi-scale descriptions

1. Solid Harmonic Scattering Transform

1. 7 scale indices j.
2. Same Fourier representation of all scales.

2. **Ad-hoc multi-scale method**, Deringer and Csányi (2017) :

1. Short-range SOAP kernel: $3\dot{A}$.
2. Medium-range SOAP kernel: $6\dot{A}$.
3. Long-range pair potential: $10\ \dot{A}$.

$$E_{\text{l-r}}(x) = \sum_{r_{ij}<10\dot{A}} f(r_{ij})$$

$\rightarrow$ Strong assumption on long-range energies

Energy regression results

1. **Multi-scale Scattering**: 49.8 meV MAE.
2. **Ad-hoc multi-scale method** : 52.8 meV MAE.

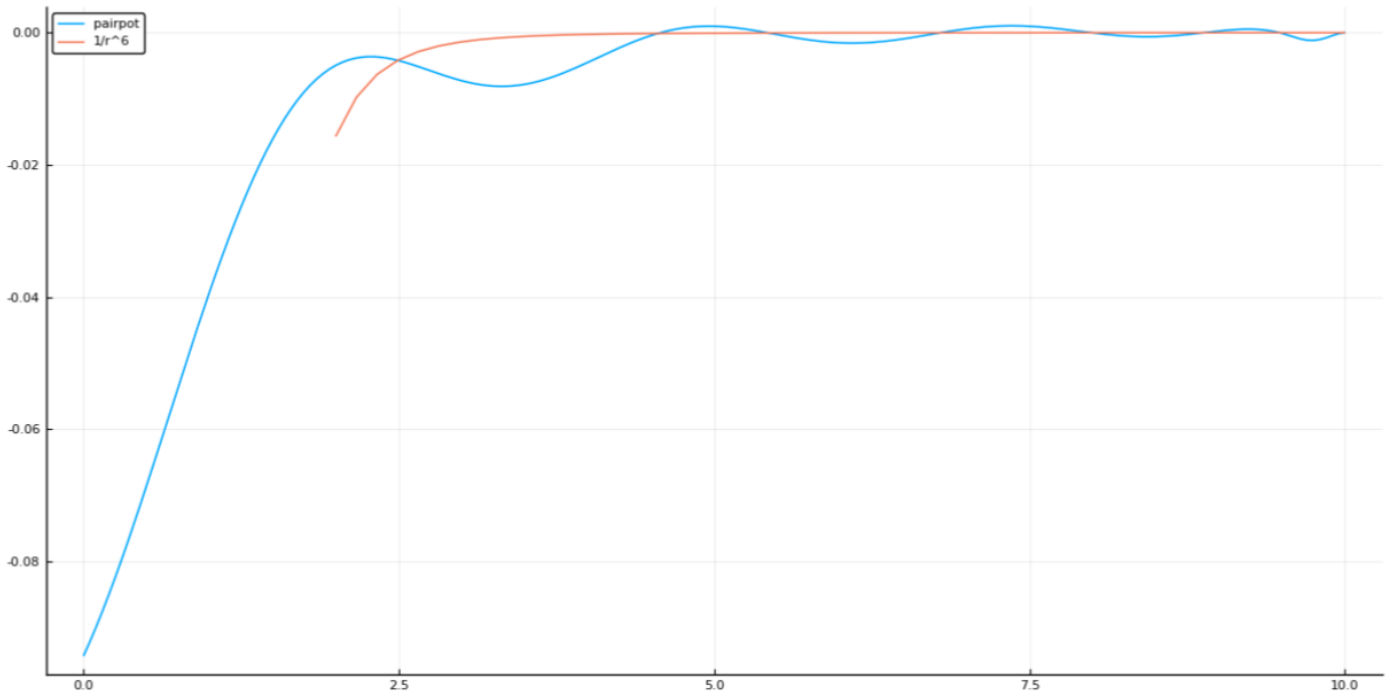
Comments:

- 50 meV = 1.2 kcal/mol, 3 meV = 0.07 kcal/mol.
- Ad-hoc 3 scales description is as efficient as Solid Scattering with 7 scale description.
- Long-range energy terms are essentially two-body
- Short-range efficiently captured with local SOAP descriptor

Energy components analysis

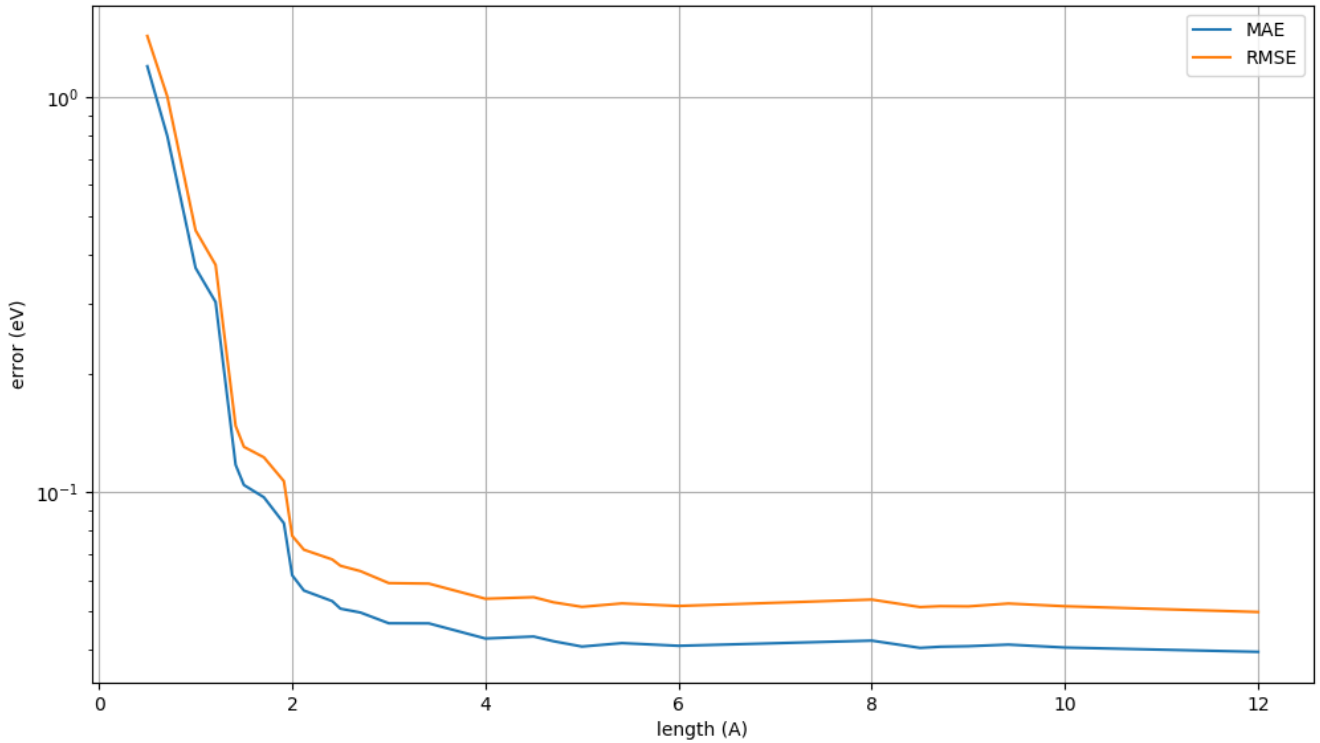
Ad-hoc multi-scale method

- Without pair potential: 70 meV MAE.
  $\rightarrow$ energy is concentrated in $6\dot{A}$ neighborhoods
- Pair potential alone, MAE : 453 meV MAE.
  $\rightarrow$ local many-body descriptor is key for good accuracy



Energy components analysis

Solid Harmonic Scattering



Error w.r.t Solid Harmonic Scattering coefficients max. length.

$\rightarrow$ energy is concentrated in $5\dot{A}$ neighborhoods

Solid Harmonic Scattering for energy regression

Solid Harmonic Scattering

- Scale and scale-interaction representation
- Invariant to rotations and translation
- Angular Fourier spectrum description of scales

Energy regression

- Energy of small organic molecules is apparently local
- Van-der-waal graphite energies: ad-hoc multi-scale method efficient
- Long-range energies are two-body
- Energy: scale components are better described separately

Free energy and vibrational entropy

Free-energy

$$A(r_1, \ldots, r_{N_a}) = E(r_1, \ldots, r_{N_a}) - T \times S(r_1, \ldots, r_{N_a})$$

Free-energy computations $\rightarrow$ $C15$ Iron phase discovery (Marinica et al., 2012).

Vibrational entropy: function of hessian eigenvalues ω_j

$$S = k_B \sum_j \left[\ln \left(\frac{k_B T}{\hbar \omega_j} \right) + 1 \right], \quad \frac{\hbar \omega_j}{k_B T} \ll 1$$

Computational cost $\mathcal{O}(N_{\text{atoms}}^3) \rightarrow$ Free-energy landscape exploration is computationally infeasible.

Vibrational entropy regression

No existing vibrational entropy regression technique

- Can we regress accurately vibrational entropy?
- Hessian is a global quantity. Do we need a multi-scale description of the configuration to regess a function of the Hessian eigenvalues?
- Can we regress a function of the Hessian eigenvalues with solely local description of the configuration ?

Configuration database

No data available $\rightarrow$ C. Marinica and C. Lapointe (CEA Saclay) created a new database.

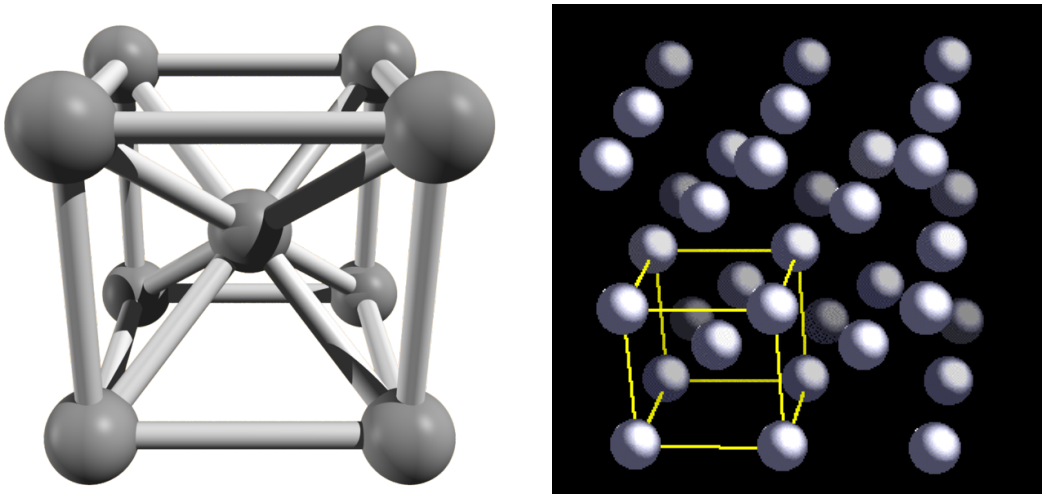


Figure: Body-centered cubic Iron

- Body-centered cubic Iron
- Defects with 1-4 removed or additional atoms
- 31,000 configurations with 1000 $\rightarrow$ 3500 atoms

Configurations representation

1. Multi-scale representation

- Solid Harmonic Scattering transform
- 9 scales and 9 Fourier indices

2. Local representation

- Angular Fourier Series (Bartók et al., 2013)

$$\mathcal{A}_{n,l}(\mathcal{N}_i) = \sum_{j,k \in \mathcal{N}_i} f\left(r_{ij}, r_{ik}, \theta_{jik}\right)$$

- **Roto-translation invariant**: function of pairwise distance r_ij and triplet angles θ_{jik}
- Global descriptor: sum of local descriptors

$$A_{n,l} = \sum_i \mathcal{A}_{n,l}(\mathcal{N}_i)$$

- Neighborhood radius $5\dot{A}$

Entropy extensivity property

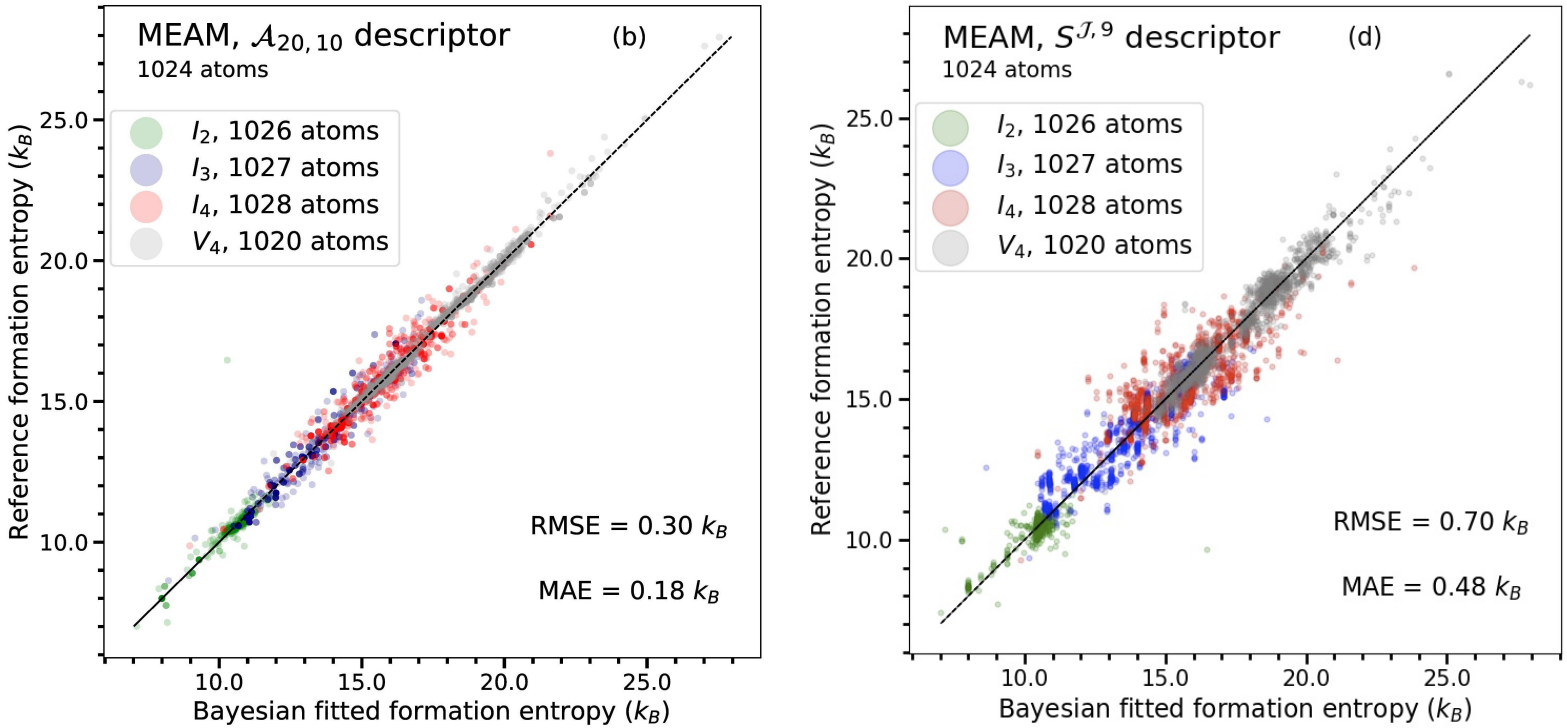
Classical thermodynamics entropy extensivity:

Twice the number of atoms yields twice bigger entropy.

- Solid Harmonic Scattering Transform is extensive with $\|.\|_q$ pooling
- Angular Fourier Series is extensive since it's a sum of local descriptors.

$\rightarrow$ **Use a linear regression**. Multi-linear regression would cancel the extensivity property.

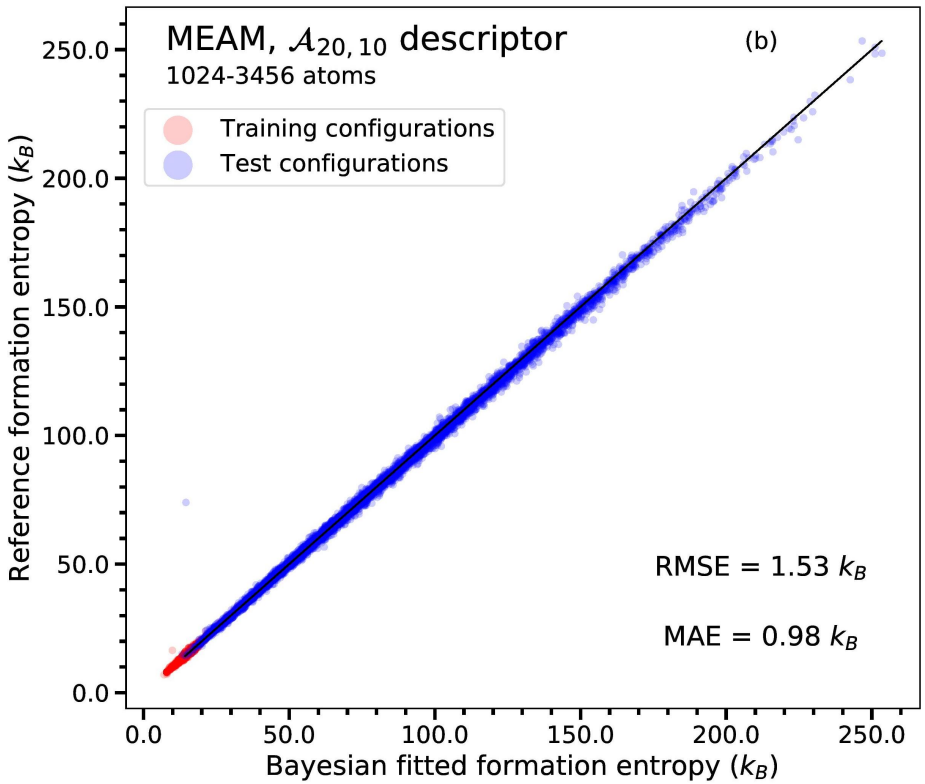
Vibrational entropy regression



Predictions with (left) AFS and (right) Solid Harmonic Scattering

- Solid Harmonic Scattering: 0.48 k_B MAE
- Angular Fourier Series: 0.18 k_B MAE

Extrapolation capacities of AFS



- Entropies range: train $5 - 25k_B$, test $10 - 250k_B$
- Extrapolation: extensivity property

Vibrational entropy regression

- Accurate direct vibrational entropy regression method
- Local AFS description performs significantly better than Multi-scale Scattering
- Ensuring extensivity property allows to extrapolate predictions
- Allows fast free-energy landscape exploration

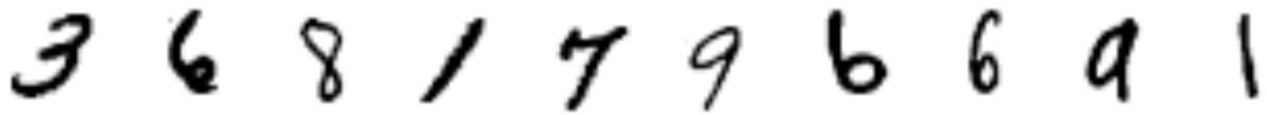
Intra-class and extra-class variability in Image Classification

1. **Intra-class variability:** varability in the set of of images of a given class

$$\mathcal{S}_y = F^{-1}(y)$$

Ex: Handwritten digits, intraclass variability is the *local* group $\mathcal{G}$ of small deformations

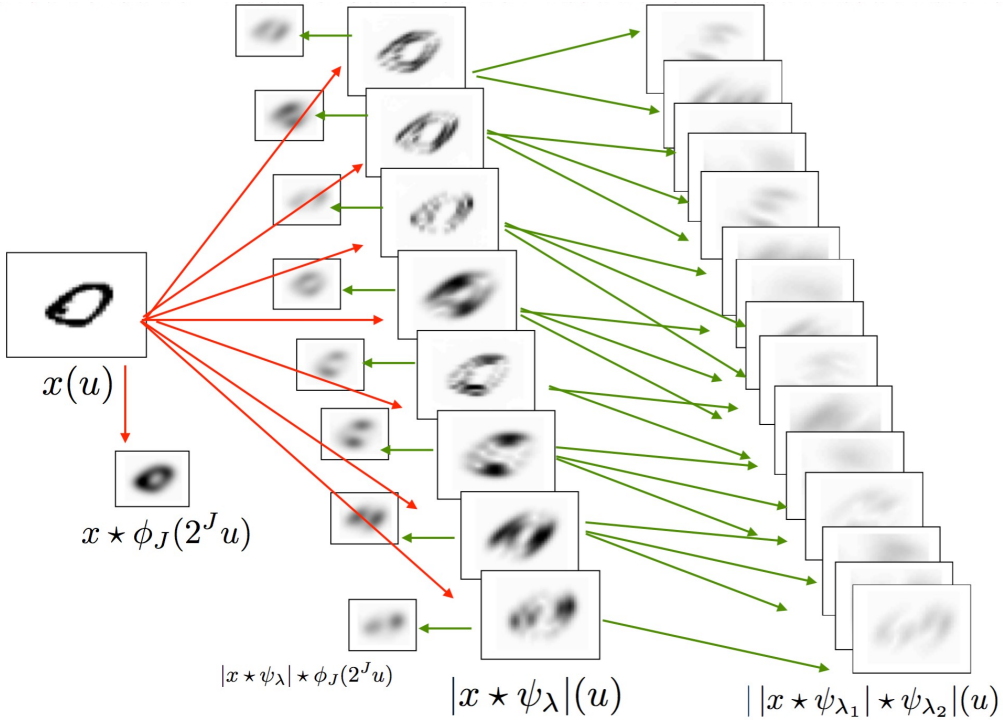
$$\mathcal{S}_y = \mathcal{G}.x = \{g.x,\ g \in \mathcal{G}\}$$



2. **Extra-class variability:** varability between the sets of different image class.

Class separation $\rightarrow$ reduce intra-class variability and preserve extra-class variability

Scattering Transform



Scattering transform (Mallat, 2012; Bruna and Mallat, 2013)

- $Sx(u)$: multi-scale descriptor of $2^J \times 2^J$ patch.
- Invariant to small geometric deformations:
  $\rightarrow$ **reduces intra-class variability**
- $>$99.5 % accuracy for handwritten digits recognition

Intra-class and extra-class variability in Image Classification

CIFAR-10

- Intra-class variability: small deformations, pose, texture, background...
- Scattering Transform: 82 % accuracy.

ImageNet

- Small deformations: intra- and extra- class variability
- Scattering Transform: 42 % accuracy.

Hybrid CNN architecture

Oyallon et al. (2018) :

- Incorporate geometric invariance properties: Scattering Transform
- Learn the other sources of variability: convolutional neural network
- Non-local method
- 80 % accuracy on ImageNet

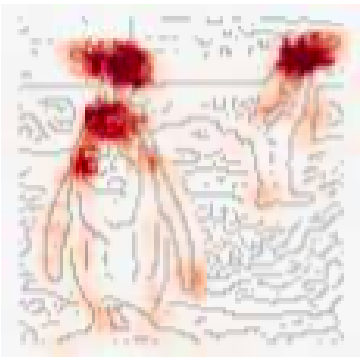
Local Convolutional Neural Network

BagNet (Brendel and Bethge, 2019):

- Local method

$$F(x) = \sum_{p \in x} f(p)$$

- f is a convolutional neural network
- Accuracy: 88 % on ImageNet.
- Explainability of the classification decision: patch evidence

majority of the patches are *filtered* $\rightarrow$ reduces intra-class variability

Hybrid Local Convolutional Neural Network

- Is locality a good hypothesis to reduce intra-class variability ?
- Can we incorporate apriori gemoetric invariance in a hybrid CNN architecture?
- Do we need to learn the spatial component of the filters ?
- What are the class separation mechanisms in such a hybrid architecture?

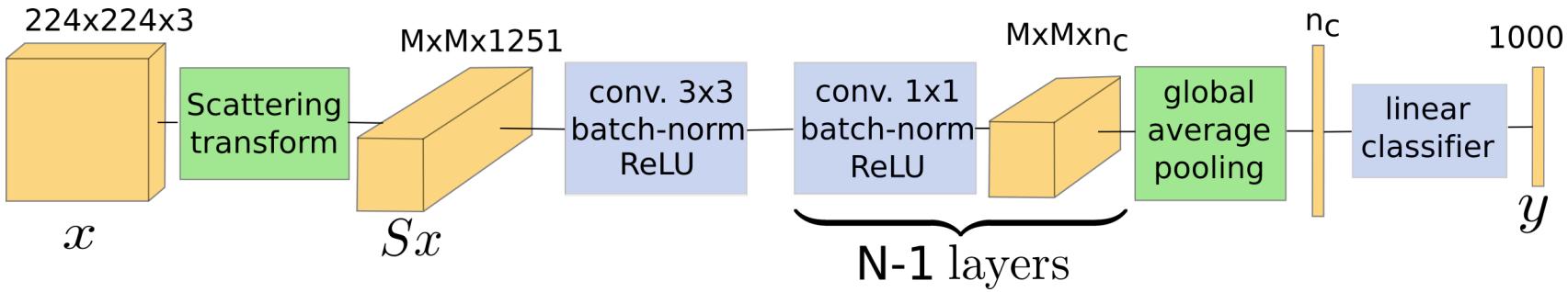
BagNet Scattering

Principles

1. Scattering transform: reduce geometric variability
2. Learn a Non-linear local encoding: reduce intra-class variability while preserving extra-class variability.
3. Global-spatial average: local method hypothesis, reduces variability.
4. Linear classification decision.

BagNet Scattering Algorithm

- Scattering Transform, $J = 4$ scales, oversampling: encoding of 16×16 patches, with 8 overlapping pixels.
- Concatenation of 3×3 neighboring descriptors: 32×32 patches, with 16 overlapping pixels
- Local encoding: sequence of N 1×1 convolutions, batch-norm, ReLU non-linearity
- Global average pooling
- Linear classifier



3×3 descriptor concatenation and first 1×1 convolution $\rightarrow$ implemented in 3×3 convolution.

ImageNet classification results

	Fisher Vectors	Alex Net	BagNet 17	BagNet 33	Scatt. + linear	**Scatt. + non-lin. enc.**
CNN	×	✓	✓	✓	✓	✓
local	✓	×	✓	✓	✓	✓
patch	24^2	-	17^2	33^2	16^2	32^2
depth	-	8	50	50	2	10
Top5	74.3	79.1	81.2	87.0	41.6	84.5

Table: Comparison with other methods

Ablation study

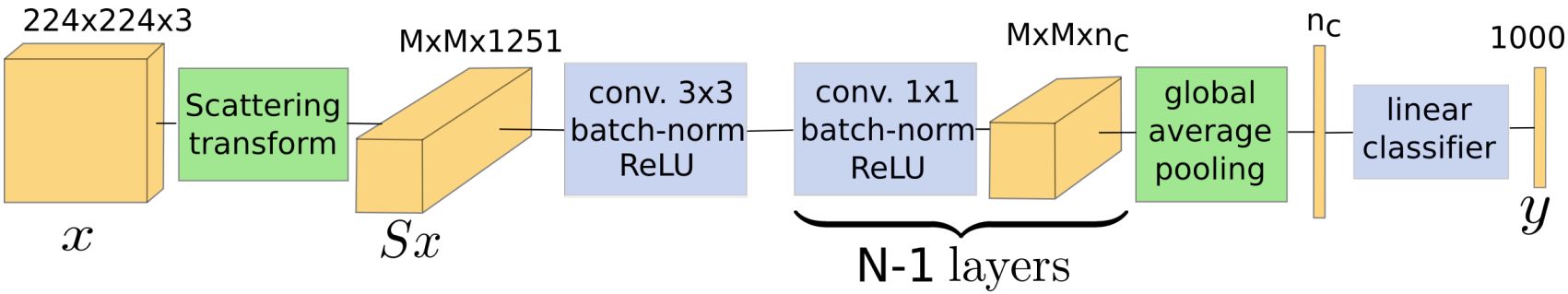


Figure: Original pipeline, accuracy 84.5%

1. Removing the concatenation of Scattering Vector: 78.8%. **Competitive accuracy without learning filters spatial component**.
2. Reducing encoding's channels number: 80.5%.
3. Reducing encoding's depth: 79.2%.

Understand the non-linear encoding

Mathematical operation implemented in sequence of 1×1 convolutions ?

- ℓ^1 sparse coding hypothesis by Zarka, Thiry, Angles, and Mallat (2019).
- Tight-frame contractions: Zarka, Guth, and Mallat (2020)
- Phase collapse: Zarka, Guth, and Mallat.

Patch K-nearest-neighbors

Motivations

- Competitive local methods.
- 16×16 patch, $D = 768 \rightarrow$ **Still high dimension**.

- Does the local hypothesis allow to reduce the intra-class variability ?
- Are there low-dimensional properties of natural image patches ?

- What is the performance of a patch K-nearest-neighbor-based classifier?

- How does it compare with predefined invariance-based representations like Scattering Transform?

The unreasonable effectiveness of patches in Convolutional Kernel Methods, (Thiry et al., 2021).

Naive K-nearest-neighbors

1. Image Level: 40 % accuracy on CIFAR-10

2. Patch Level:

$$F(x) = \sum_{p \in x} \sum_{n \in \mathrm{KNN}(p)} 1_{\mathrm{class(n)}}$$

- Performs poorly: $\sim$ 30% with CIFAR-10 subset.
- Heavy nearest-neighbor search (millions of patches)
- Does not ignore non-informative patches

Informative patches in BagNet (Brendel and Bethge, 2019)

Our Patch K-nearest-neighbors

Goals:

- As close as possible of nearest neighbor classifier
- Reduce the nearest-neighbor search computational cost
- Filter non-informative patches

$\rightarrow$ Learn the class evidence w_n of the patches:

$$F(x) = \sum_{p \in x} \left(\sum_{n \in \mathrm{KNN}(p)} w_n \right)$$

Our Patch K-nearest-neighbors

Algorithm

- Select N patches of size P^2 randomly in the training set
- Mahanalobis Euclidean distance: patches whitening operation
- Patches nearest-neighbors one-hot encoding spatial map

$$\Phi(x) = \left(1_{\mathrm{KNN}(p[i,j])}\right)_{i,j}$$

- linear regression

$$F(x) = \langle W, \Phi(x) \rangle = \sum_{p[i,j] \in x} \left(\sum_{n \in \mathrm{KNN}(p[i,j])} w_n^{i,j} \right)$$

Our Patch K-nearest-neighbors

- High-dimensional embedding

$$\Phi(x) = \left(1_{\mathrm{KNN}(p[i,j])}\right)_{i,j}$$

- Finite dimensional convolutional kernel method

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle$$

- Regularization: low-rank classifier factorization

$$W = W_1 W_2$$

Classification results

CIFAR-10 linear classification

Method	N patches	P	Acc.
Scattering (Oyallon et al. 2015)	-	8	82.2
SimplePatch ℓ^2 (Ours)	$10k$	6	65.4
SimplePatch Mahanalobis (Ours)	$10k$	6	85.6
SimplePatch Mahanalobis (Ours)	$60k$	6	**86.9**

- Mahanalobis distance is key aspect
- Surprisingly good accuracy

ImageNet linear classification

Method	N patches	P	Res.	Top5
Scattering	-	16	224	42.3
Ours	$2k$	12	128	**57.6**

Classification results

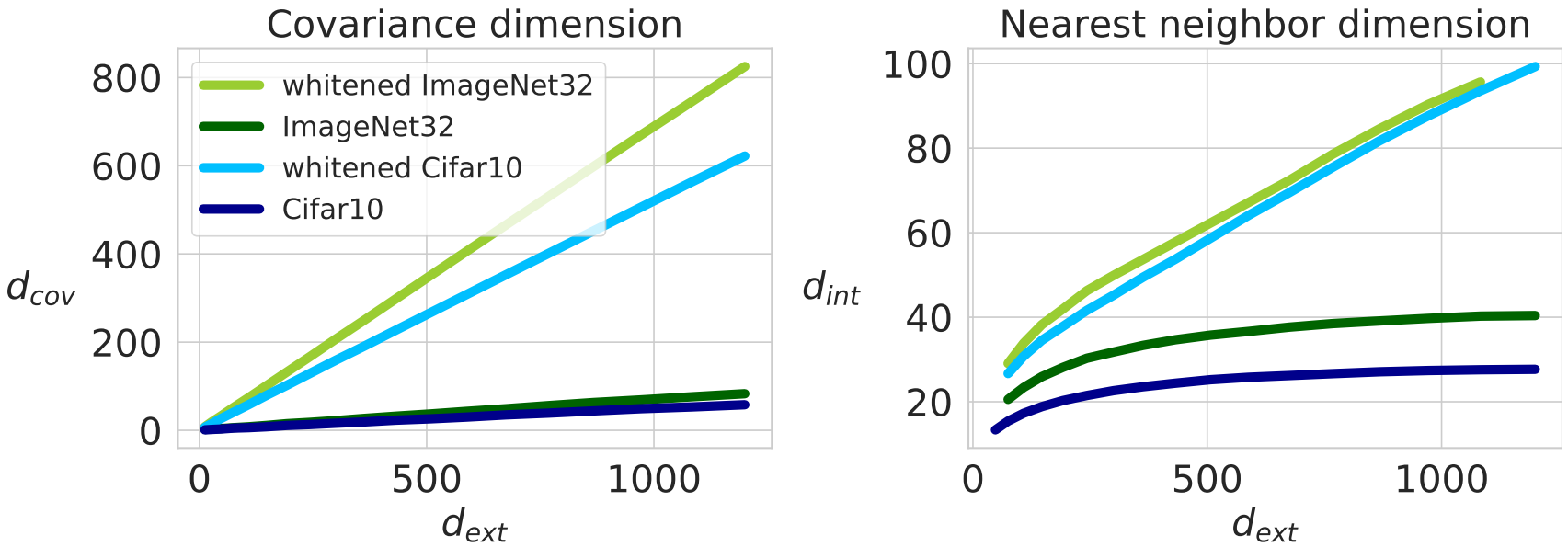
CIFAR-10 Convolutional Kernel Classification

Method	Classifier	Acc.
SimplePatch (Ours)	linear	86.9
SimplePatch (Ours)	1-hidden-layer	88.5
NKWT (Li et al. 2019)	kernel	89.1
NK (Shankar et al. 2020)	kernel	89.8
CKN (Mairal et al. 2016)	kernel	89.8

- Competitive accuracy with convolutional kernel methods
- Possible line of explanation of their success

Low-dimensionality anaylsis

"Scenario where high-dimensional nearest neighbors are meaningful occurs when the underlying dimensionality of the data is much lower than the actual dimensionality.", (Beyer et al., 1999).



Dimensionality measures w.r.t. patch extrinsic dimensionality d_{ext}

Our Patch K-nearest-neighbors

- Locality hypothesis improves significantly nearest neighbors classifier
- State-of-the-art performance as **non-learned** (i.e. non-optimized) representation
- Competitive Convolutional Kernel method
- Very small patch subsets:
  - $60,000$ out of 35 millions CIFAR-10 patches
  - $2,000$ out of 10 billions ImageNet patches
- Patches low-dimensional properties

Images generated with A.I. algorithm in the Art Market

Mario Klingemann

Follow

Imposture Series - The Butcher's Son, 2017
Giclée printing with long-lasting mineral pigments on cotton paper
Hahnemühle Museum Etching 350 gms.
29 9/10 × 19 7/10 in
76 × 50 cm
Edition 0/1 + 1AP

This is a unique work.

€6,500

Contact gallery

ONKAOS

Madrid

Certificate of authenticity
This work includes a certificate of authenticity.

(left) Mario Klingelmann's *The Butcher's son*, Lumen prize gold award in 2018 and (right) Obivous' *Edmond de Bellamy*, sold for 432,500 dollars at Christies.

Images generated with A.I. algorithm in the Art Market

Mario Klingemann

Follow

Imposture Series - The Butcher's Son, 2017
Giclée printing with long-lasting mineral pigments on cotton paper
Hahnemühle Museum Etching 350 gms.
29 9/10 × 19 7/10 in
76 × 50 cm
Edition 0/1 + 1AP

This is a unique work.

€6,500

Contact gallery

ONKAOS

Madrid

Certificate of authenticity
This work includes a certificate of authenticity.

(left) Mario Klingelmann's *The Butcher's son*, Lumen prize gold award in 2018 and (right) Obivous' *Edmond de Bellamy*, sold for 432,500 dollars at Christies.

Is it a prank?

Artification of A.I. art

Art emerges over time as the sum total of institutional activities, everyday interactions, technical implementations, and attributions of meaning.
Roberta Shapiro & Nathalie Heinich, *When is Artification?*, 2012

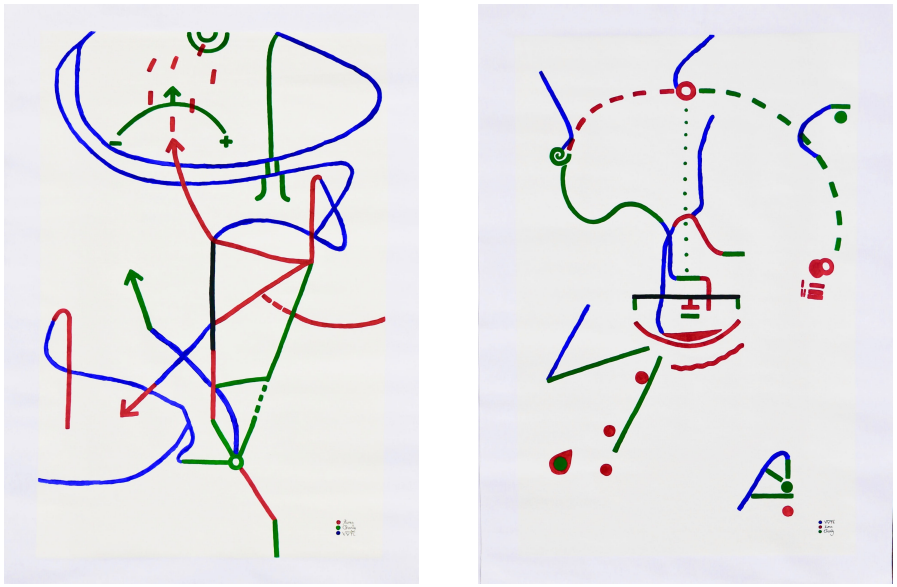
1. Institutional activities: "AI" art exhibitions in major museums (Centre Pompidou, Jeu de Paume), Big Tech companies artists residencies (Google Art and Culture, Nokia Bell Labs)
2. Everyday interactions: press articles, smartphone applications
3. Technical implementations: open-source software, "A.I. artists" have software engineering bakcground.
4. Attributions of meaning: artists narratives.

Can we propose a narrative around the creative interaction rather than centered on the algorithm?

Dialog on a canvas with a machine

Cabannes, Kerdreux, Thiry, Campana, and Ferrandes (2019)

- *Tina&Charly* artist duo.
- Three-way dialogue between Charly (**green**), Tina (**red**) and an algorithm (**blue**).
- Creativity: Human-machine interaction rather than an algorithm solely.

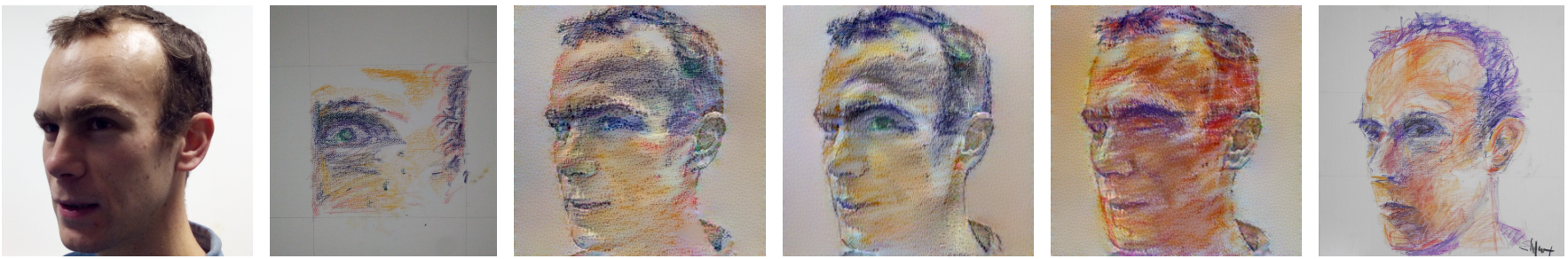


(left)*Actif* and (right) *Passif* from the series *Peinture Algorithmée*.

Neural style transfer with artists

Kerdreux, Thiry, and Kerdreux (2020).

- *Erwan Kerdreux*: Professor at ENS-Paris Saclay design departement.
- Neural Style Transfer (Gatys et al., 2015): artistic style transfer algorithm.
- Interaction of the artist with *its own style*.

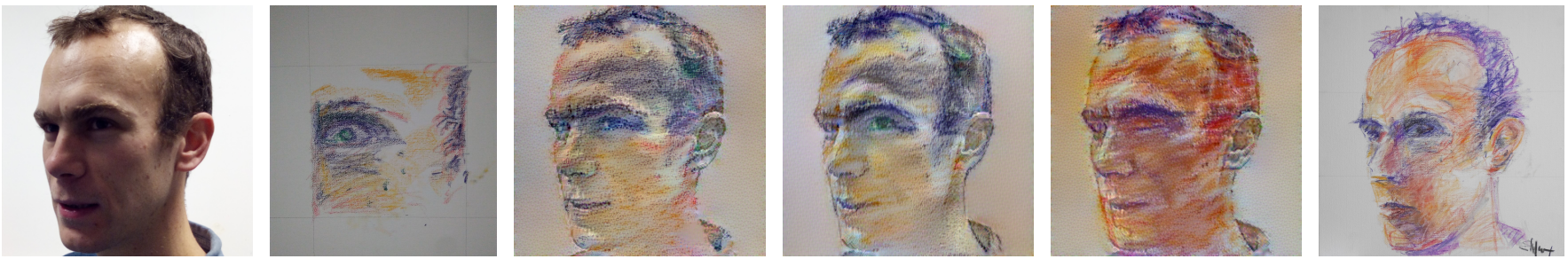


(Left to right) Original photograph, first iteration, first, fifth and last projections, and final canvas

Neural style transfer with artists

Kerdreux, Thiry, and Kerdreux (2020).

- *Erwan Kerdreux*: Professor at ENS-Paris Saclay design departement.
- Neural Style Transfer (Gatys et al., 2015): artistic style transfer algorithm.
- Interaction of the artist with *its own style*.



(Left to right) Original photograph, first iteration, first, fifth and last projections, and final canvas

Testomony of the effects of our daily interactions with increasingly powerful machines

Future perspecitves

- **Solid Harmonic Scattering Transform**: multi-scale invariant descriptor. **Not restricted to atoms, can be used for densities.** Multi-scale exchange-correlation for Density Functional Theory ?
- **Energy regression**: energy is multiscale, but **local components are extremely dominant** in our case-studies.
- **Entropy regression**: allows **free-energy landscape exploration** computationally unfeasible before.
- **Structured CNN architecture**: mathematical analysis of the operations. ℓ^1 sparsity hypothesis (Zarka et al., 2019) is not satisfactory, other hypotheses?
- **Patches K-nearest-neighbor classifier**: rethink high-dimensional learning assumption? Characterize more precisely patches dimensionality ? Refine the Euclidean metric used in KNN ?


Questions ?

References

Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1872–1886, August 2013.

Vivien Cabannes, Thomas Kerdreux, Louis Thiry, Tina Campana, and Charly Ferrandes. Dialog on a canvas with a machine. In *NeurIPS 2019 Workshop on Machine Learning for Creativity and Design*, 2019.

Volker L Deringer and Gábor Csányi. Machine learning based interatomic potential for amorphous carbon. *Physical Review B*, 95(9):094203, 2017.

Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, Stéphane Mallat,