

# Tutorial on Interpreting and Explaining Deep Models in Computer Vision



Wojciech Samek  
(Fraunhofer HHI)



Grégoire Montavon  
(TU Berlin)



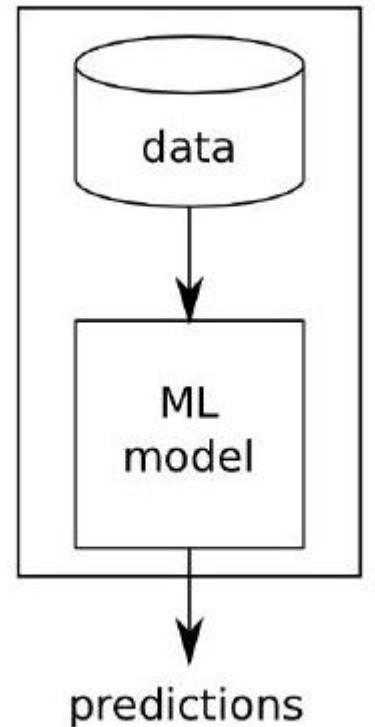
Klaus-Robert Müller  
(TU Berlin)

08:30 - 09:15      Introduction KRM  
09:15 - 10:00      Techniques for Interpretability GM  
10:00 - 10:30      Coffee Break ALL  
10:30 - 11:15      Applications of Interpretability WS  
11:15 - 12:00      Further Applications and Wrap-Up **KRM**



# Is the Generalization Error all we need?

Standard ML



*Generalization error*

# Application: Comparing Classifiers

Test error for various classes:

	aeroplane	bicycle	bird	boat	bottle	bus	car
Fisher	79.08%	66.44%	45.90%	70.88%	27.64%	69.67%	80.96%
DeepNet	88.08%	79.69%	80.77%	77.20%	35.48%	72.71%	86.30%
	cat	chair	cow	diningtable	dog	horse	motorbike
Fisher	59.92%	51.92%	47.60%	58.06%	42.28%	80.45%	69.34%
DeepNet	81.10%	51.04%	61.10%	64.62%	76.17%	81.60%	79.33%
	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Fisher	85.10%	28.62%	49.58%	49.31%	82.71%	54.33%	59.99%
DeepNet	92.43%	49.99%	74.04%	49.48%	87.07%	67.08%	72.12%

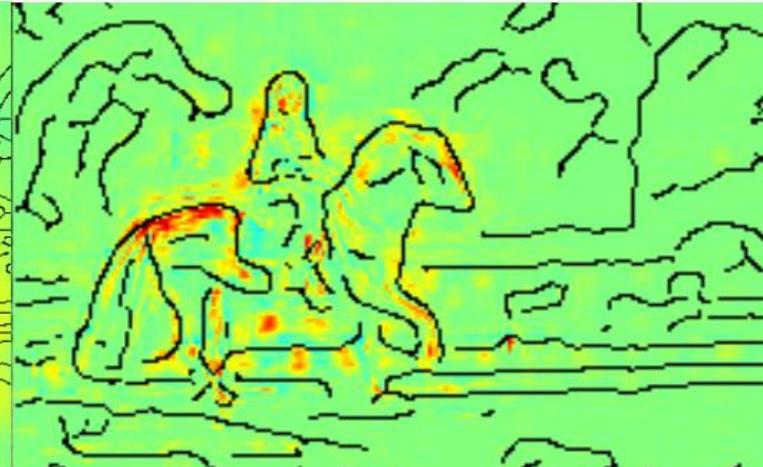
Image



FV



DNN

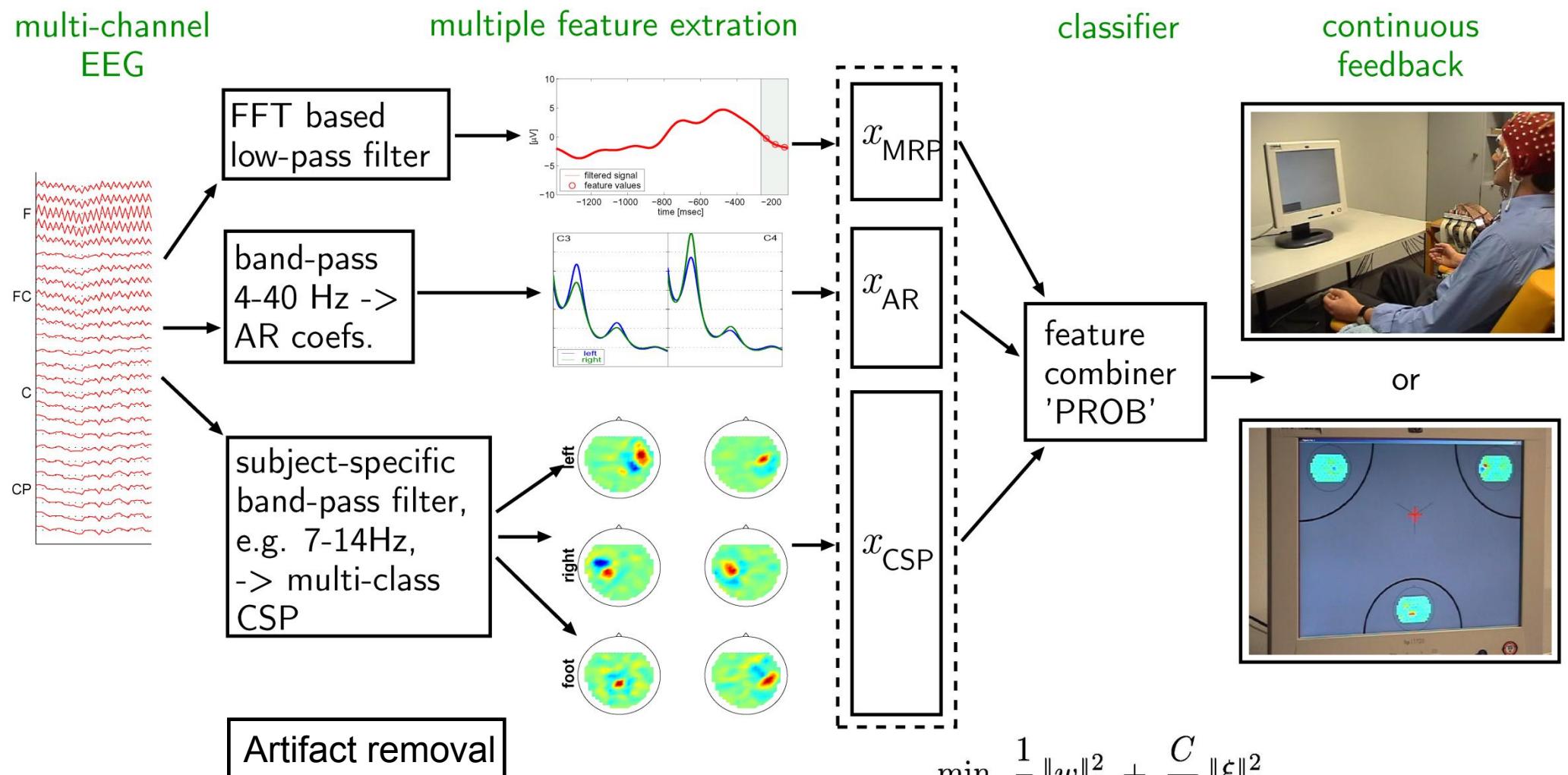


C. Lothar Lenz  
www.pferdefotoarchiv.de

# Machine Learning in the Sciences

# Machine Learning in Neuroscience

# BBCI Set-up: Let the machines learn



$$\min_{w,b,\xi} \frac{1}{2} \|w\|_2^2 + \frac{C}{K} \|\xi\|_2^2$$

$$\text{subject to } y_k(w^\top x_k + b) = 1 - \xi_k \quad \text{for } k = 1, \dots, K$$

[cf. Müller et al. 2001, 2007, 2008, Dornhege et al. 2003, 2007, Blankertz et al. 2004, 2005, 2006, 2007, 2008]

# Brain Computer Interfacing: ,Brain Pong‘



## Berlin Brain Computer Interface

- ML reduces patient training from 300h -> 5min

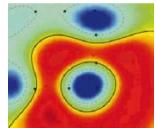
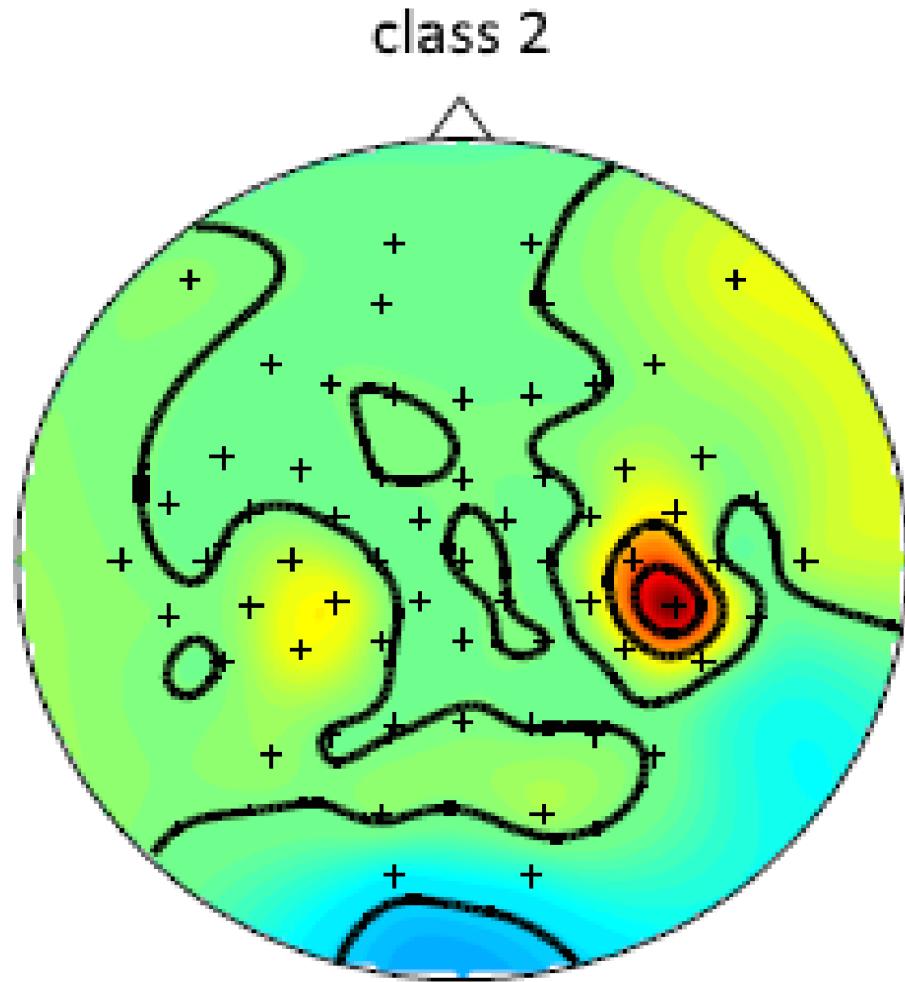
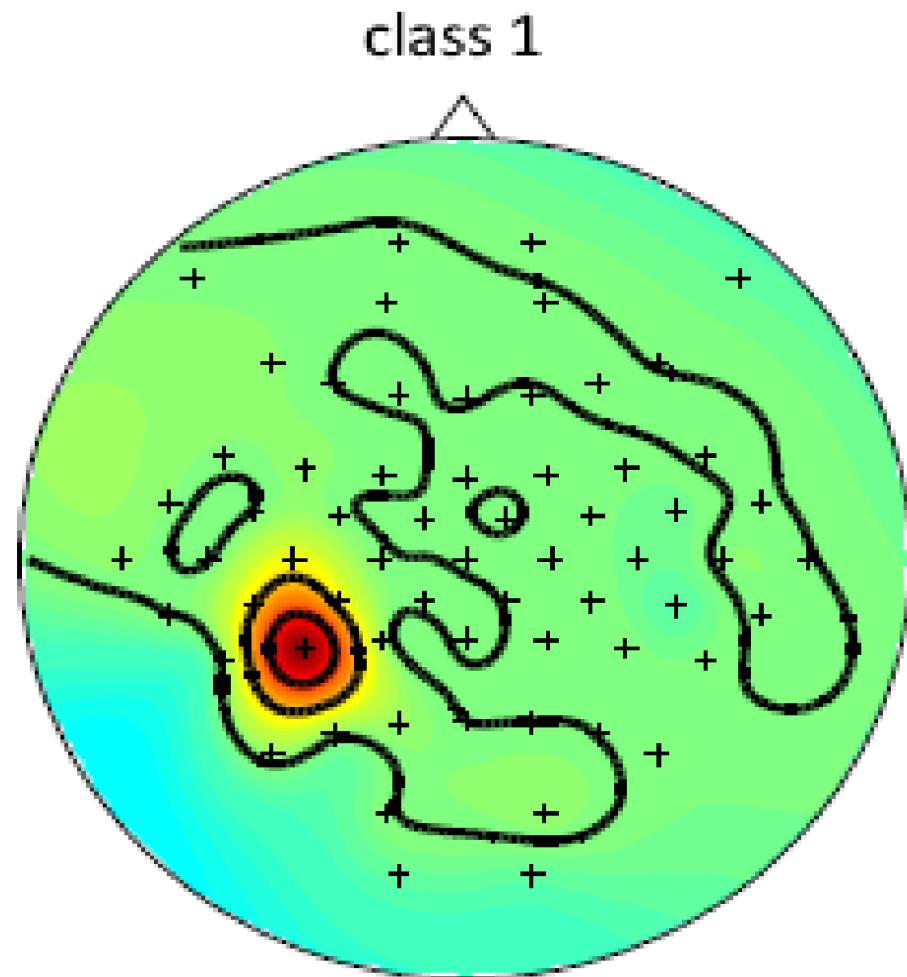
## Applications

- help/hope for patients (ALS, stroke...)
- neuroscience
- neurotechnology (video coding, gaming, monitoring driving)

**Leitmotiv: »let the machines learn«**

# DNN Explanation Motor Imagery BCI

---



**Note:** Explanation available for single Trial (Sturm et al 2016)

# Explaining in Physics

# Machine Learning in Chemistry, Physics and Materials

Matthias Rupp, Anatole von Lilienfeld,  
Alexandre Tkatchenko, Klaus-Robert Müller

# Machine Learning for chemical compound space

---

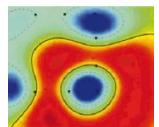
Ansatz:

$$\{Z_I, \mathbf{R}_I\} \xrightarrow{\text{ML}} E$$

instead of

$$\hat{H}(\{Z_I, \mathbf{R}_I\}) \xrightarrow{\Psi} E$$

$$\hat{H}\Psi = E\Psi$$



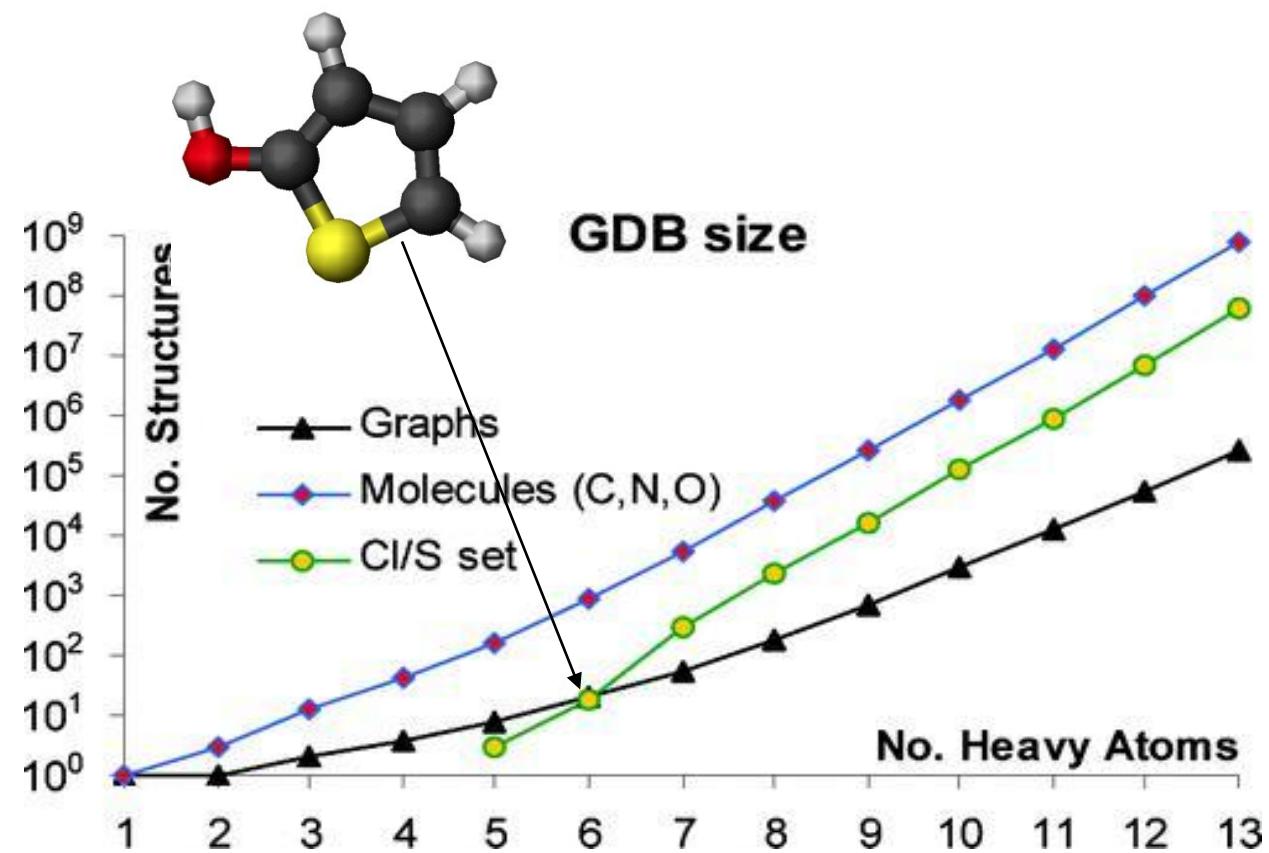
[from von Lilienfeld]

# The data

GDB-13 database of all organic molecules (within stability & synthetic constraints) of 13 heavy atoms or less: 0.9B compounds

Table 1. Structure Generation Statistics for GDB-13

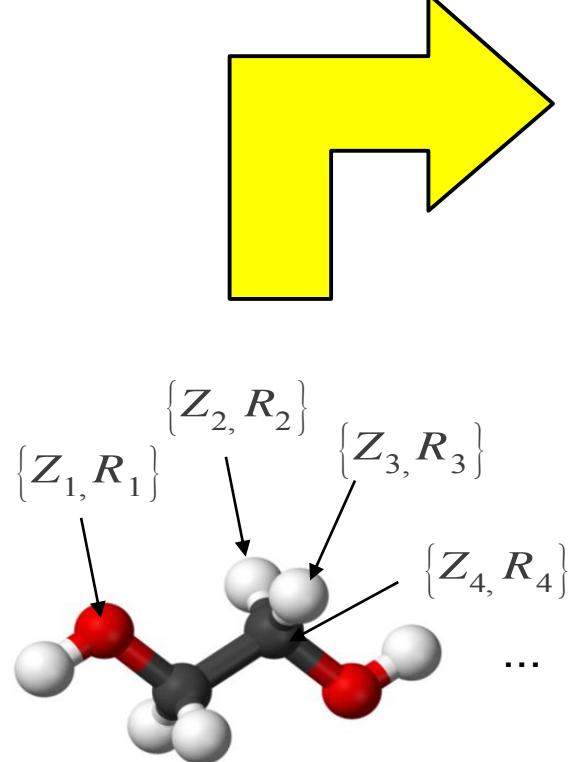
nodes <sup>a</sup>	graphs <sup>b</sup>	GDB <sup>c</sup>	CI/S <sup>d</sup>	CPU time (h) <sup>e</sup>
1	1	1	0	0.00
2	1	3	0	0.00
3	2	12	0	0.00
4	4	43	0	0.00
5	8	155	3	0.01
6	20	934	19	0.02
7	57	5 726	315	0.05
8	194	37 151	2 438	0.33
9	706	255 542	17 056	2.68
10	2 831	1 784 626	130 465	25.26
11	12 011	12 961 686	938 704	223.49
12	53 789	99 821 343	724 0108	3 023.79
13	250 268	795 244 451	59 027 533	36 606.45
Total	319 892	910 111 673	67 356 641	39 882.08



Blum & Reymond, JACS (2009)

[from von Lilienfeld]

# Coulomb representation of molecules

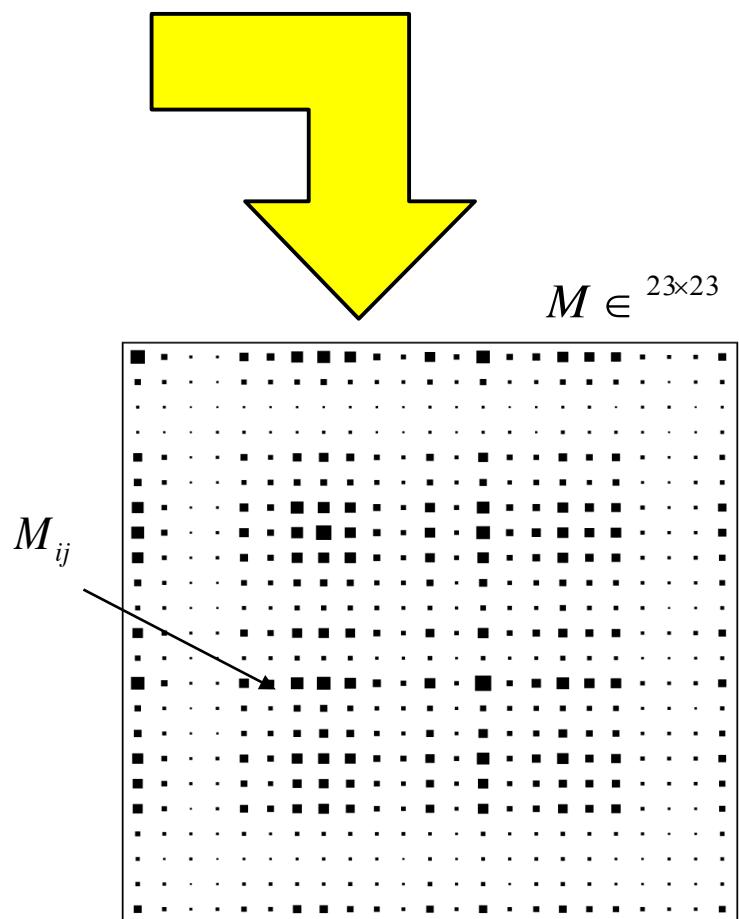


+ phantom atoms

$$\{0, R_{21}\} \quad \{0, R_{22}\} \quad \{0, R_{23}\}$$

Two yellow arrows point from the molecular structure to a light purple rectangular box containing the Coulomb matrix equations. The first arrow points to the diagonal element equation, and the second arrow points to the off-diagonal element equation.

$$M_{ii} = Z_i^{2.4}$$
$$M_{ij} = \frac{Z_i Z_j}{\|R_i - R_j\|}$$



Coulomb Matrix (Rupp, Müller et al 2012, PRL)

$$d(\mathbf{M}, \mathbf{M}') = \sqrt{\sum_{IJ} |M_{IJ} - M'_{IJ}|^2}$$

# Kernel ridge regression

Distances between  $\mathbf{M}$  define Gaussian kernel matrix  $\mathbf{K}$

$$k(\mathbf{M}, \mathbf{M}') = \exp\left(-\frac{d(\mathbf{M}, \mathbf{M}')^2}{2\sigma^2}\right)$$

Predict energy as sum over weighted Gaussians

$$E^{est}(\mathbf{M}) = \sum_i \alpha_i k(\mathbf{M}, \mathbf{M}_i) + b$$

using weights that minimize error in training set

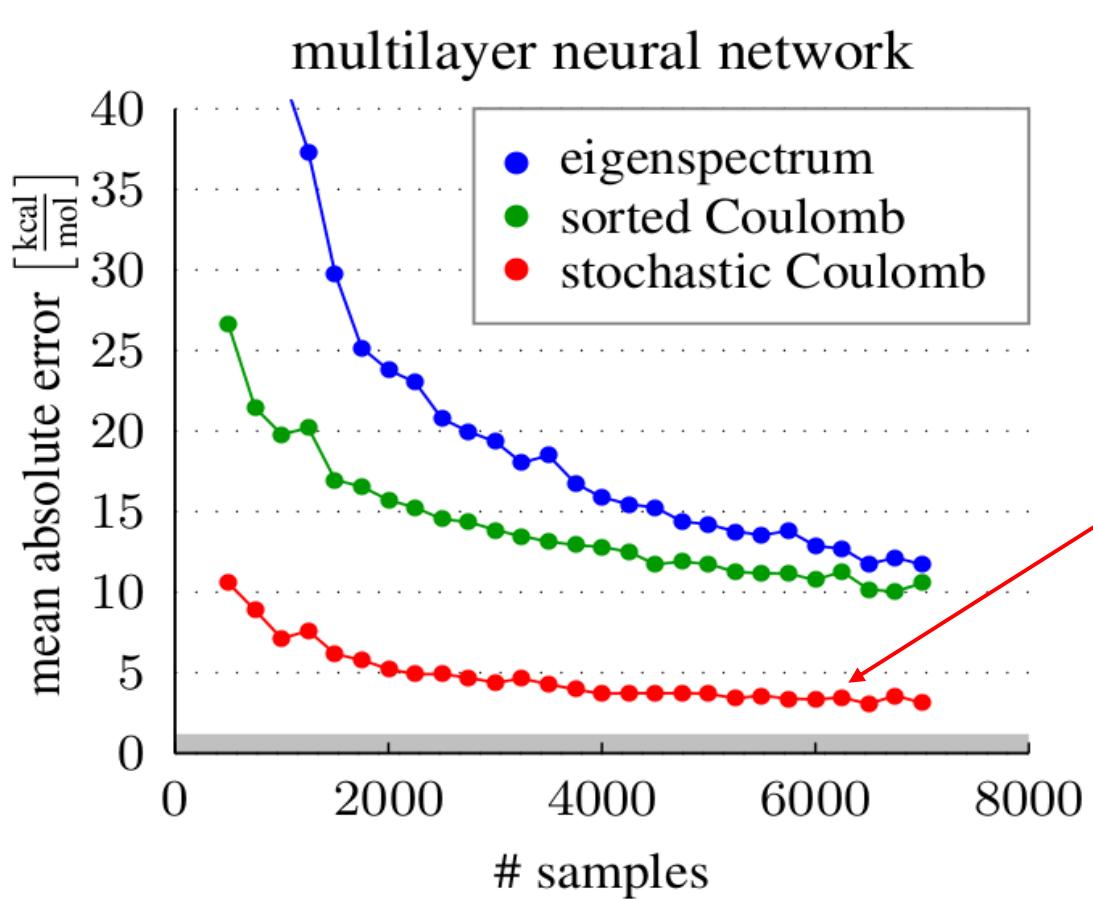
$$\min_{\alpha} \quad \sum_i (E^{est}(\mathbf{M}_i) - E_i^{ref})^2 + \lambda \sum_i \alpha_i^2$$
$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

Exact solution

As many parameters as molecules + 2 global parameters, characteristic length-scale or  $kT$  of system ( $\sigma$ ), and noise-level ( $\lambda$ )

# Predicting Energy of small molecules: Results

---



March 2012

Rupp et al., PRL

**9.99 kcal/mol**

(kernels + eigenspectrum)

December 2012

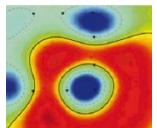
Montavon et al., NIPS

**3.51 kcal/mol**

(Neural nets + Coulomb sets)

2015 Hansen et al 1.3kcal/mol at  
**10 million times faster than the**  
state of the art

Prediction considered chemically  
accurate when MAE is below **1  
kcal/mol**



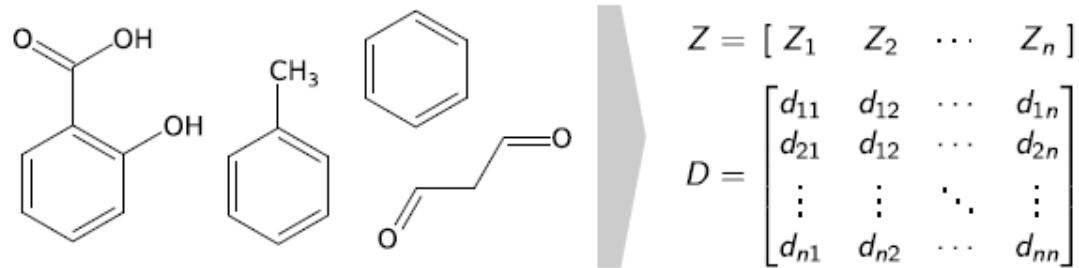
Dataset available at <http://quantum-machine.org>

# Learning Atomistic Representations with Deep Tensor Neural Networks

Kristof Schütt, Farhad Arbabzadah,  
Stefan Chmiela, Alexandre Tkatchenko,  
Klaus-Robert Müller

# Deep Tensor Neural Network (DTNN) for representing molecules

**Input:** Atomic numbers and interatomic distances



Embedding of based on atom types

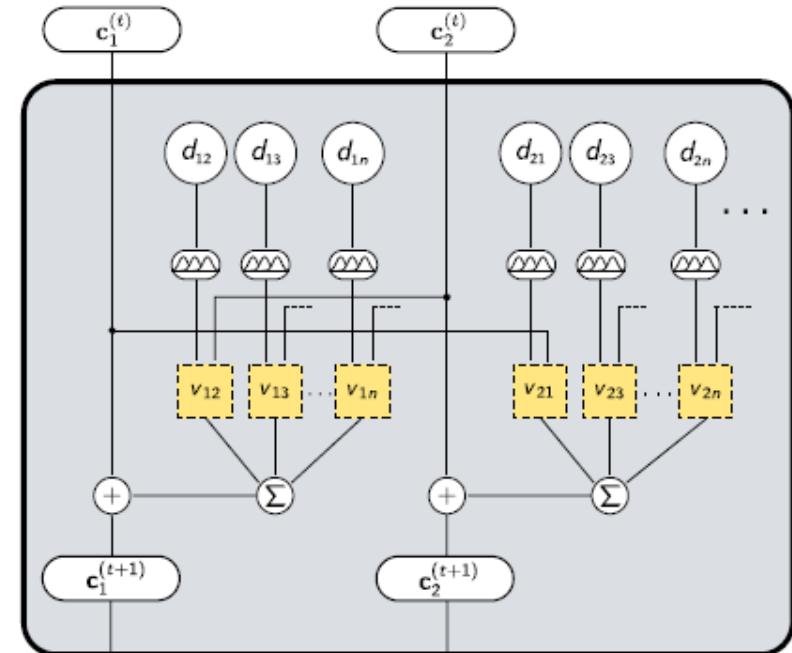
$$\mathbf{x}_i^{(0)} = \mathbf{x}_{Z_i} \in \mathbb{R}^d$$

Add interaction with environment using  $t = 1 \dots T$   
sequential refinements  $\mathbf{v}_i^{(t)}$

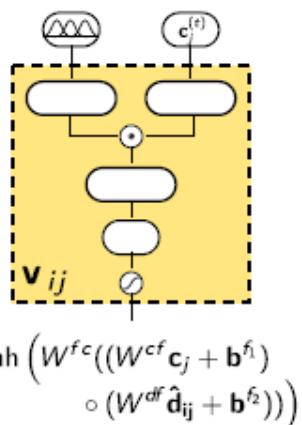
$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t)} \left( \mathbf{x}_1^{(t)}, \dots, \mathbf{x}_{n_{\text{atoms}}}^{(t)}, d_{i1}, \dots, d_{in_{\text{atoms}}} \right)$$

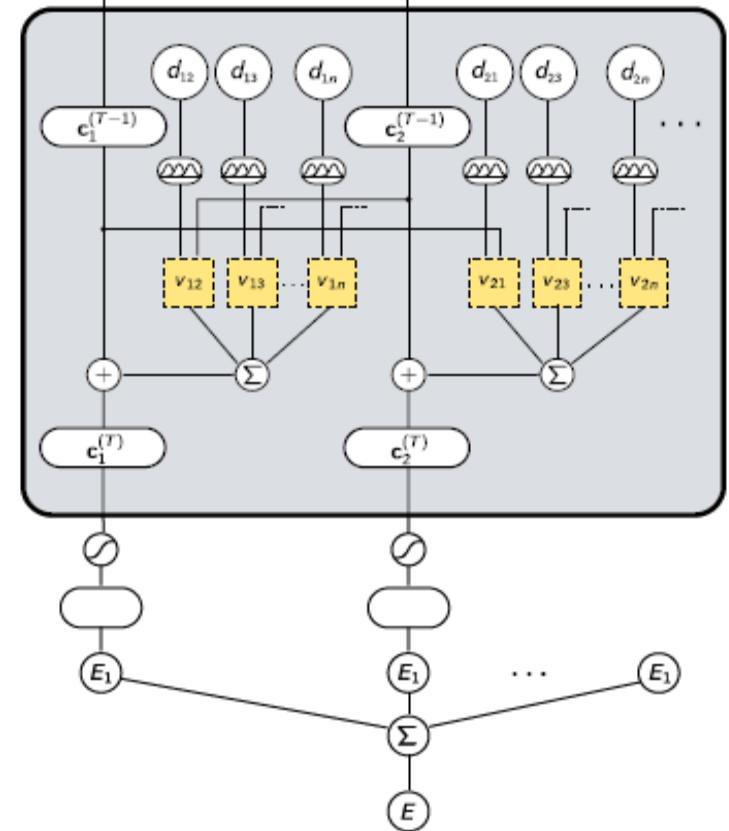
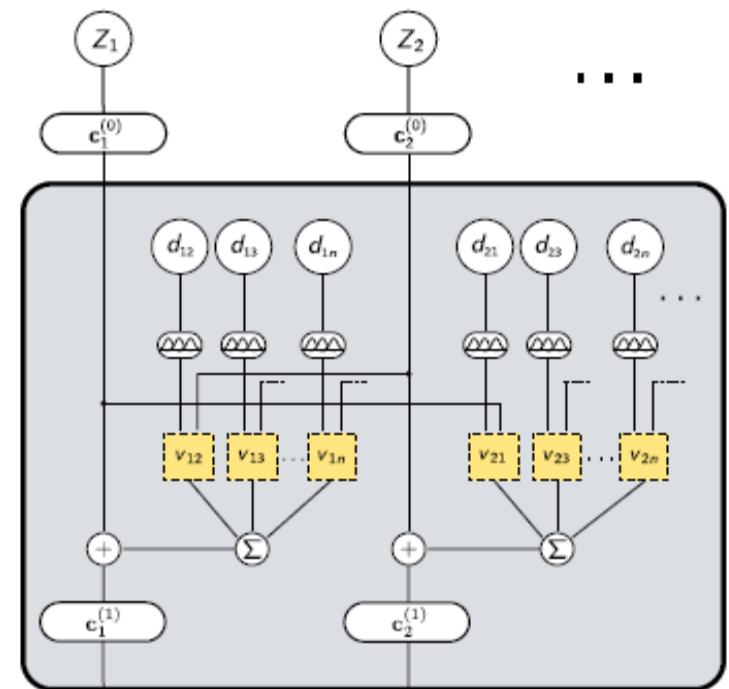
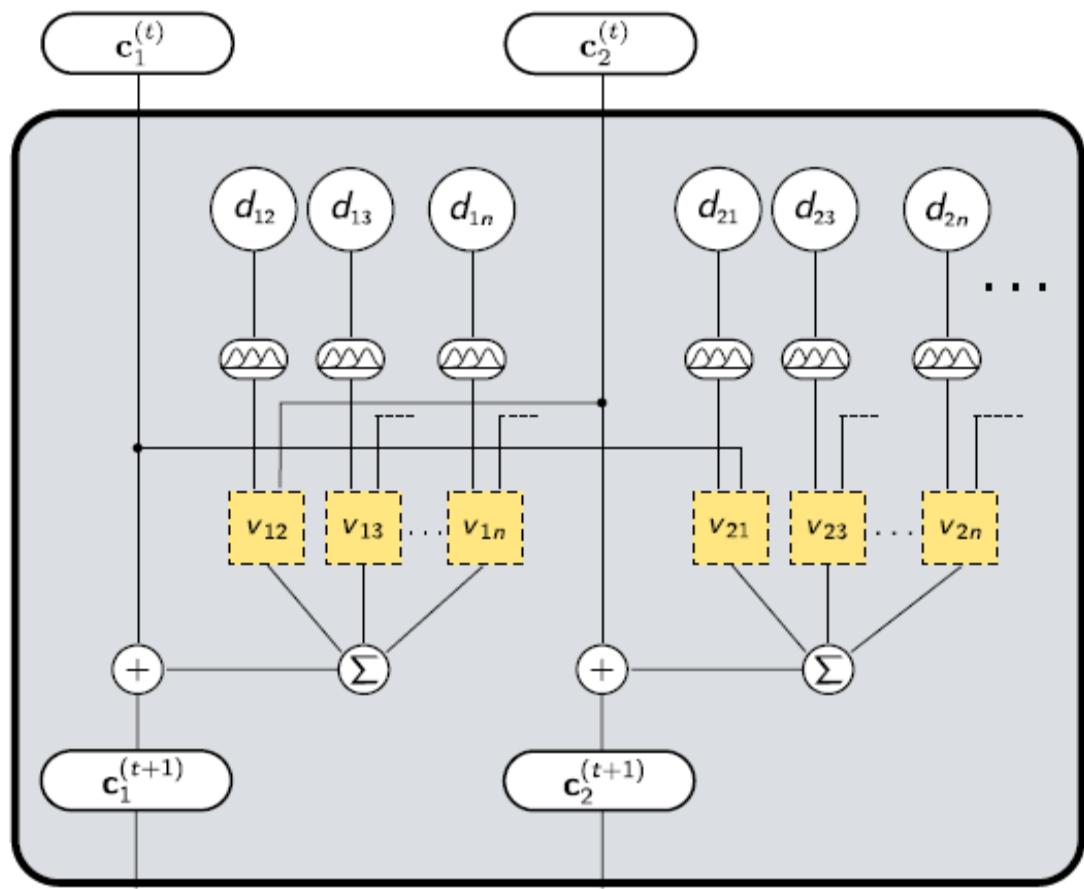
Prediction via atom-wise contributions:

$$\hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(\mathbf{x}_i^{(T)})$$

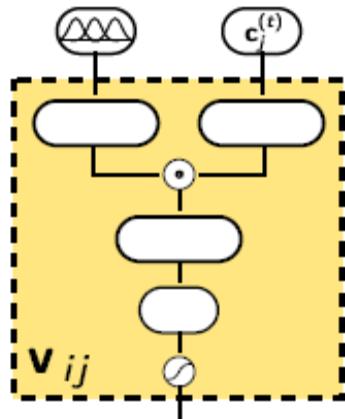


- ( Gaussian expansion)
- ( hyperbolic tangent)
- ( element-wise product)
- ( element-wise sum)





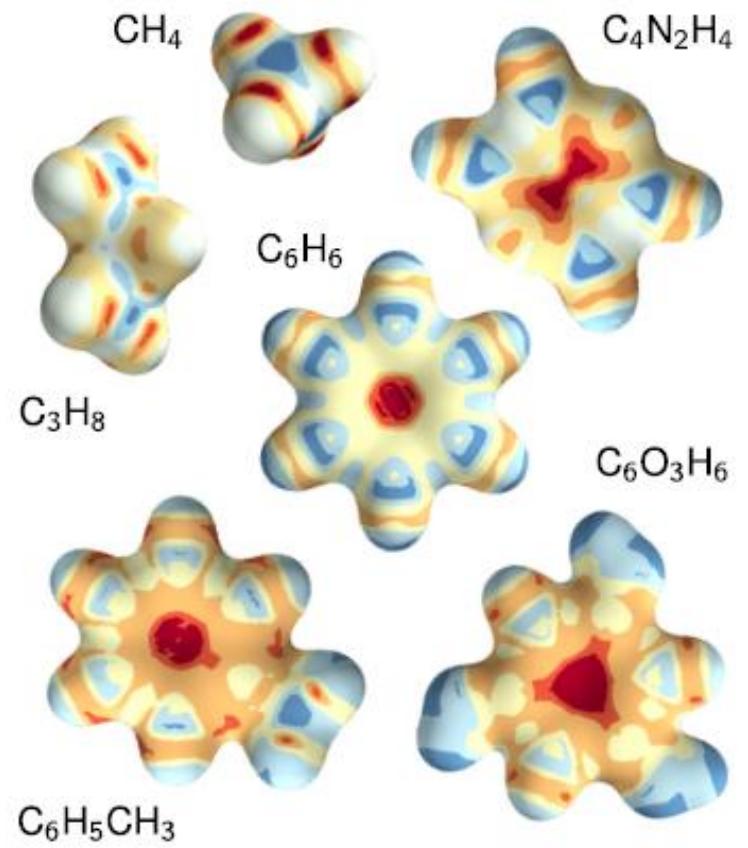
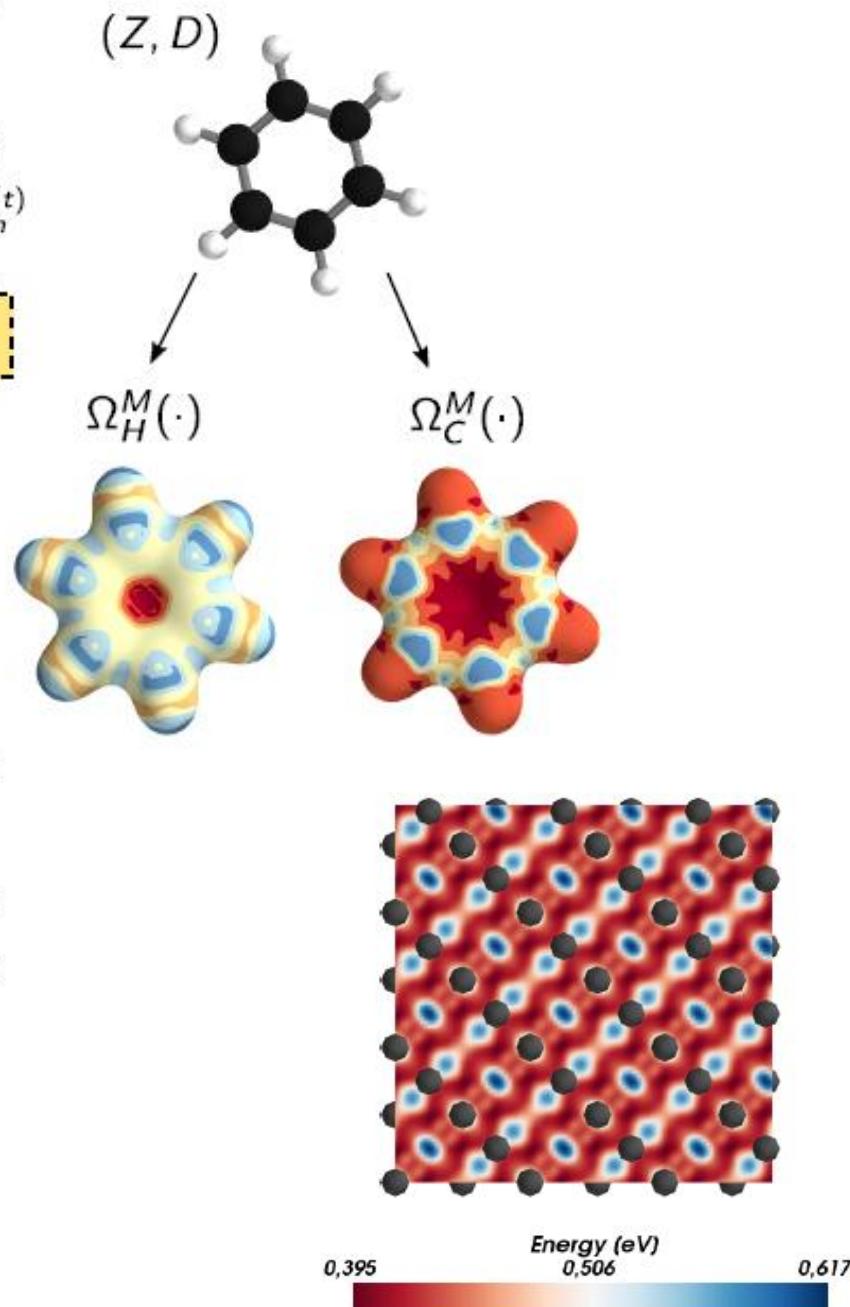
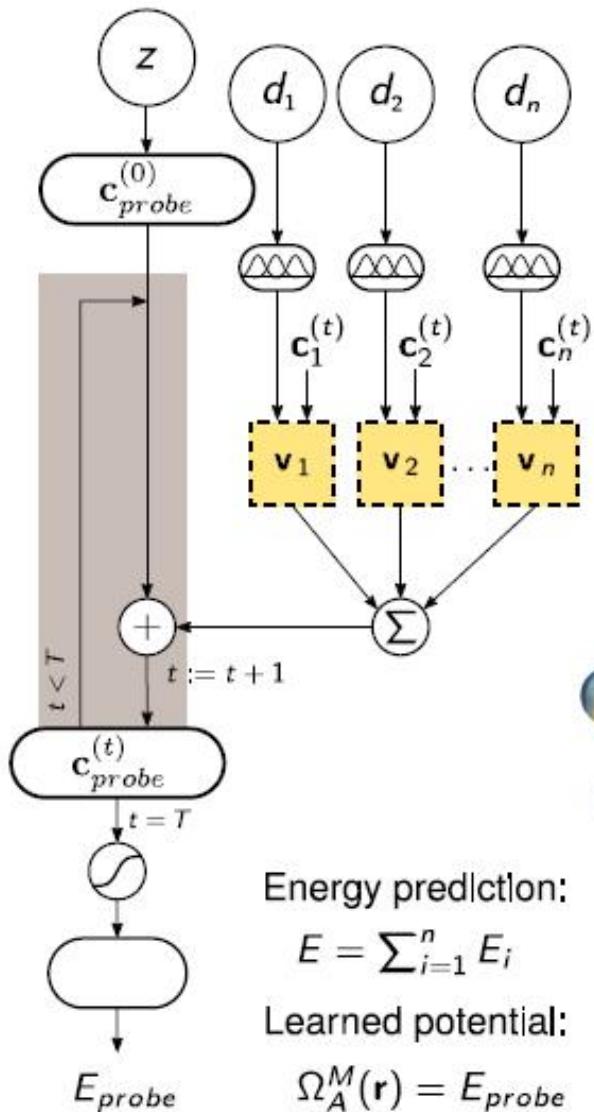
- Gaussian expansion
- hyperbolic tangent
- element-wise product
- element-wise sum



$$\tanh \left( W^{fc} ((W^{cf} \mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df} \hat{\mathbf{d}}_{ij} + \mathbf{b}^{f_2})) \right)$$

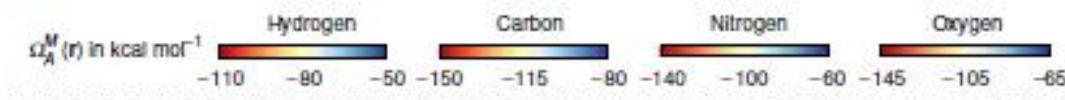
# Gaining insights for Physics

# Toward Quantum Chemical Insights: supervised

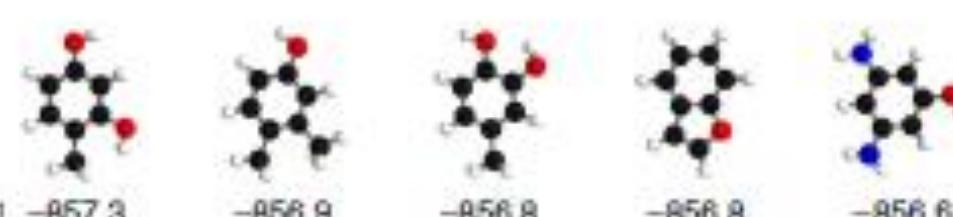
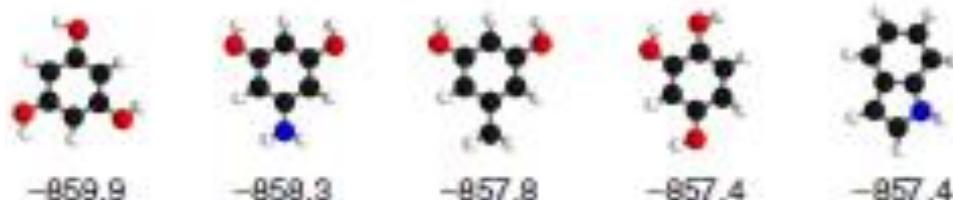
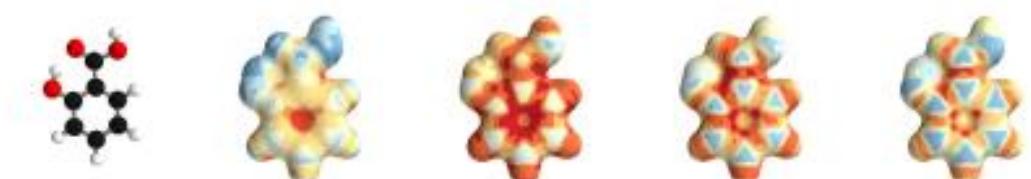


[Schütt et al. Nat Comm. 2017,  
 Schütt et al JCP 2018]

# Quantum chemical insights

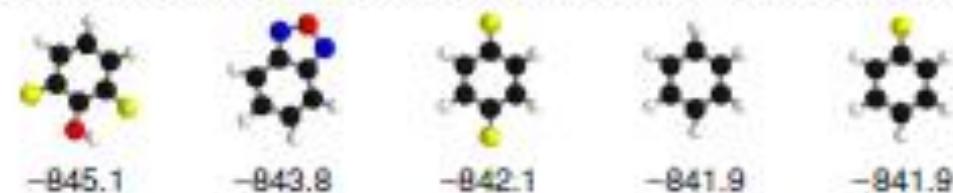


Inferred chemical potentials

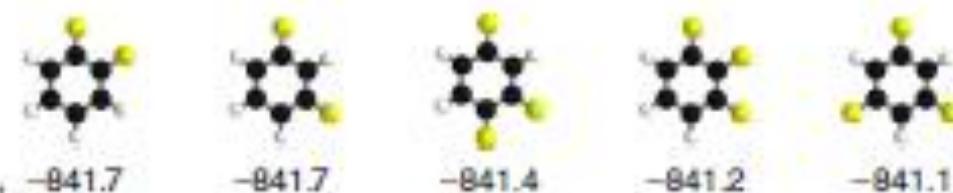


Inferred stable and unstable  
carbon ring classification  
**‘aromaticity’**

# 281 – 290



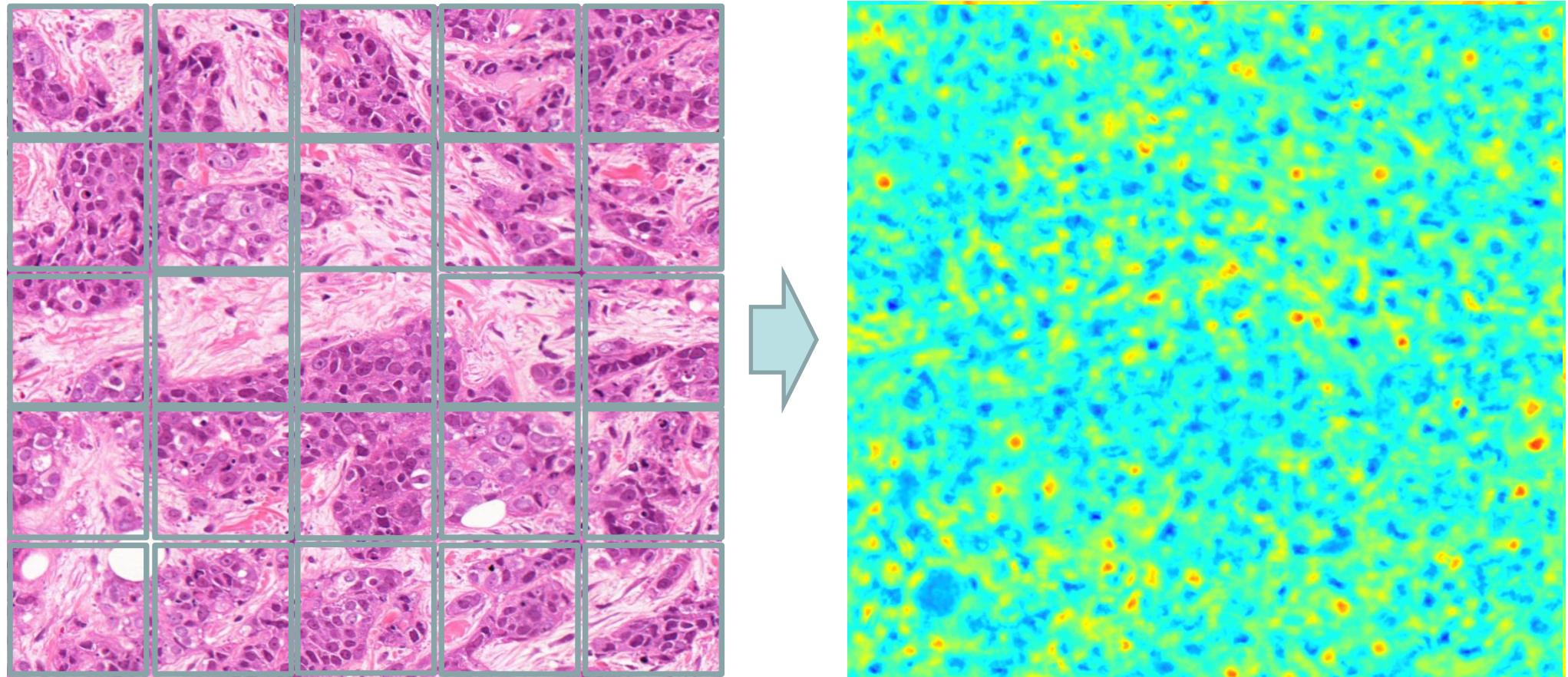
$E_{\text{ring}}$  in kcal mol<sup>-1</sup>



# Machine Learning for morpho-molecular Integration

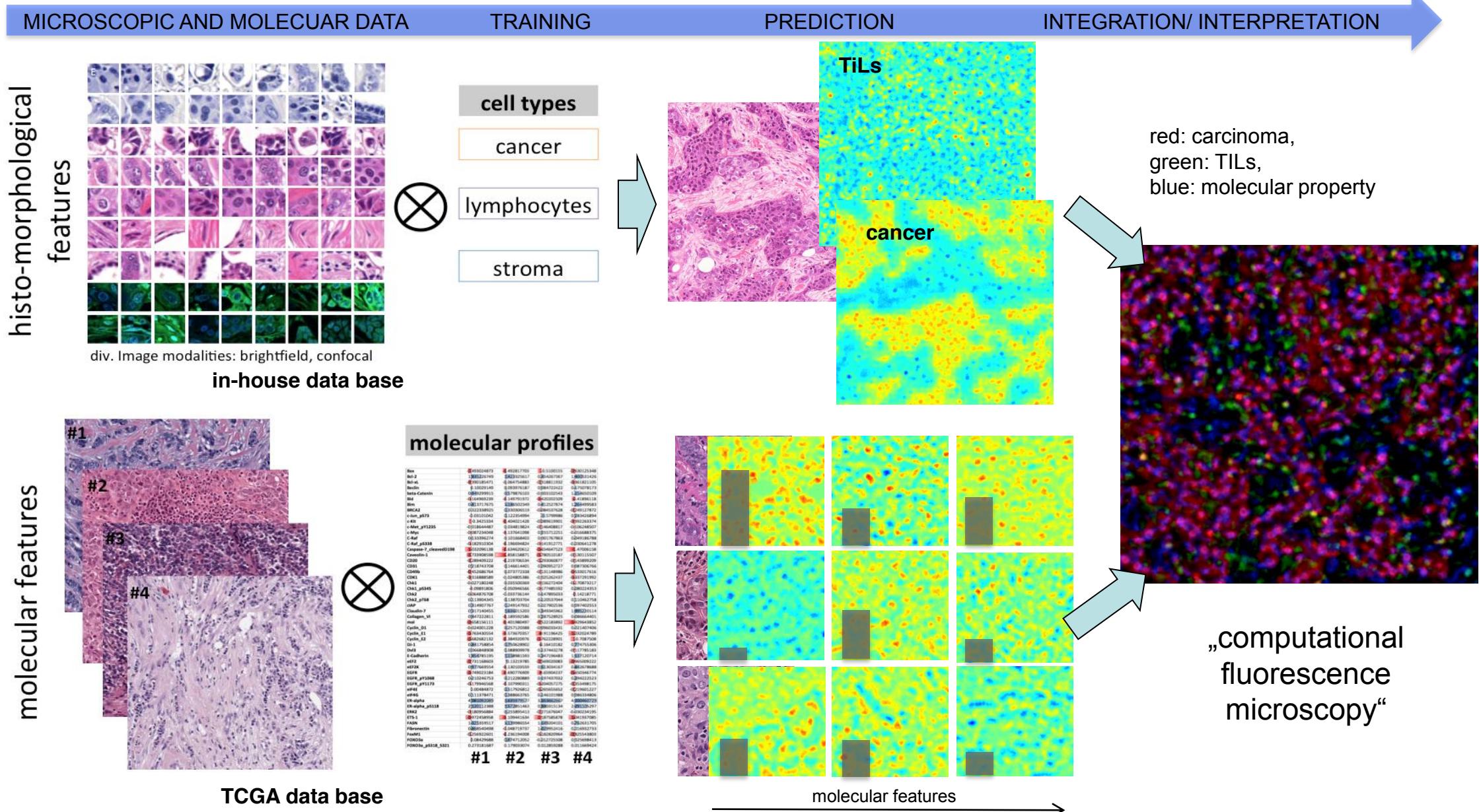
Alexander Binder<sup>1,6</sup>, Michael Bockmayr<sup>2,10</sup>, Miriam Hägele<sup>1</sup>, Stephan Wienert<sup>2</sup>, Daniel Heim<sup>2</sup>, Katharina Hellweg<sup>3</sup>, Albrecht Stenzinger<sup>4</sup>, Laura Parlow<sup>2</sup>, Jan Budczies<sup>2</sup>, Benjamin Goeppert<sup>4</sup>, Denise Treue<sup>2</sup>, Manato Kotani<sup>5</sup>, Masaru Ishii<sup>5</sup>, Manfred Dietel<sup>2</sup>, Andreas Hocke<sup>3</sup>, Carsten Denkert<sup>2,7</sup>, Klaus-Robert Müller<sup>1,8,9,\*</sup> and Frederick Klauschen<sup>2,7,\*</sup>

# Interpretable ML



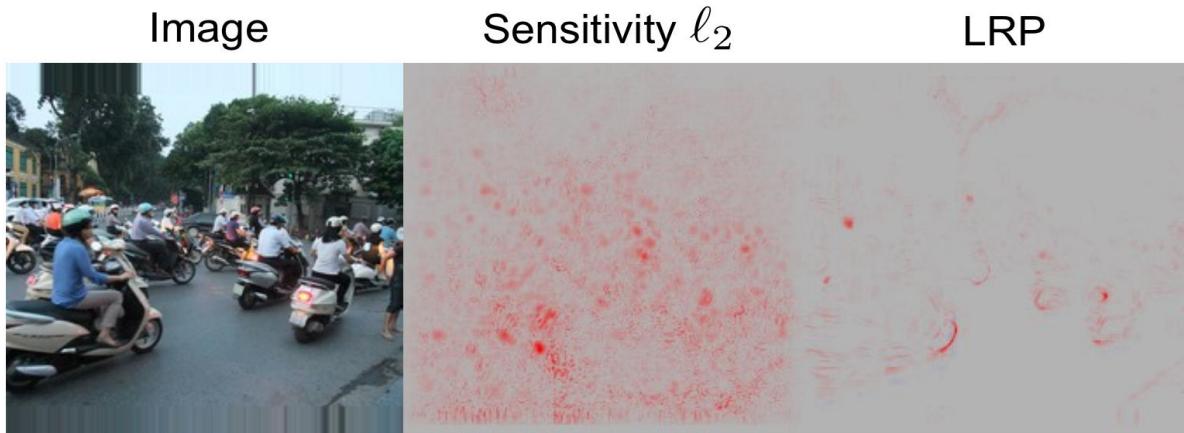
Bach et al., PLoS1 2015  
Klauschen et al., US Patent #9558550  
Binder et al., *in revision*

# Machine learning based integration of morphological and molecular tumor profiles



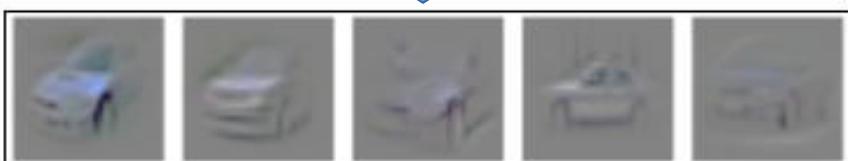
# Take Home messages

# Sensitivity analysis is not the question that you would like to ask!



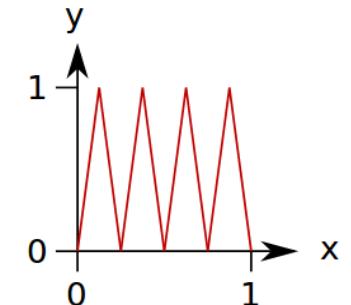
Explanation for simple models does  
not necessary work for deep models

What works for simple models doesn't work for deep models.



gradient-based  
methods

vulnerable to  
shattered  
gradients



Our LRP method is robust to this.

# Layer-Wise Relevance Propagation

Desirable properties  
of an explanation

{ positivity  
conservation  
selectivity  
continuity }

“Tricks of  
the trade”

$$R_i = \sum_j \frac{\partial R_j}{\partial a_i} \cdot (a_i - \tilde{a}_i^{(j)})$$

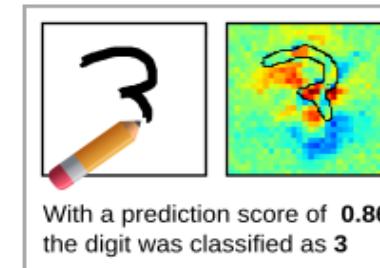
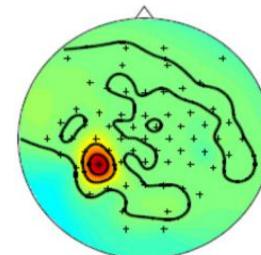
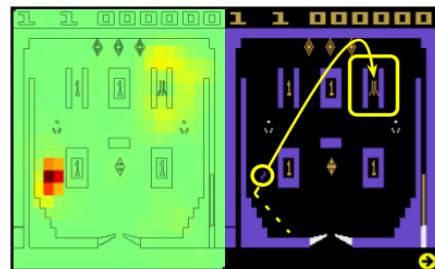
Underlying theory  
for consistency

Deep Taylor  
Decomposition

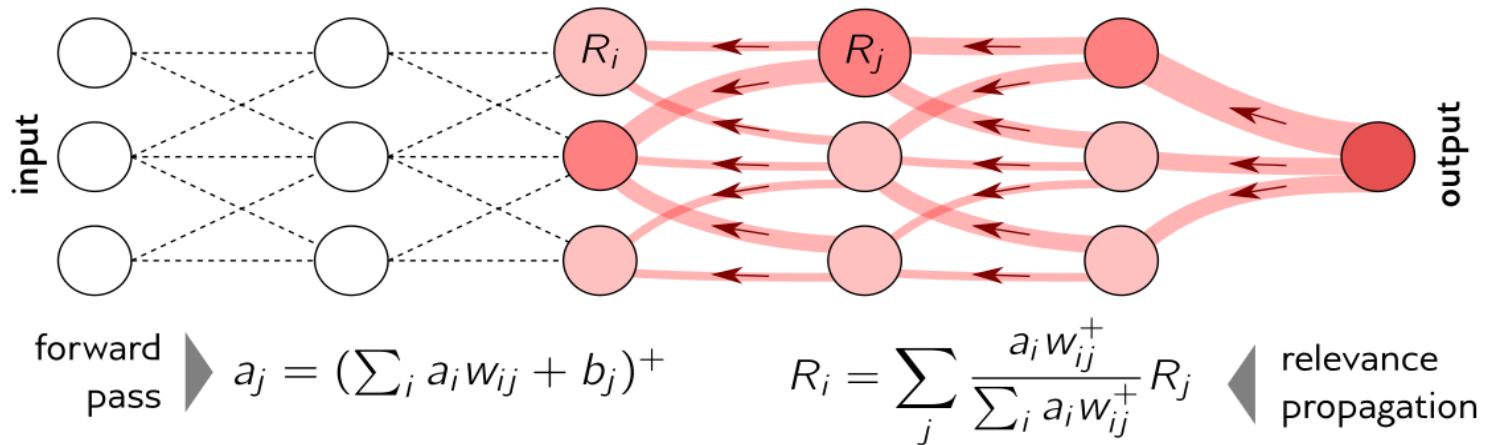
## LRP Explanation Framework

e people are more prone to g  
The mental part is usually  
y is up or down, ie: the Shu  
ointed towards Earth, so the  
astronauts. About 50% of t  
s, and NASA has done numerou

(software, tutorials, demos,  
insights, applications)



# LRP works 4 all: deep models, LSTMs, kernel methods ...



# A Clarification on LRP

$LRP \neq \text{Gradient} \times \text{Input}$

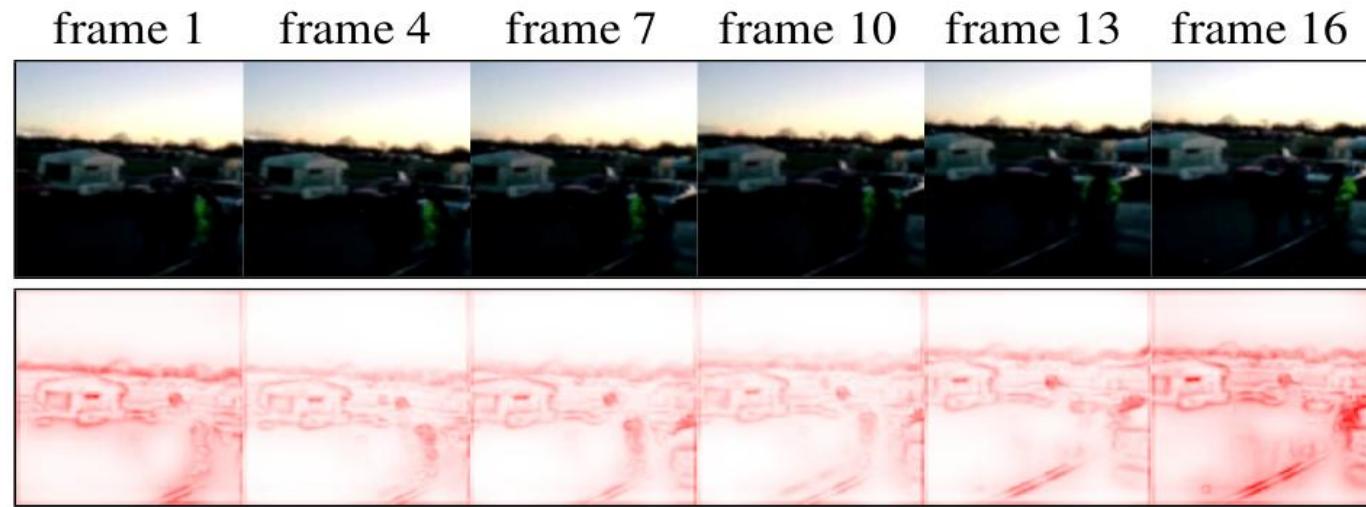
... except for special cases. LRP was developed among others because gradient-based methods aren't satisfying.

When comparing with LRP, please use appropriate LRP parameters (Like when comparing different ML techniques).

**Good news:** No need to reimplement LRP, check our software at [www.heatmapping.org](http://www.heatmapping.org).

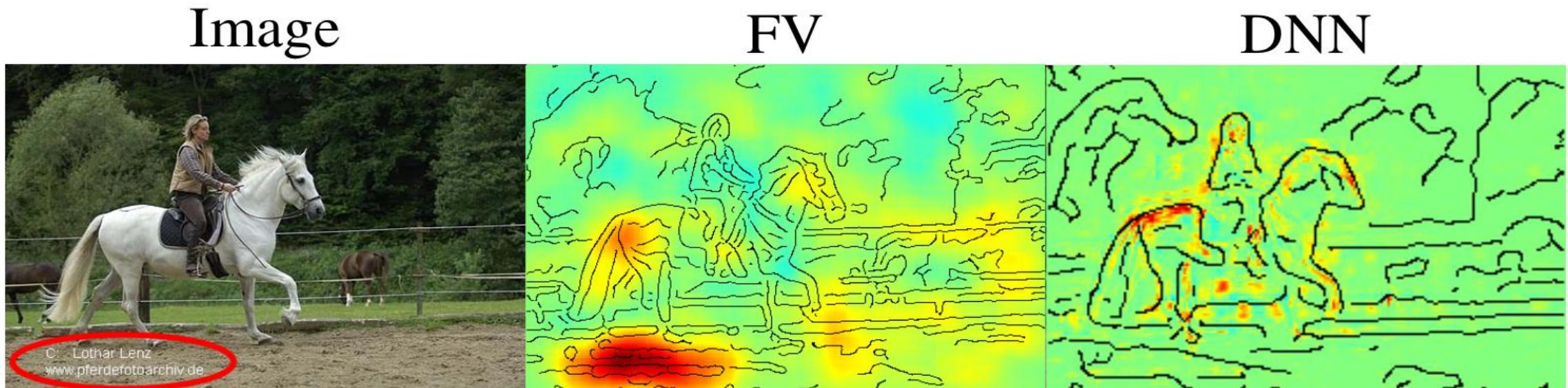
Explanations can be evaluated:  
Pixel flipping (model agnostic)  
And beyond LRP and DTD

# Explanation helps to improve models



## Explaining ML, Now What?

# Explanation helps to find flaws in models



[Lapuschkin et al CVPR 2016]

# Getting new Insights in the Sciences

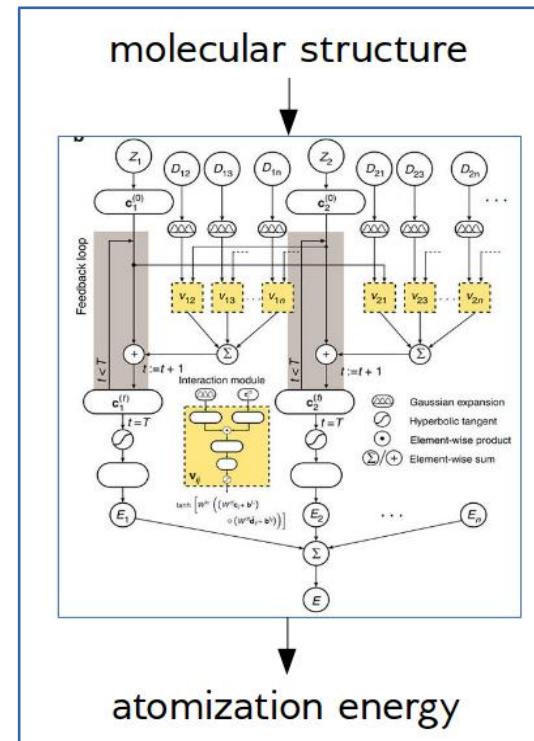
**Example:** Understanding physical systems at the quantum level.

time-independent Schrödinger Equation

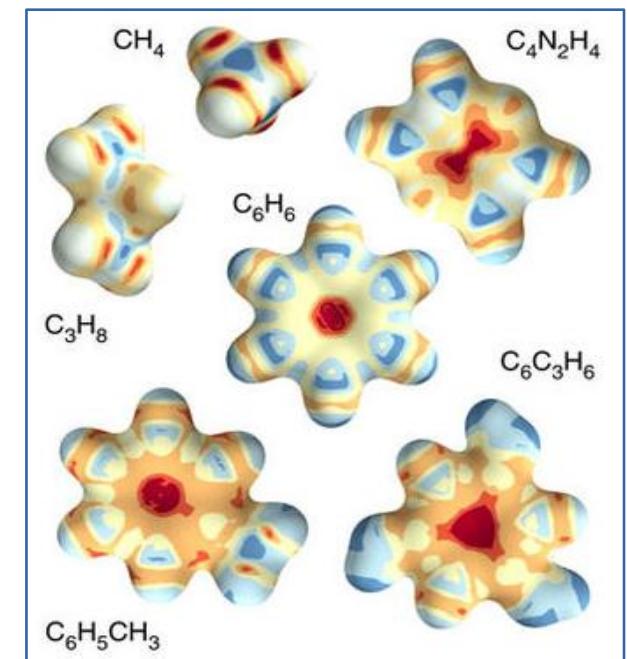
$$\hat{H}\Psi = E\Psi$$

Hamiltonian      energy

equation describing general physical systems



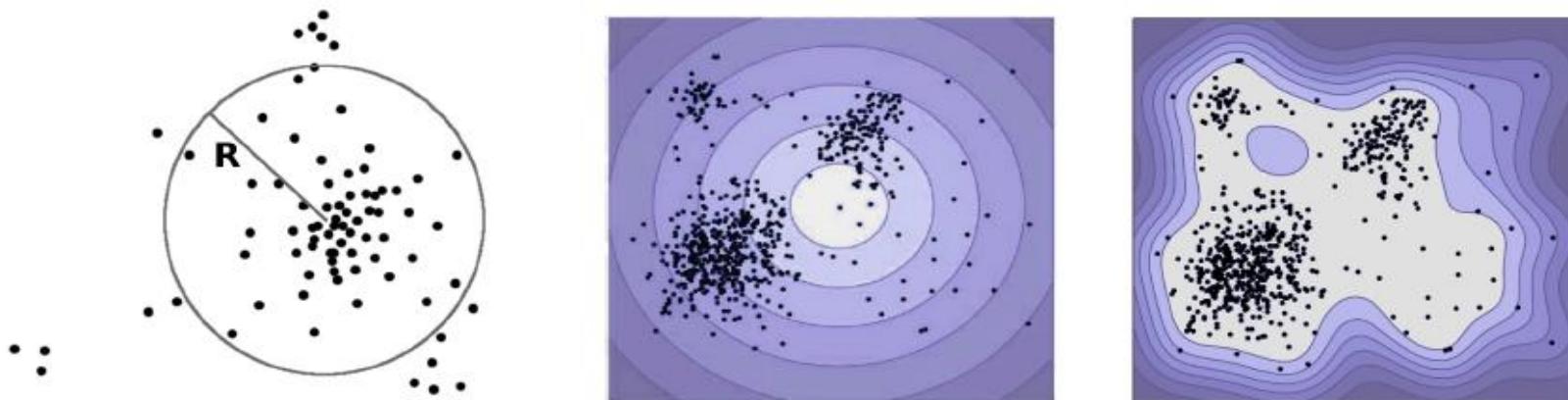
DNN approximation for organic molecules



Interpretation of the trained DNN model

[Schütt et al. Nat Comm. 2017, Schütt et al JCP 2018, Chmiela et al. Sci. Adv. 2017,...]

# Support Vector Data description



## Support Vector Data Description (SVDD)

- Compute minimal enclosing sphere with center  $\mathbf{c}$  and radius  $R$
- Anomaly score as the distance to center  $\mathbf{c}$ , that is  $f(\mathbf{x}) = \|\phi(\mathbf{x}) - \mathbf{c}\|$
- Accept data point  $\mathbf{x}$  if  $f(\mathbf{x}) \leq R$  and ...  
... reject  $\mathbf{x}$  if  $f(\mathbf{x}) > R$

# Explaining one-class

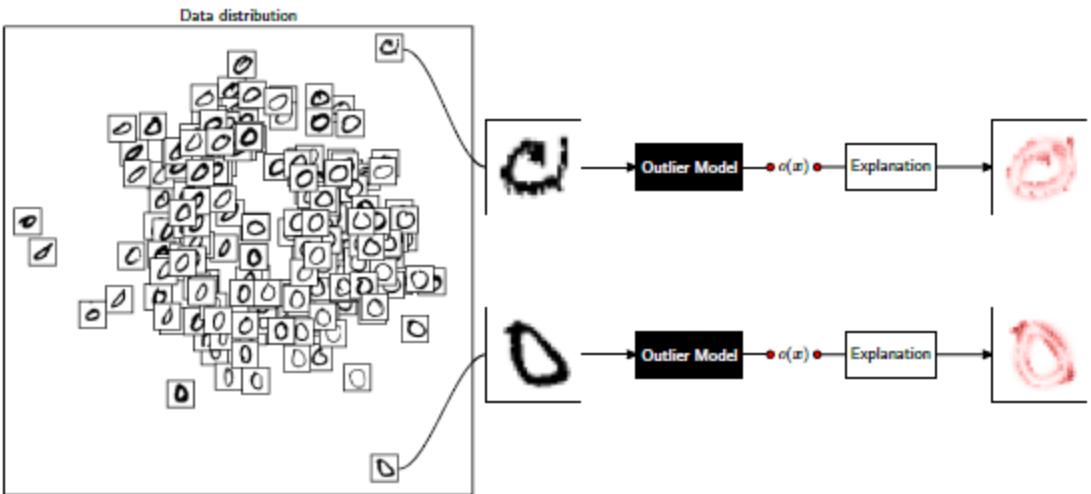


Figure 1: Illustration of the outlier detection and explanation setting. *Left:* Data is generated from an unknown distribution, we are for example interested in potential outliers; *Middle:* Unsupervised machine learning techniques estimate the data generating distribution and assign an outlier score  $o(x)$  to unlikely data points; *Right:* Our explanation method assigns a relevance score to every input variable that reflects the contribution of input variable  $x_i$  to the model decision. We apply dithering to all heatmaps for printing reliability.

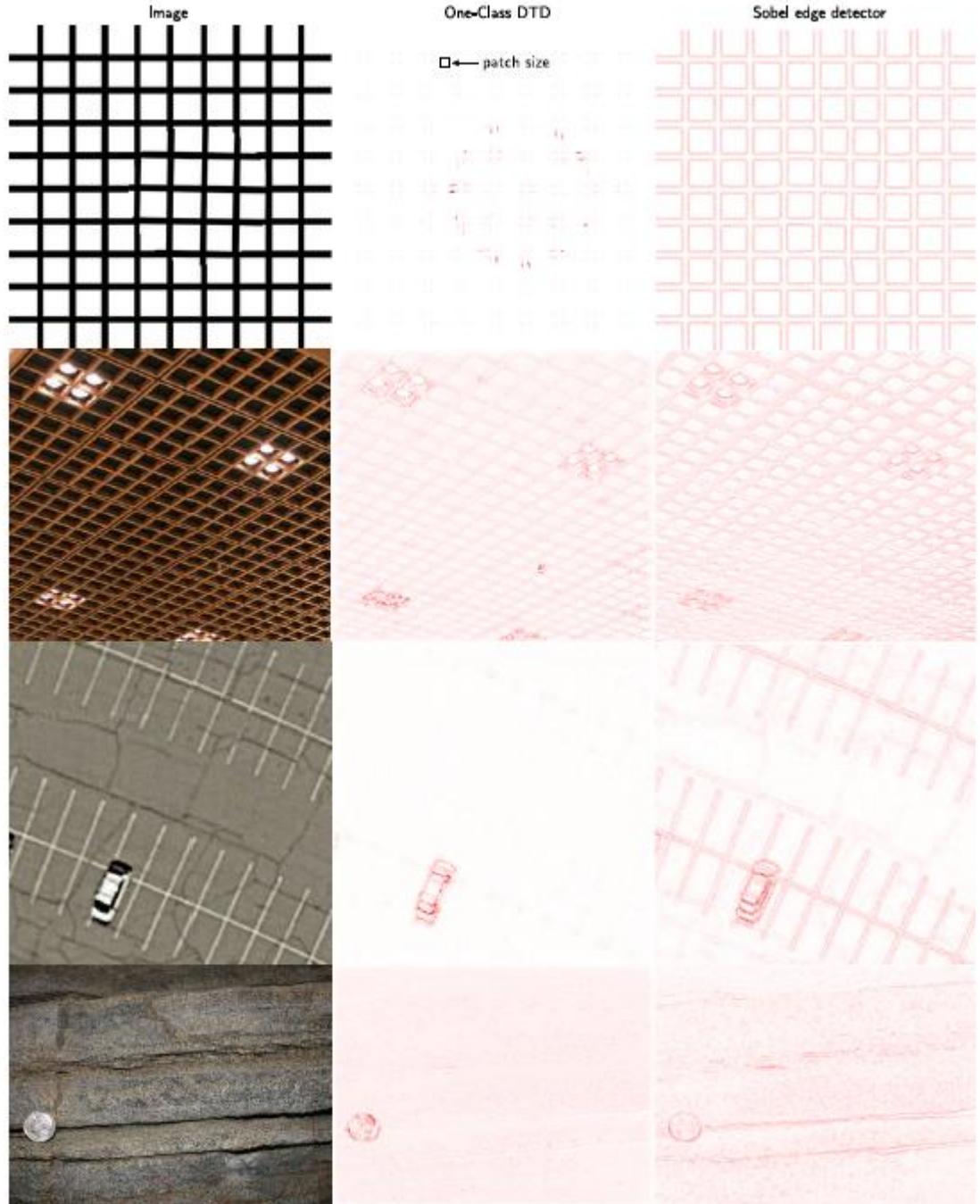


Figure 5: A One-Class SVM is trained on small  $7 \times 7$  patches of the very image itself. Parameter  $\nu = 0.1$  is set to allow at most 10% outliers. Images from a texture data set [11] (row one, two and four) and PatternNet [61]; top image is altered by us. For every image, we show *Left:* input image; *Middle* decomposition of one-class SVM; *Right* Sobel filter for reference. All images were resized to 256 pixels width.

## Semi-final Conclusion

- explaining & interpreting nonlinear models is essential
- orthogonal to improving DNNs and other models
- need for opening the blackbox ...
- understanding nonlinear models is essential for Sciences & AI
- new **theory**: LRP is based on deep taylor expansion
- compare the right thing

[www.heatmapping.org](http://www.heatmapping.org)

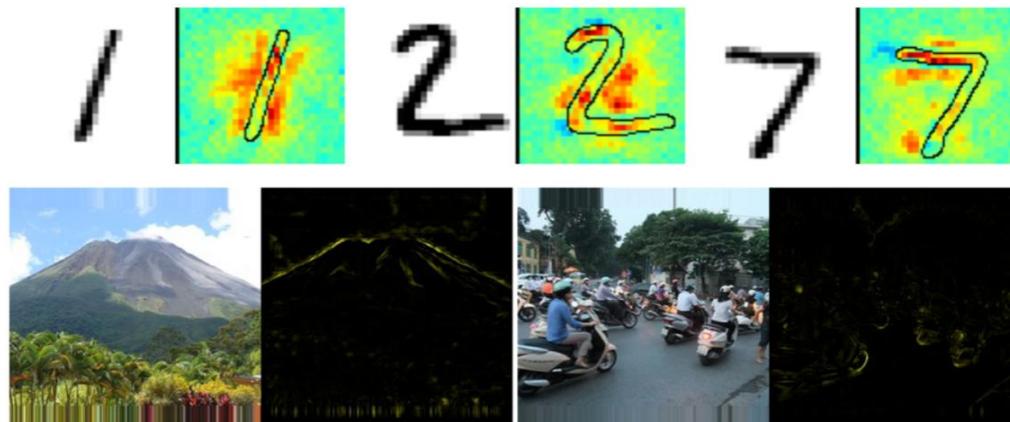
# Thank you for your attention

---

Visit:

<http://www.heatmapping.org>

- ▶ Tutorials
- ▶ Software
- ▶ Online Demos



## Tutorial Paper

Montavon et al., “Methods for interpreting and understanding deep neural networks”, Digital Signal Processing, 73:1-5, 2018

## Keras Explanation Toolbox

<https://github.com/albermax/investigate>

State-of-the-Art  
Survey

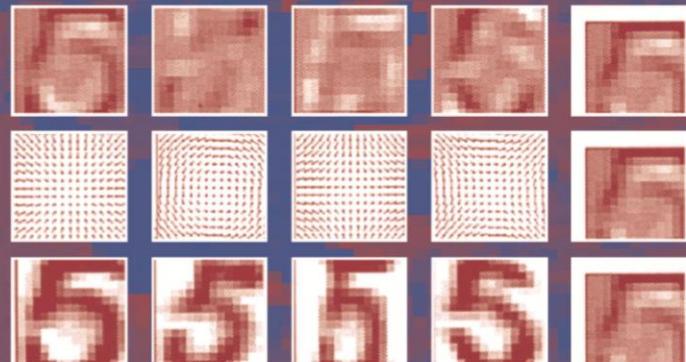
LNCS 7700

Grégoire Montavon  
Genevieve B. Orr  
Klaus-Robert Müller (Eds.)

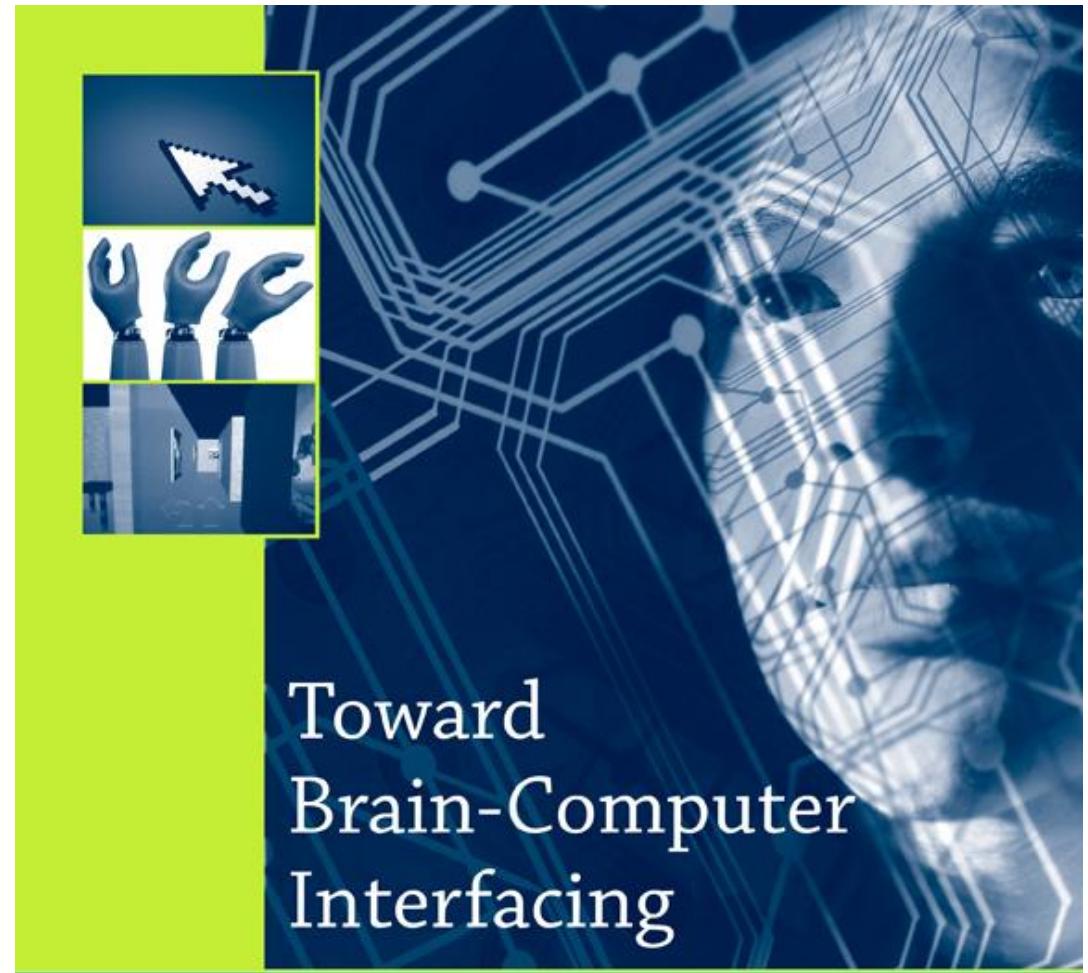
# Neural Networks: Tricks of the Trade

Second Edition

RELOADED



 Springer



edited by  
Guido Dornhege, José del R. Millán,  
Thilo Hinterberger, Dennis J. McFarland,  
and Klaus-Robert Müller

foreword by Terrence J. Sejnowski



BERLIN BIG  
DATA CENTER



# Further Reading I

- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one*, 10, e0130140 (7).
- Bach, S., Binder, A., Montavon, G., Müller, K.-R. & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2912-2920 (2016).
- Binder et al. Machine Learning for morpho-molecular Integration, *arXiv:1805.11178* (2018)
- Blum, L. C., & Reymond, J. L. (2009). 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25), 8732-8733.
- Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., & Müller, K. R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), e1603015.
- Hansen, K., Montavon, G., Biegler, F., Fazli, S., Rupp, M., Scheffler, M., von Lilienfeld, A.O., 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, no. 8 (2013): 3404-3419.
- Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., von Lilienfeld, O. A., Müller, K. R., & Tkatchenko, A. (2015). Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space, *J. Phys. Chem. Lett.* 6, 2326–2331.
- Harmeling, S., Ziehe, A., Kawanabe, M., & Müller, K. R. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5), 1089-1124.
- Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, KR Muller (1999), Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, 1999. *Proceedings of the 1999 IEEE Signal Processing Society Workshop*, 41-48.

# Further Reading II

- Müller, K. R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2), 181-201.
- Montavon, G., Braun, M. L., & Müller, K. R. (2011). Kernel analysis of deep networks. *The Journal of Machine Learning Research*, 12, 2563-2581.
- Montavon, Grégoire, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert Müller. "Learning invariant representations of molecules for atomization energy prediction." In *Advances in Neural Information Processing Systems*, pp. 440-448 . (2012).
- Montavon, G., Braun, M., Krueger, T., & Muller, K. R. (2013). Analyzing local structure in kernel-based learning: Explanation, complexity, and reliability assessment. *IEEE Signal Processing Magazine*, 30(4), 62-74.
- Montavon, G., Orr, G. & Müller, K. R. (2012). Neural Networks: Tricks of the Trade, Springer LNCS 7700. Berlin Heidelberg.
- Montavon, Grégoire, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. "Machine learning of molecular electronic properties in chemical compound space." *New Journal of Physics* 15, no. 9 (2013): 095003.
- Montavon, G., Lapuschkin, S., Binder, A., Samek, W. and Müller, K.R., Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65, 211-222 (2017)
- Montavon, G., Samek, W., & Müller, K. R., Methods for interpreting and understanding deep neural networks, *Digital Signal Processing*, 73:1-5, (2018).

# Further Reading III

- Rupp, M., Tkatchenko, A., Müller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5), 058301.
- K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. U. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties *Phys. Rev. B* 89, 205118 (2014)
- K.T. Schütt, F Arbabzadah, S Chmiela, KR Müller, A Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nature Communications* 8, 13890 (2017)
- K.T. Schütt, H.E. Sauceda, , P.J. Kindermans, , A. Tkatchenko and K.R. Müller, SchNet—A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), p.241722. (2018)
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.R., Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11), pp.2660-2673 (2017)
- .