EMBC Tutorial on Interpretable and Transparent Deep Learning

Wojciech Samek
(Fraunhofer HHI)

Grégoire Montavon
(TU Berlin)

Klaus-Robert Müller
(TU Berlin)

13:30 - 14:00	Introduction KRM
14:00 - 15:00	Techniques for Interpretability GM
15:00 - 15:30	Coffee Break ALL
15:30 - 16:15	Evaluating Interpretability & Applications WS
16:15 - 17:15	Applications in BME & the Sciences and Wrap-Up KRM

Berlin EMBConference 2019

TU berlin

KOREA UNIVERSITY LIBERTAS JUSTITIA VERITAS

mpii max planck institut informatik

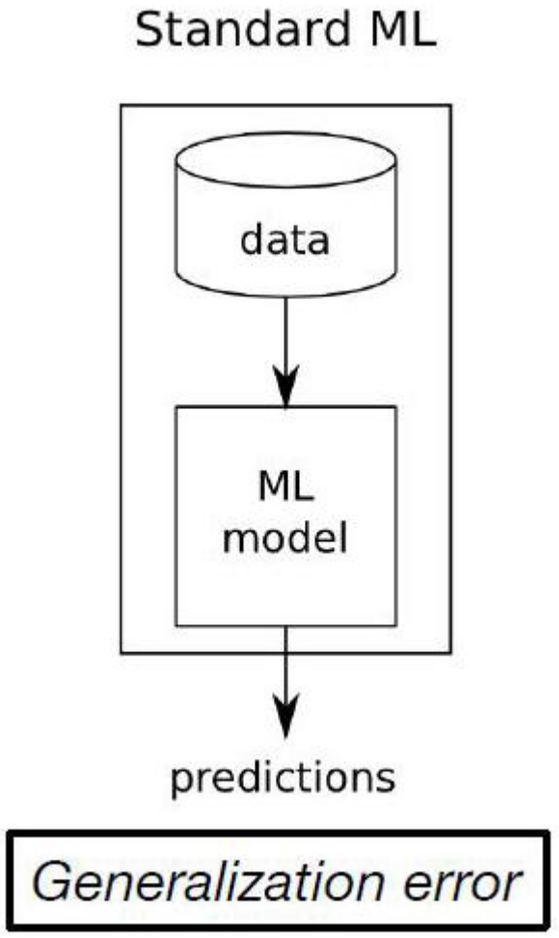
HHI

BBDC BERLIN BIG DATA CENTER

Berliner Zentrum für MASCHINELLES LERNEN

Perspectives

Is the Generalization Error all we need?



Application: Comparing Classifiers (Lapuschkin CVPR'16)

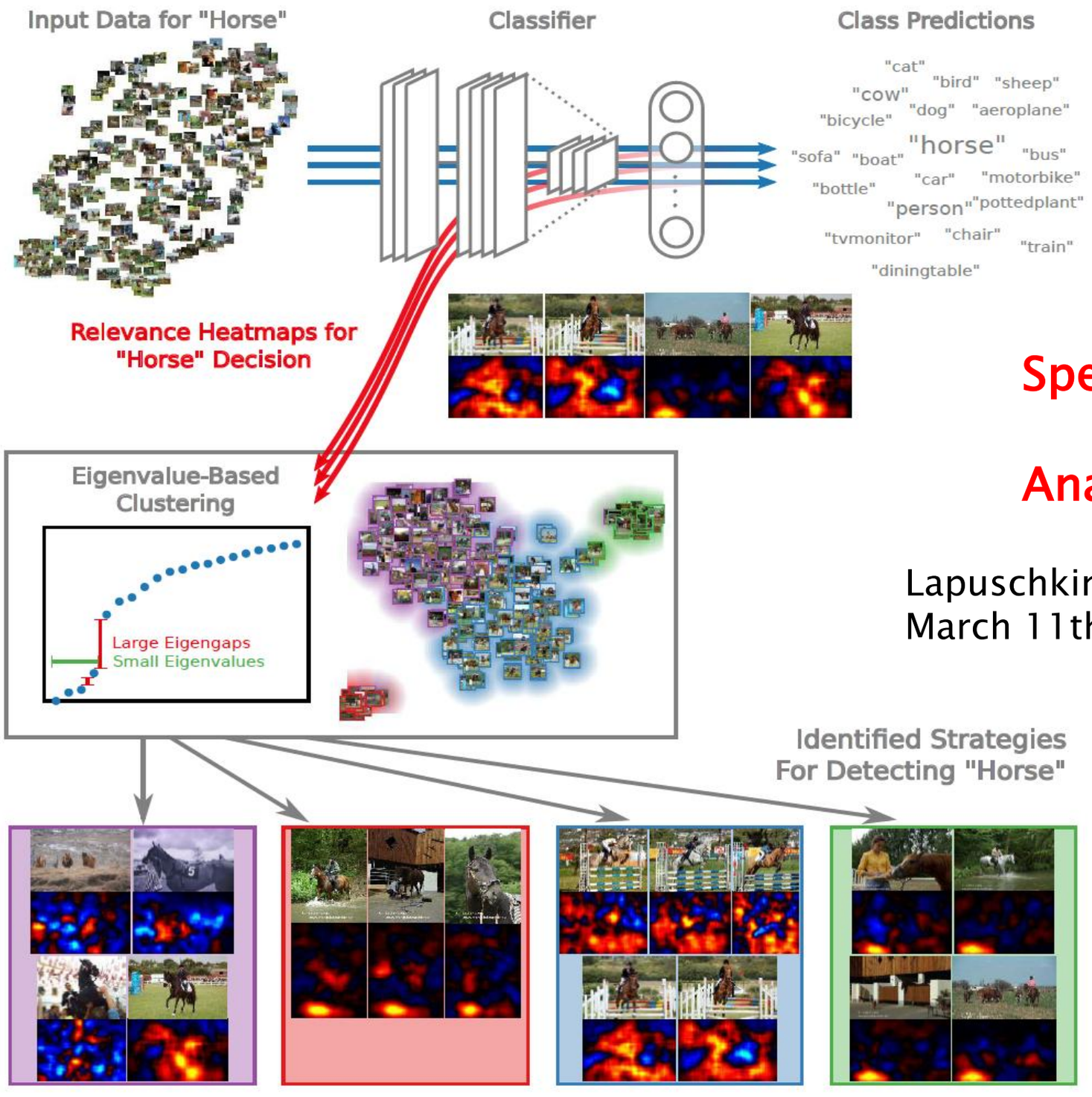
Test error for various classes:

	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


xarif

Explaining problem solving strategies in scale



Spectral Relevance Analysis (SpRAy)

Lapuschkin et al. Nat Comms, March 11th 2019

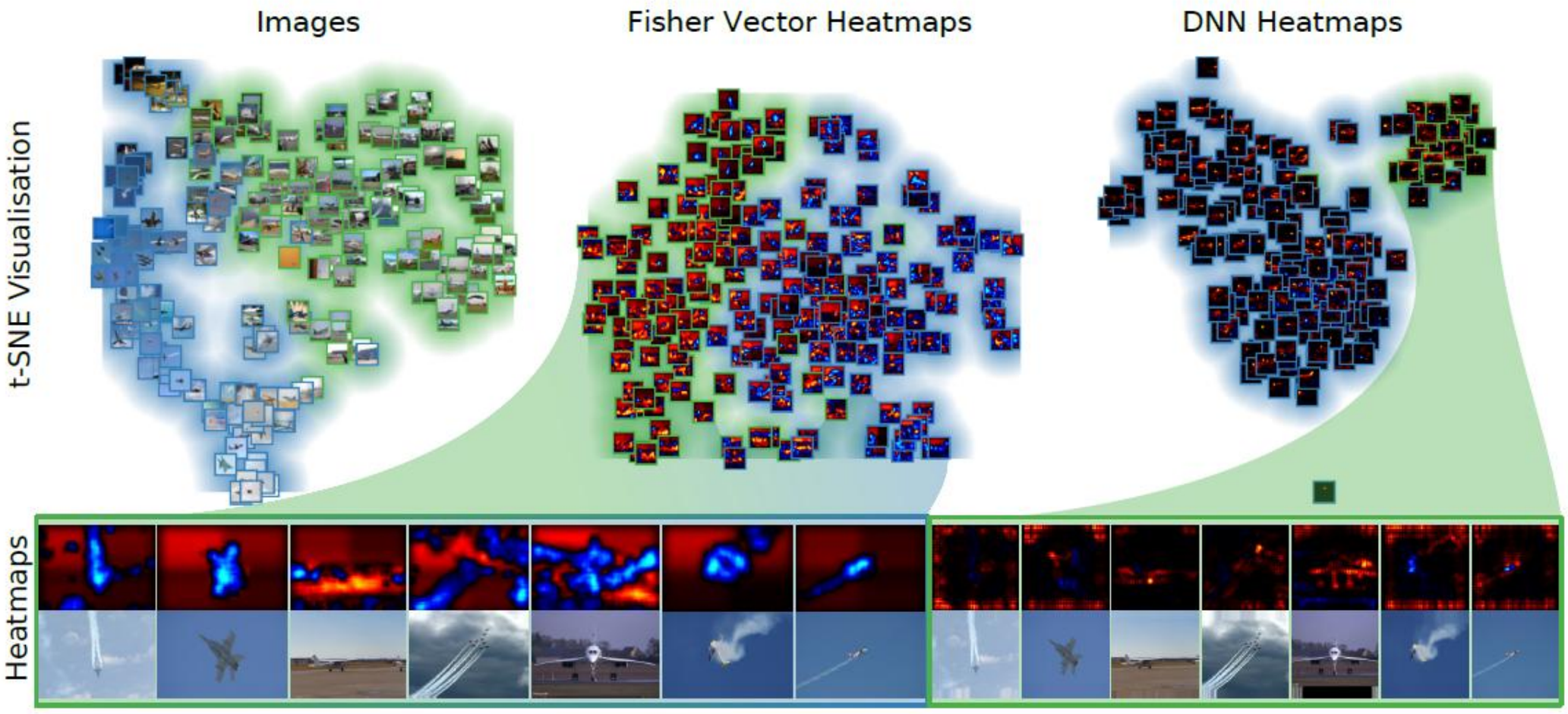
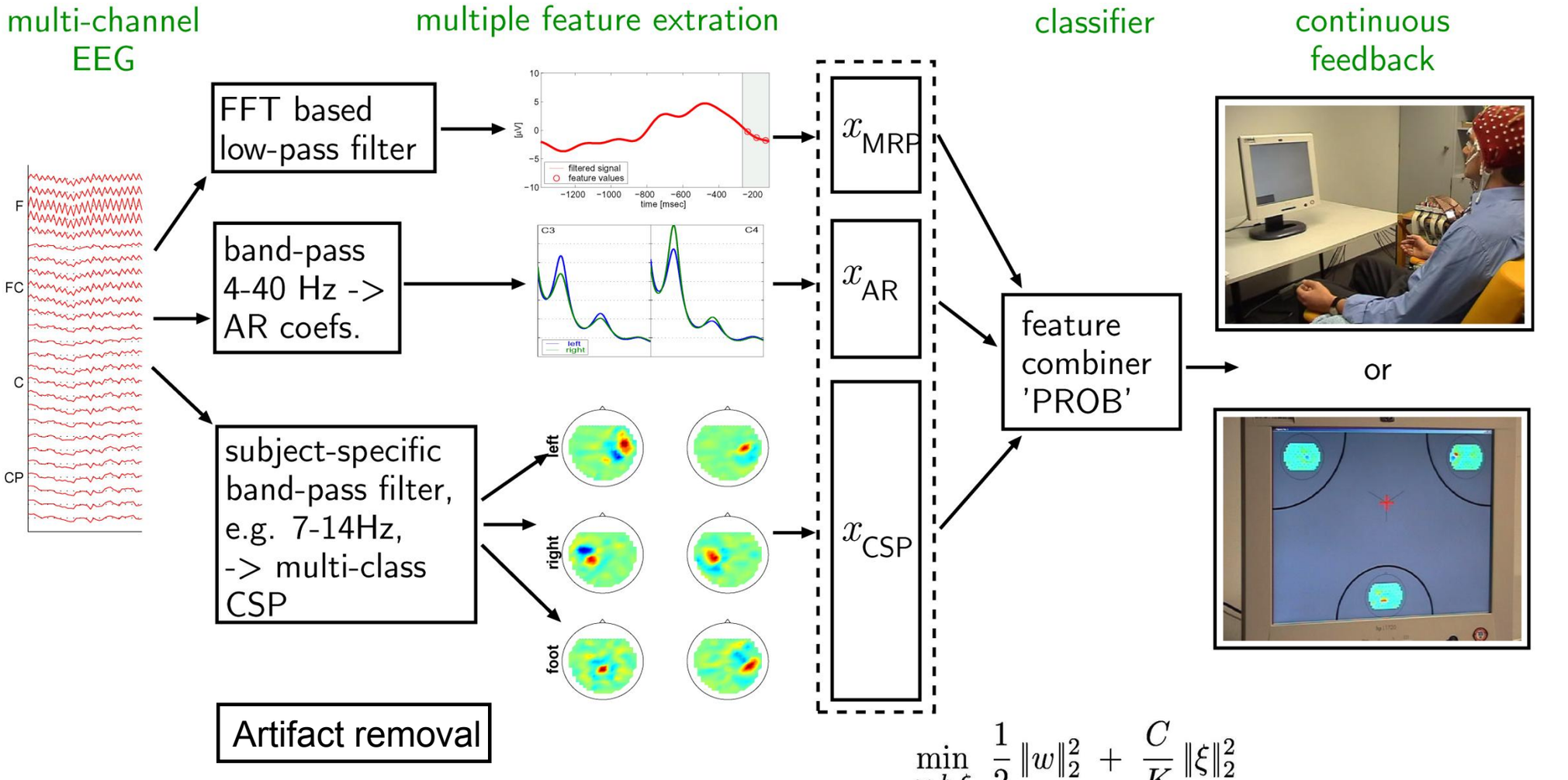


Figure 28: Cluster label assignments for class "aeroplane" via SC for input images, FV model relevance maps and DNN relevance maps. Embedding coordinates in $\mathbb{R}^2$ for visualization have been computed on pair-wise distances derived from the weighted affinity matrix W used for SC. The samples at the bottom right (square images) show DNN relevance maps and images with strong reaction of the DNN models to the border padding. FV relevance maps for the same images are shown to the left. Enlarged relevance maps and images are shown without preprocessing.

Machine Learning in the Sciences

Machine Learning in Neuroscience

BBCI Set-up: *Let the machines learn*



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

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

[cf. Müller et al. 2001, 2007, 2008, Dornhege et al. 2003, 2007, Blankertz et al. 2004, 2005, 2006, 2007, 2008, 2011, Lemm et al 2011]

Brain Computer Interfacing: ‚Brain Pong‘

Berlin Brain Computer Ínterfac

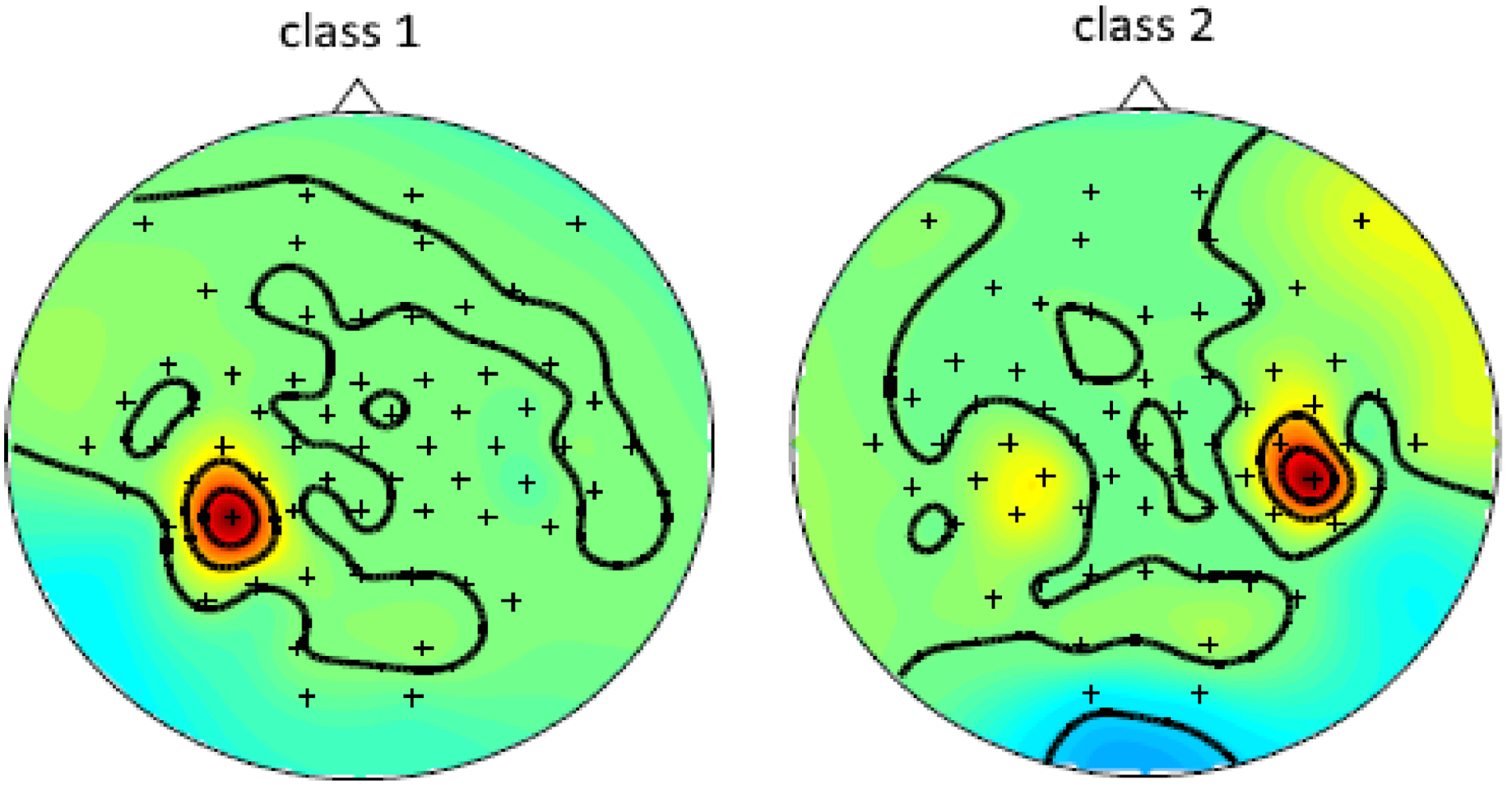
- ML reduces patient training from 300h -> 5min

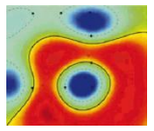
Applications

- 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)

Machine Learning in Chemistry, Physics and Materials

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

[Rupp et al. Phys Rev Lett 2012, Snyder et al. Phys Rev Lett 2012, Hansen et al. JCTC 2013 and JPCL 2015]

Machine Learning for chemical compound space

Ansatz:

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

instead of

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

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

[from von Lilienfeld]

Machine Learning in Chemistry, Physics and Materials

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

Machine Learning for chemical compound space

Ansatz:

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

instead of

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

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

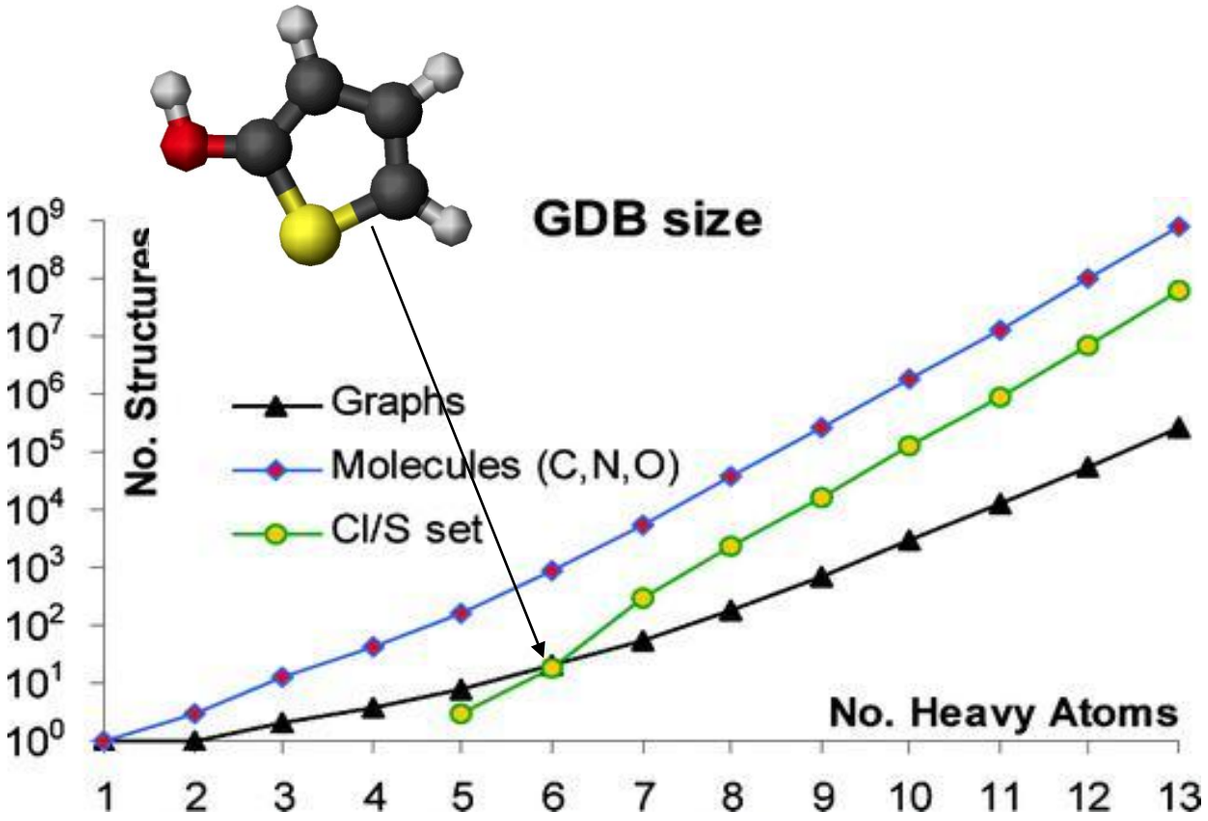
[from von Lilienfeld]

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[a]	graphs[b]	GDB[c]	Cl/S[d]	CPU time (h)[e]
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	5726	315	0.05
8	194	37151	2438	0.33
9	706	255542	17056	2.68
10	2831	1784626	130465	25.26
11	12011	12961686	938704	223.49
12	53789	99821343	7240108	3023.79
13	250268	795244451	59027533	36606.45
Total	**319892**	**910111673**	**67356641**	**39882.08**



Blum & Reymond, *JACS* (2009)

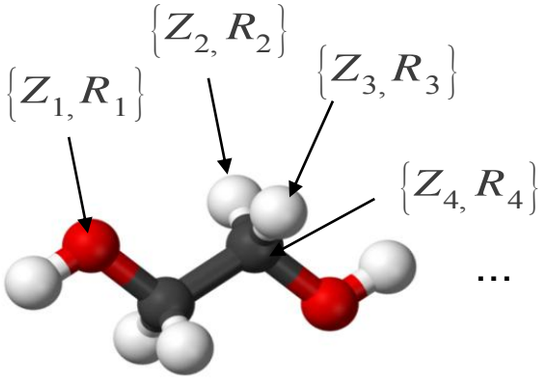
[from von Lilienfeld]

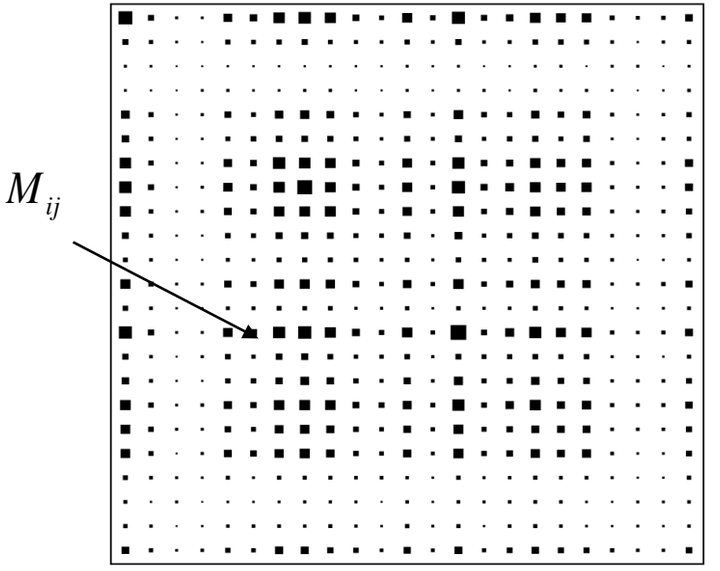
Coulomb representation of molecules

$$M_{ii} = Z_i^{2.4}$$

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

$M \in {}^{23\times 23}$





+ phantom atoms

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

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 **M** define Gaussian kernel matrix **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

$$\min_{\alpha} \quad \sum_i \left(E^{est}(\mathbf{M}_i) - E_i^{ref}\right)^2 + \lambda \sum_i \alpha_i^2$$

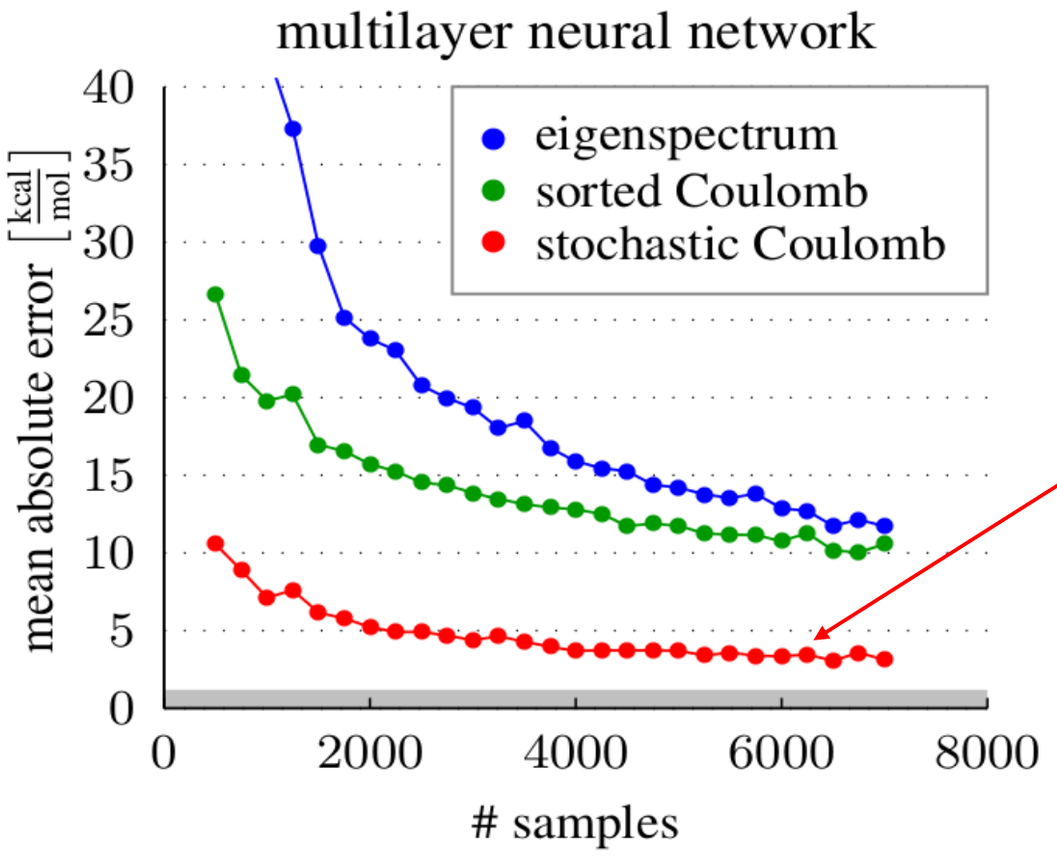
$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{ref}$$

Exact solution

As many parameters as molecules + 2 global parameters, characteristic length-scale or kT of system (σ), and noise-level (λ)

[from von Lilienfeld]

Predicting Energy of small molecules: Results

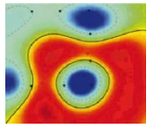


March 2012
Rupp et al., PRL
9.99 kcal/mol
(kernels + eigenspectrum)

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**



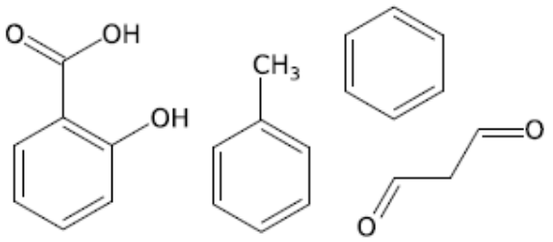
Dataset available at http://quantum-machine.org

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



$$Z = [\, Z_1 \quad Z_2 \quad \cdots \quad Z_n \,]$$

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{12} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Embedding of based on atom types

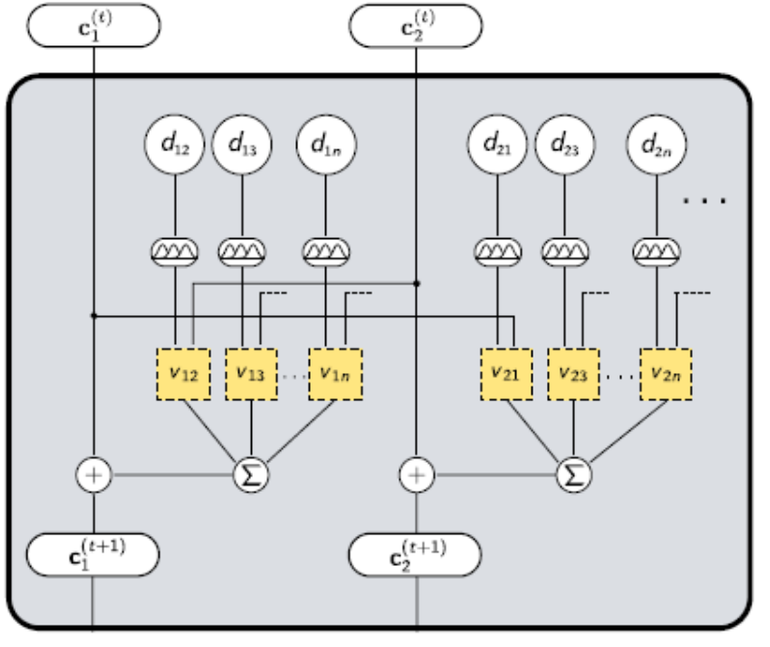
$$\mathbf{x}_i^{(0)} = \mathbf{x}_{Z_i} \in \mathbb{R}^d$$

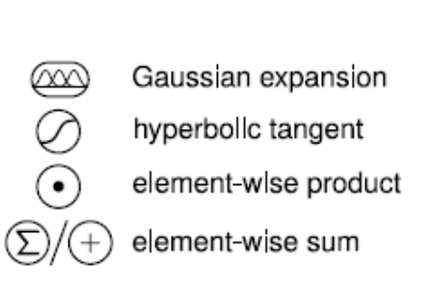
Add interaction with environment using $t = 1 \ldots T$ sequential refinements $\mathbf{v}_i^{(t)}$

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

Prediction via atom-wise contributions:

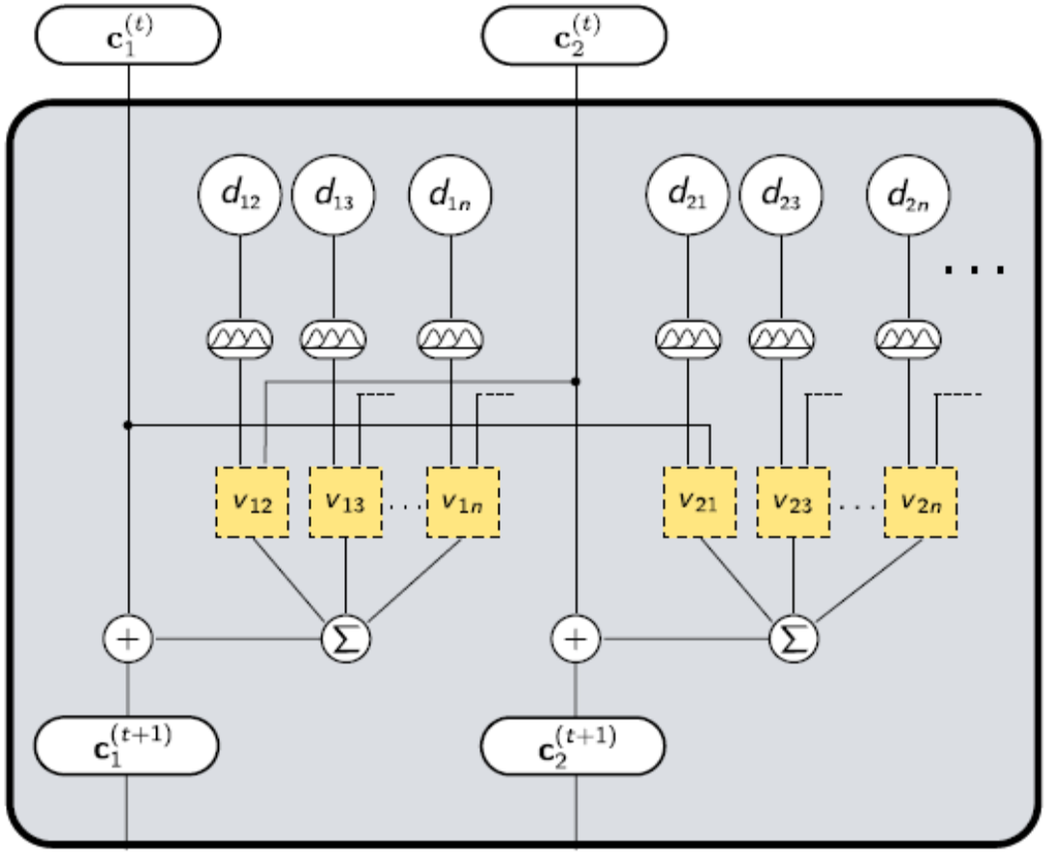
$$\hat{E} = \sum_{i=1}^{n_{\text{atoms}}} f_{\text{out}}(\mathbf{x}_i^{(T)})$$

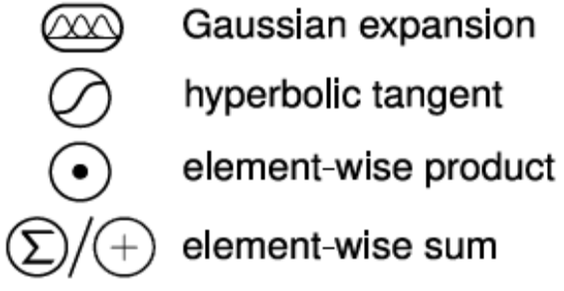


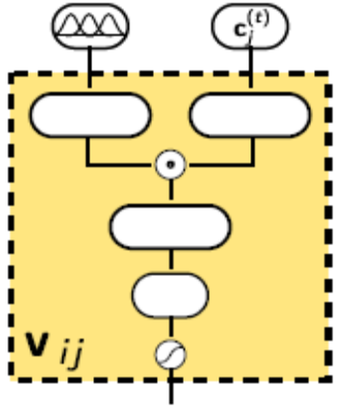


$$\tanh\left(W^{fc}((W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df}\hat{\mathbf{d}}_{ij} + \mathbf{b}^{f_2}))\right)$$

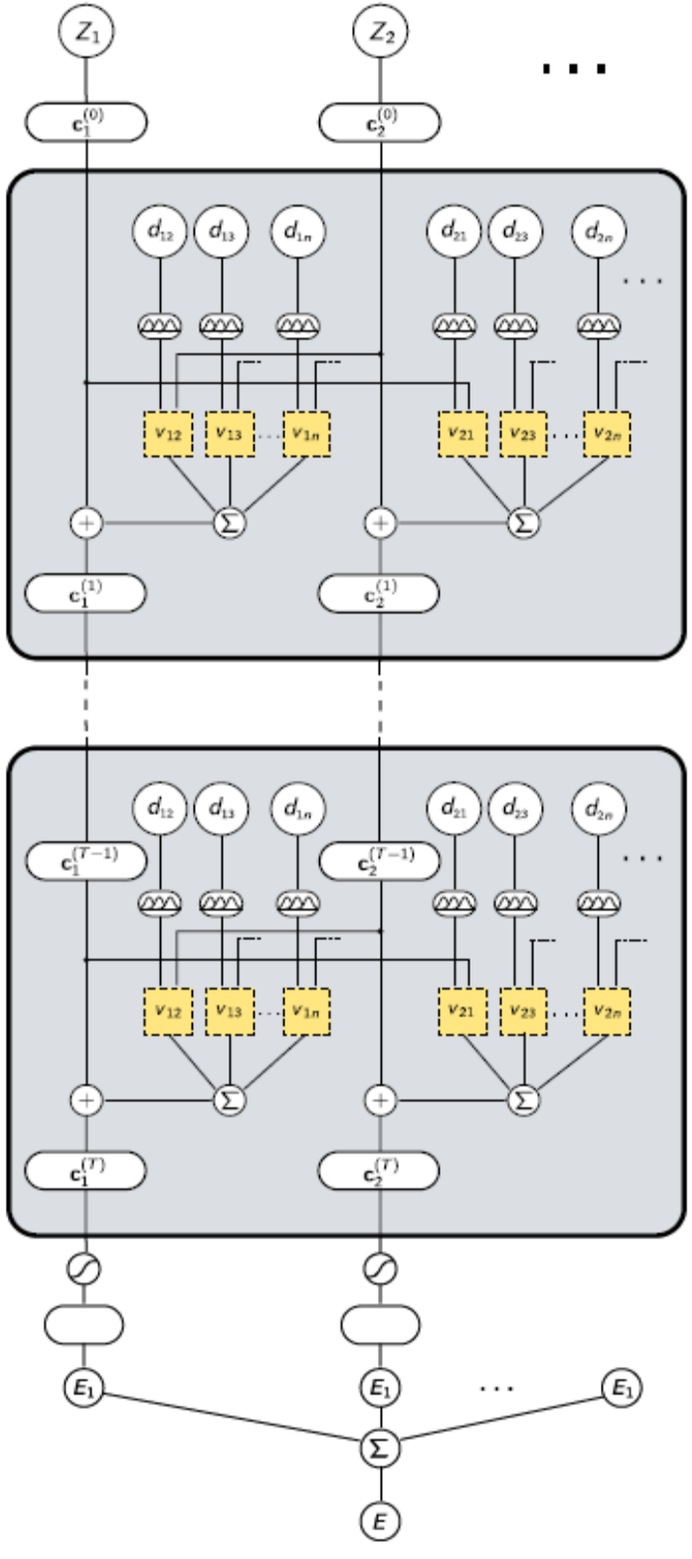
Schütt, Arbabzadah, Chmiela, Müller, Tkatchenko, Nature Communications **8**, *13890 (2017)*





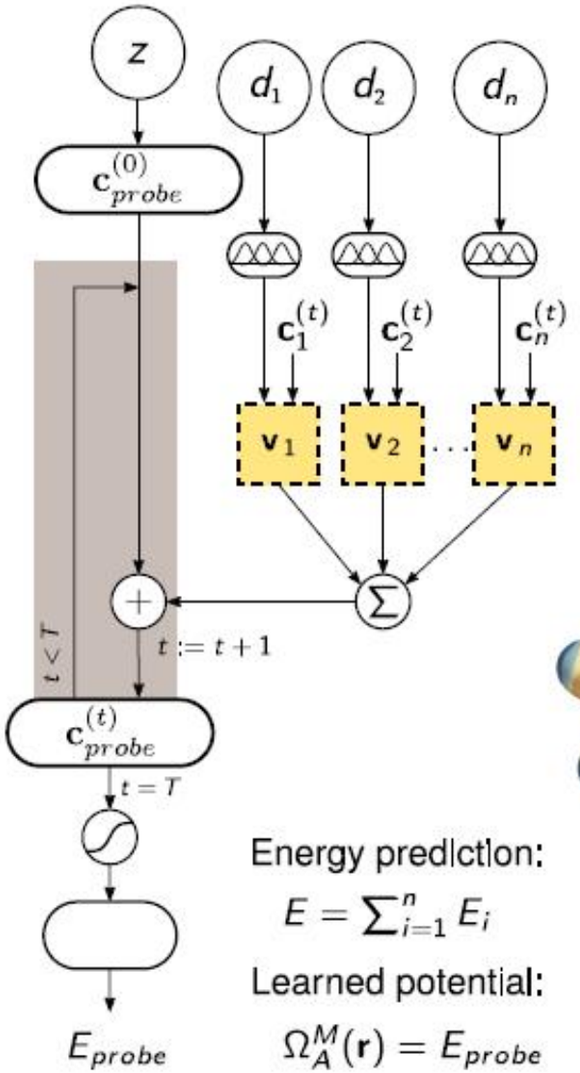


$$\tanh\left(W^{fc}((W^{cf}\mathbf{c}_j + \mathbf{b}^{f_1}) \circ (W^{df}\hat{\mathbf{d}}_{\mathbf{ij}} + \mathbf{b}^{f_2}))\right)$$



Gaining insights for Physics

Toward Quantum Chemical Insights: supervised

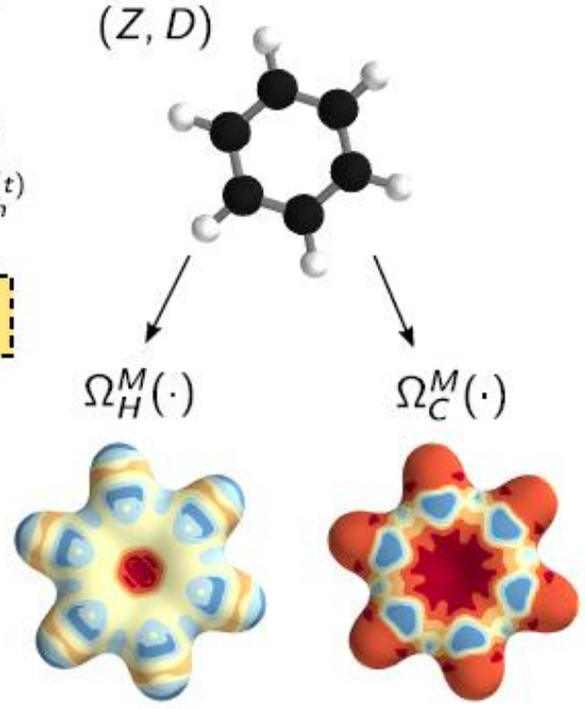


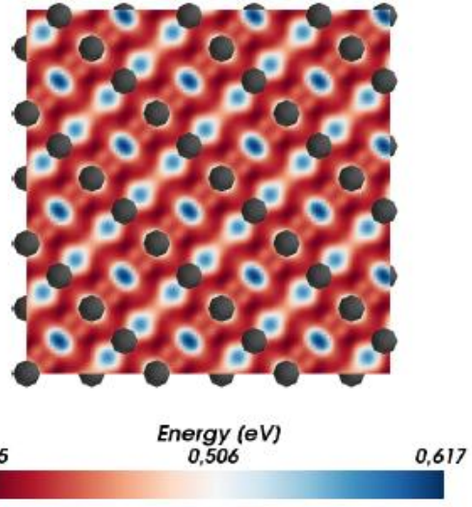
Energy prediction:

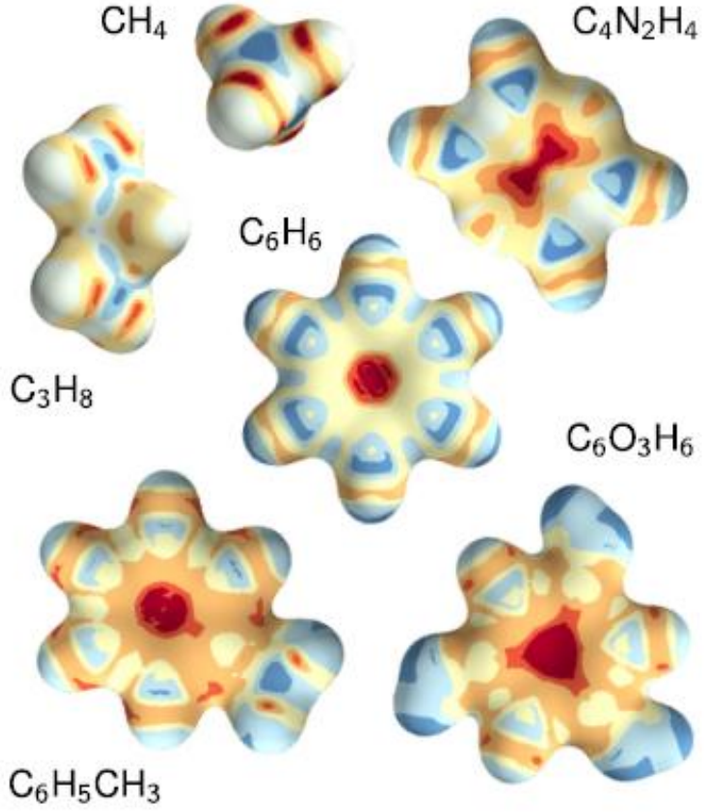
$E = \sum_{i=1}^{n} E_i$

Learned potential:

$\Omega^M_A(\mathbf{r}) = E_{probe}$

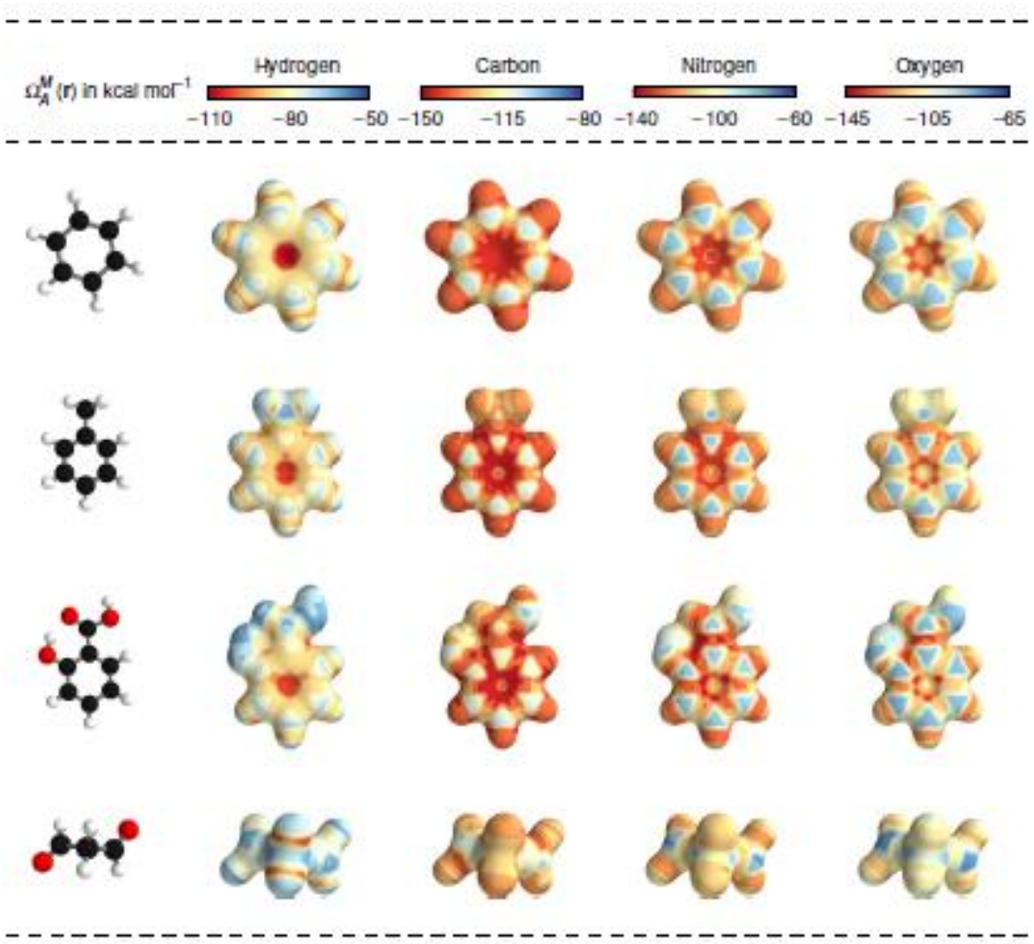




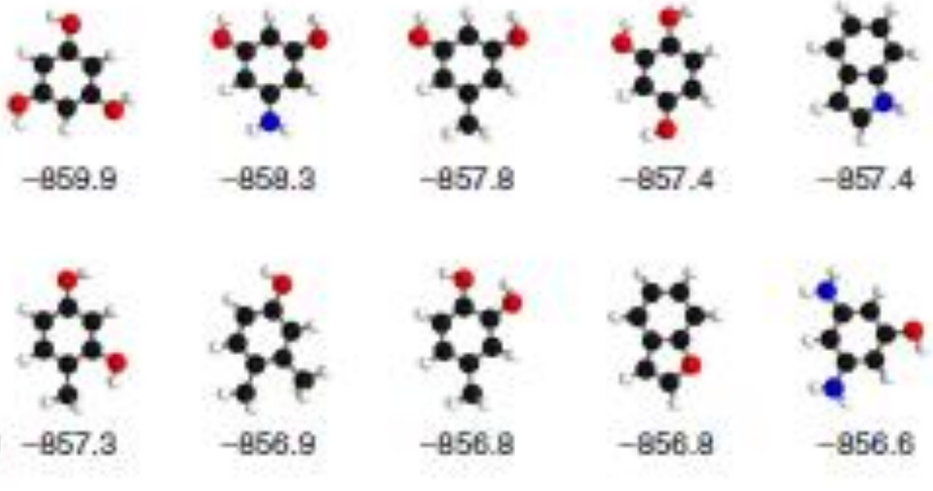


[Schütt et al. Nat Comm. 2017, Schütt et al JCP 2018]

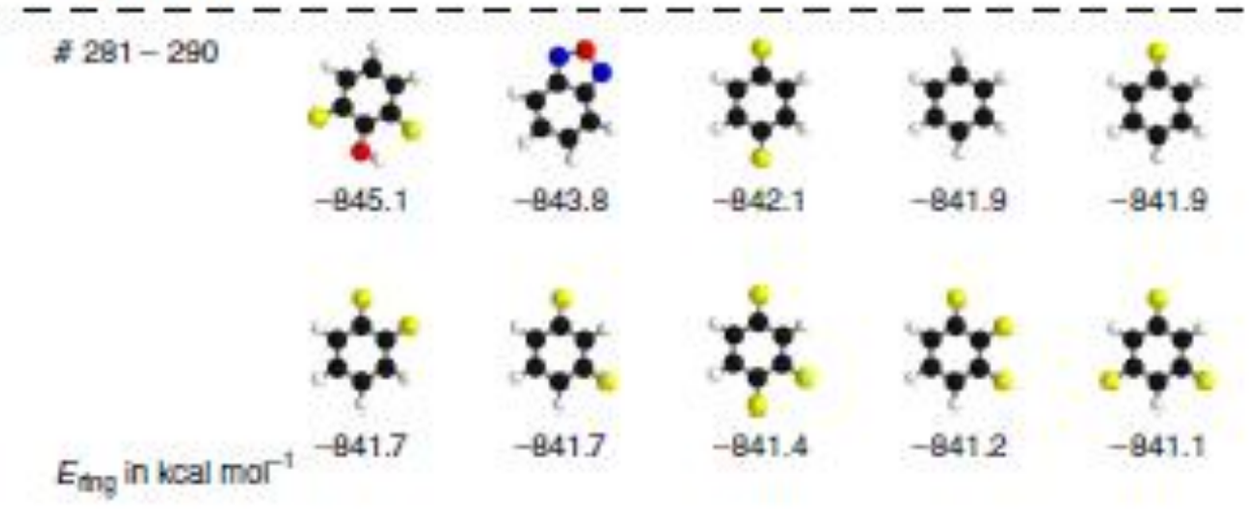
Quantum chemical insights



Inferred chemical potentials



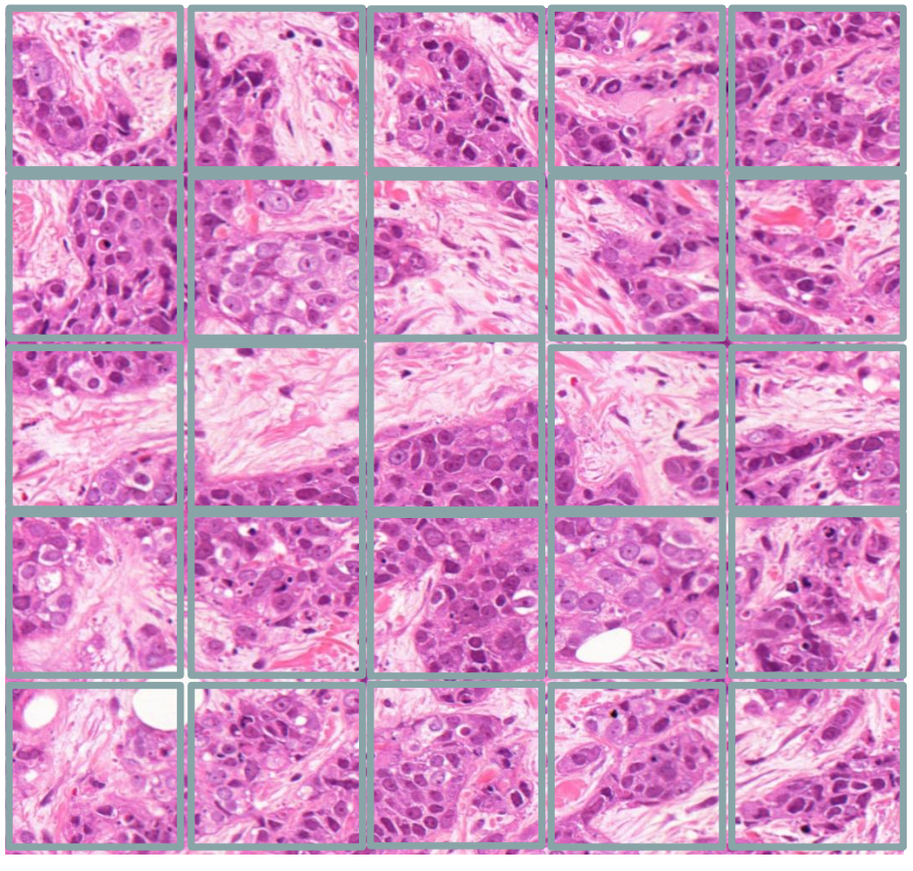
Inferred stable and unstable carbon ring classification ‚aromaticity‘

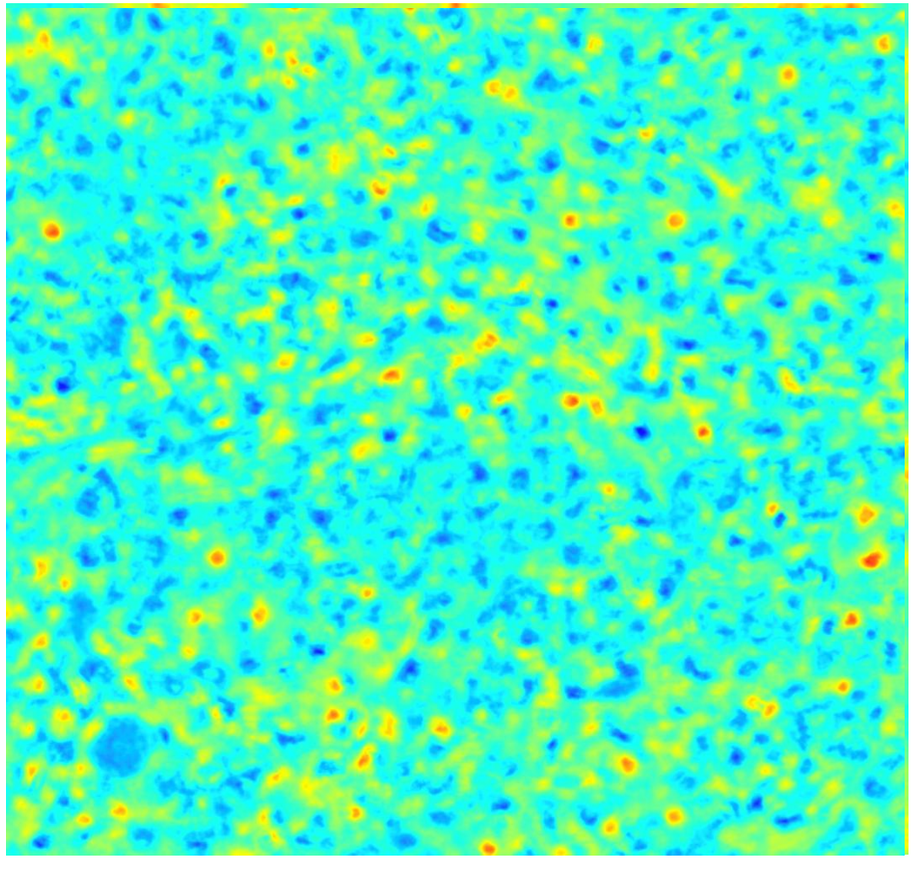


Machine Learning for morpho-molecular Integration

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

Interpretable ML



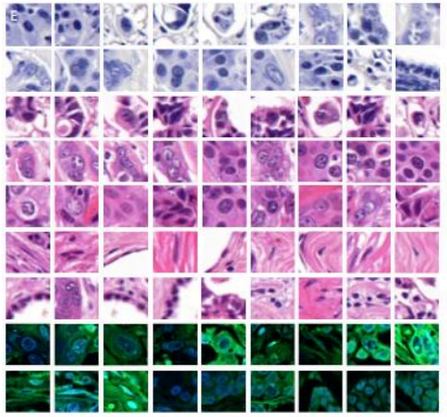


Bach et al., PLoS1 2015
Klauschen et al., US Patent #9558550
Binder et al., *in revision*

Machine learning based integration of morphological and molecular tumor profiles

MICROSCOPIC AND MOLECUAR DATA | TRAINING | PREDICTION | INTEGRATION/ INTERPRETATION

histo-morphological features



div. Image modalities: brightfield, confocal

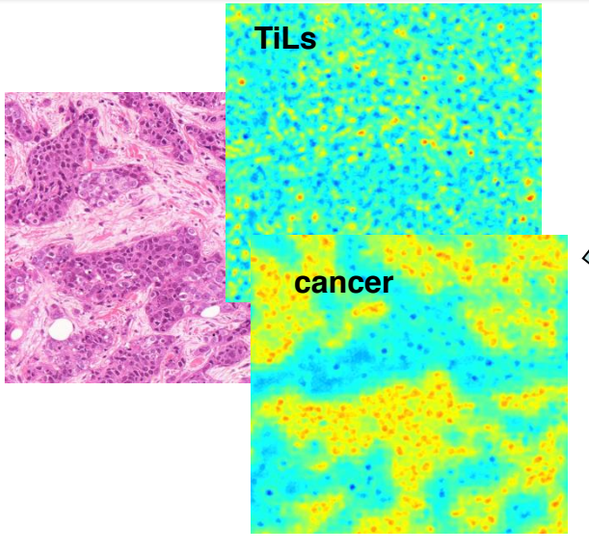
in-house data base

cell types

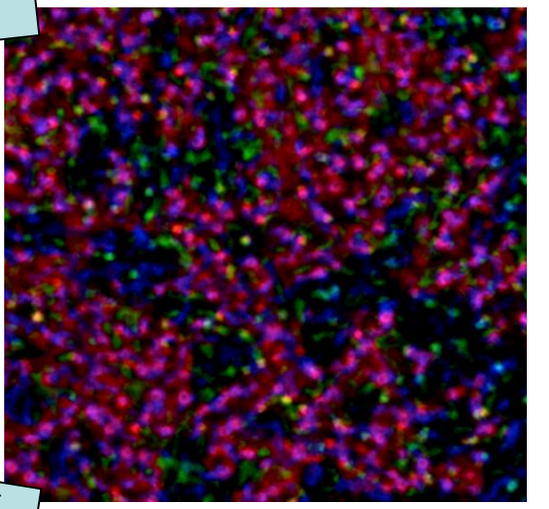
cancer

lymphocytes

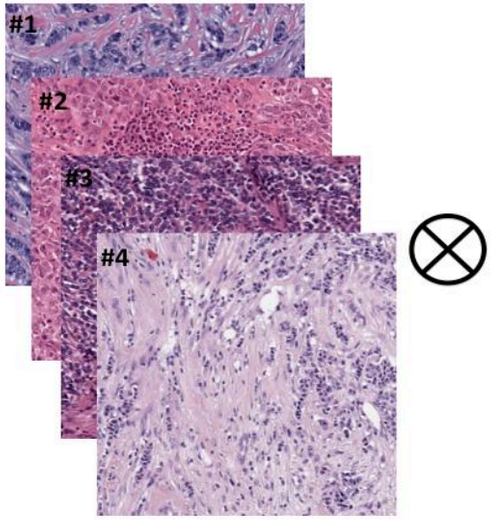
stroma



red: carcinoma,
green: TILs,
blue: molecular property

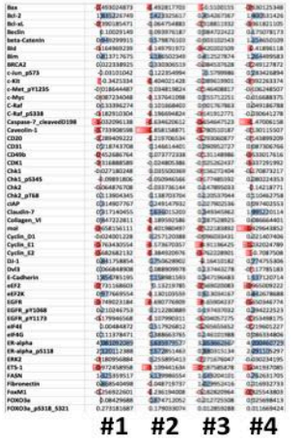


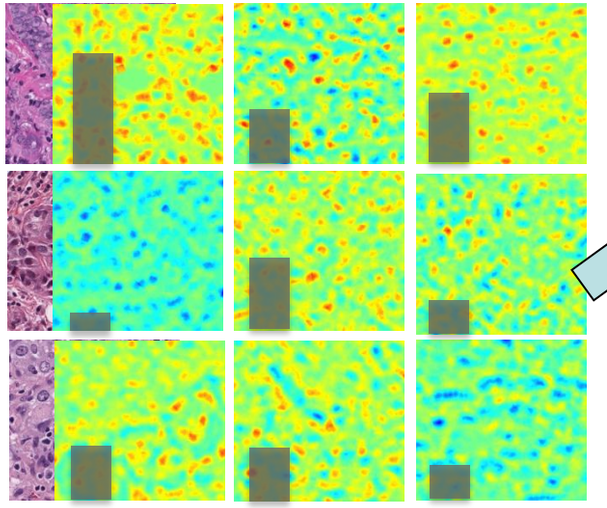
molecular features



TCGA data base

molecular profiles

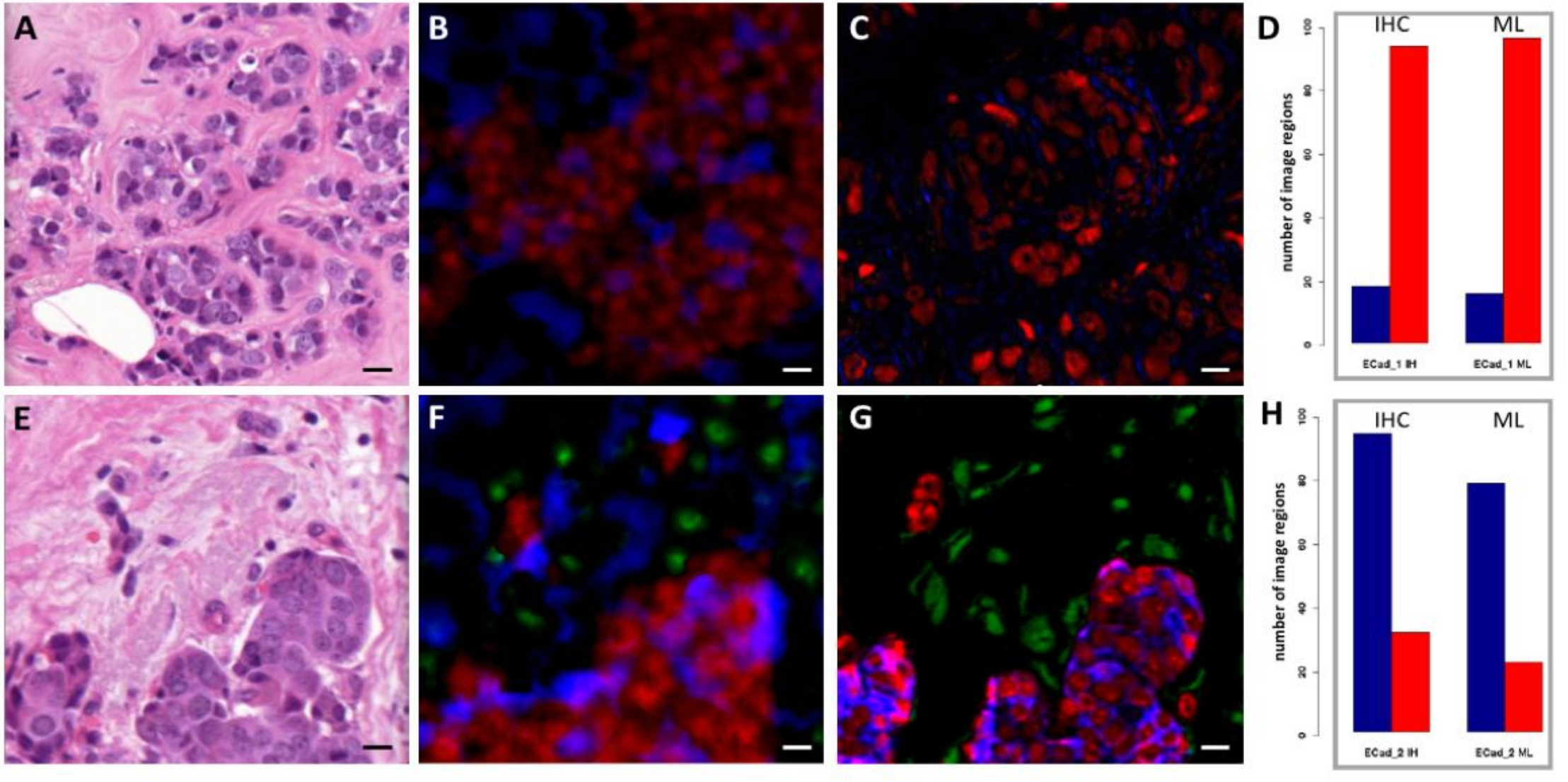




molecular features

„computational fluorescence microscopy“

Heterogenity of E-Cadherin-Expression



Morphology

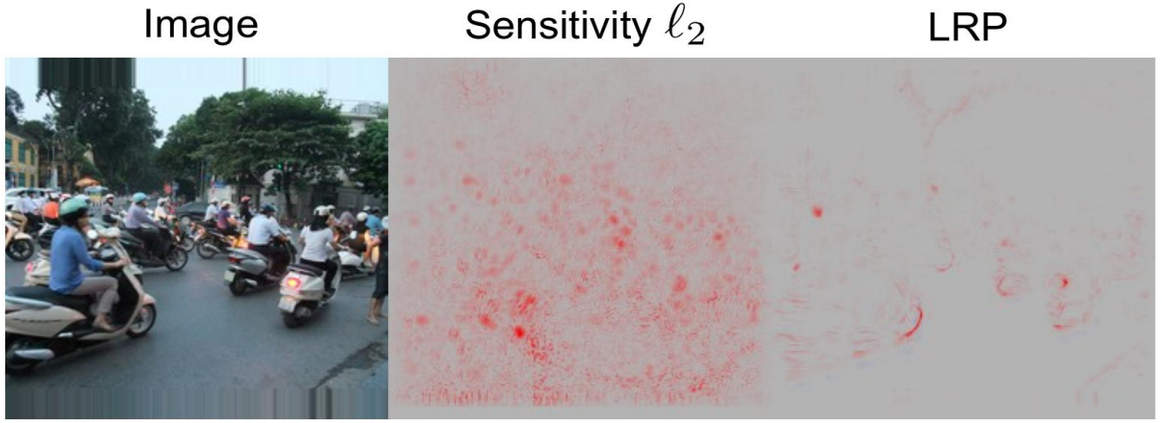
ML-Prediction:
HE+omics

Validation: IHC

Binder et al., in revision.

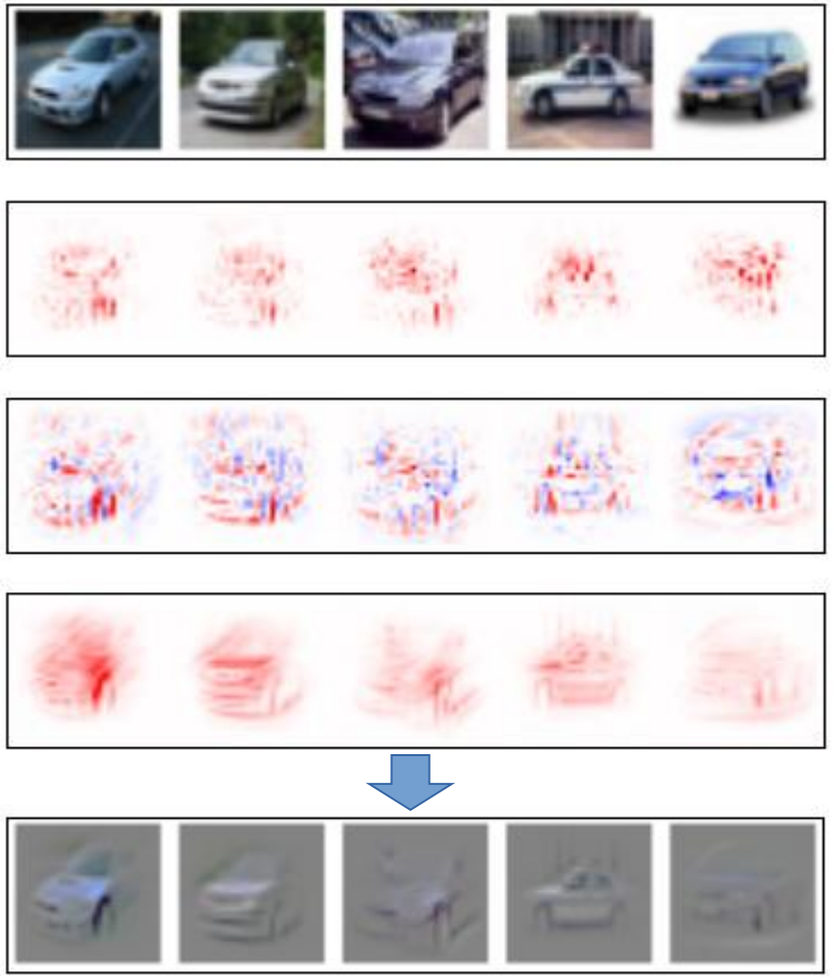
Take Home messages

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



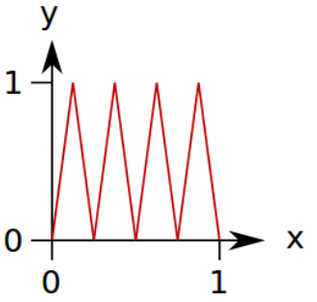
Explanation for simple models does not necessary work for deep models

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"

Underlying theory for consistency

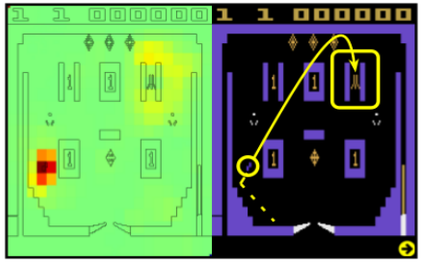
$$R_i = \sum_j \frac{\partial R_j}{\partial a_i} \cdot (a_i - \widetilde{a}_i^{(j)})$$

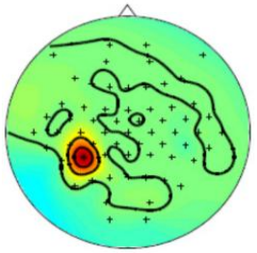
Deep Taylor Decomposition

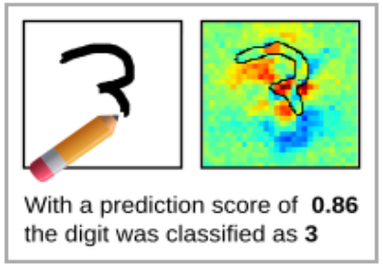
LRP Explanation Framework

(software, tutorials, demos, insights, applications)

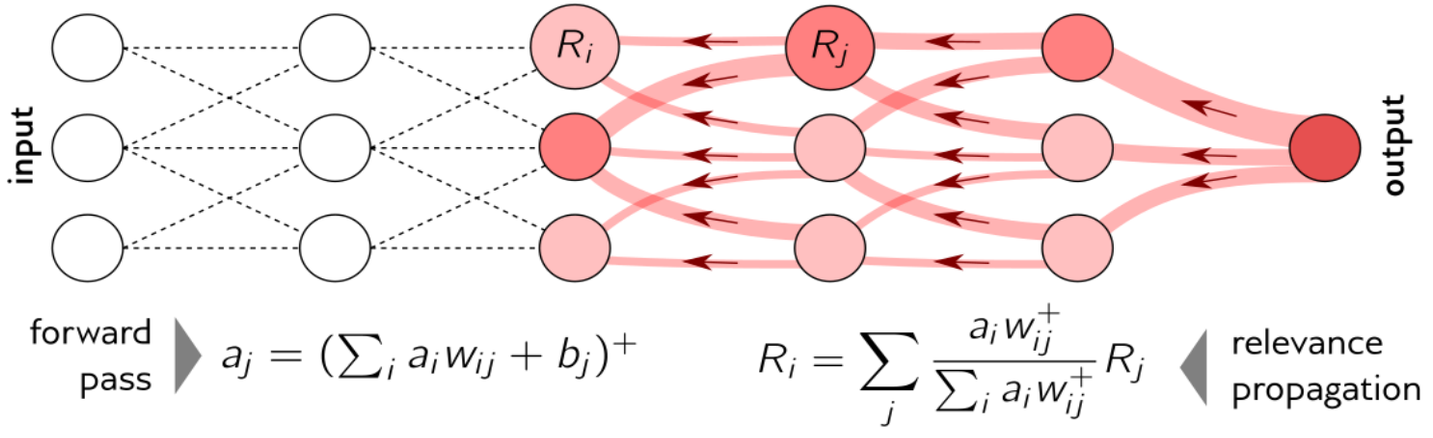
e people are more prone to g
The mental part is usually
y is up or down, ie: the Shu
ointed towards Earth, so the
astronauts. About 50% of t
s, and NASA has done numerou







LRP works 4 all: deep models, LSTMs, kernel methods …



A Clarification on LRP

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

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

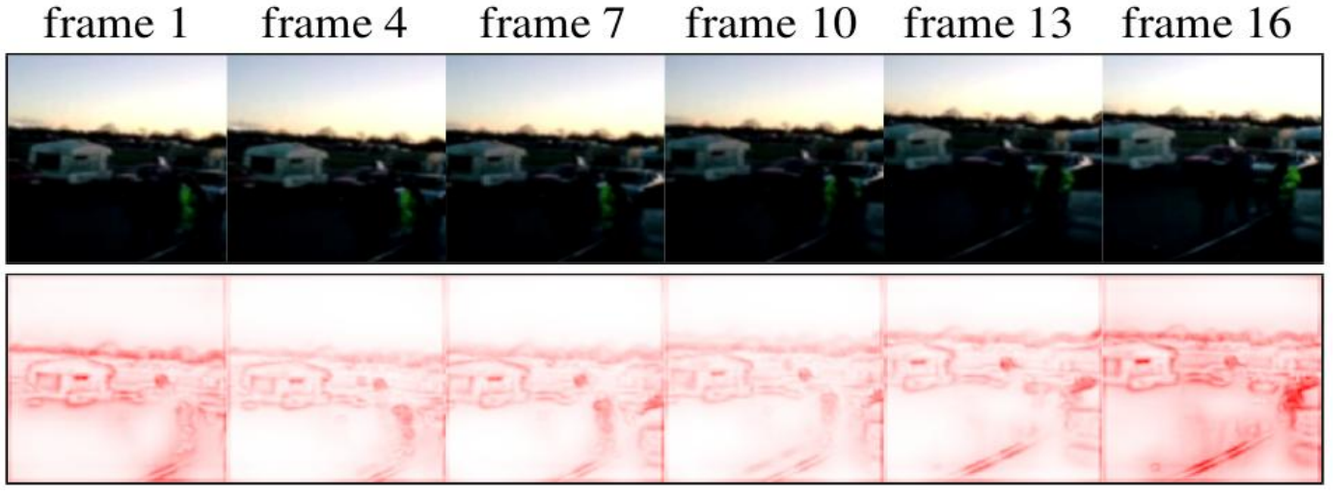
When comparing with LRP, please use appropriate LRP parameters (Like when comparing different ML techniques).

Good news: No need to reimplement LRP, check our software at www.heatmapping.org.

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

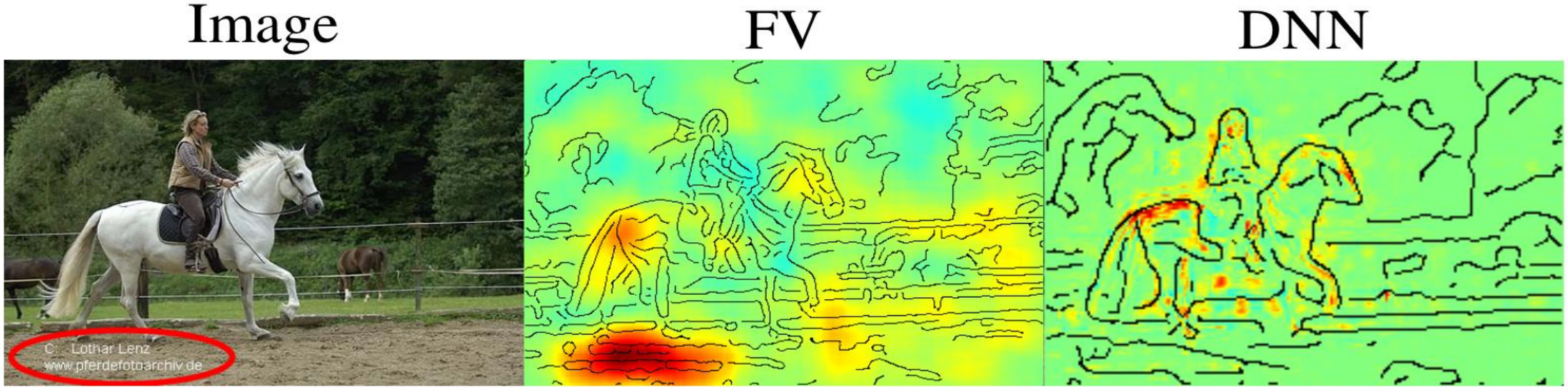
[Samek et al. IEEE TNNLS 2017]

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

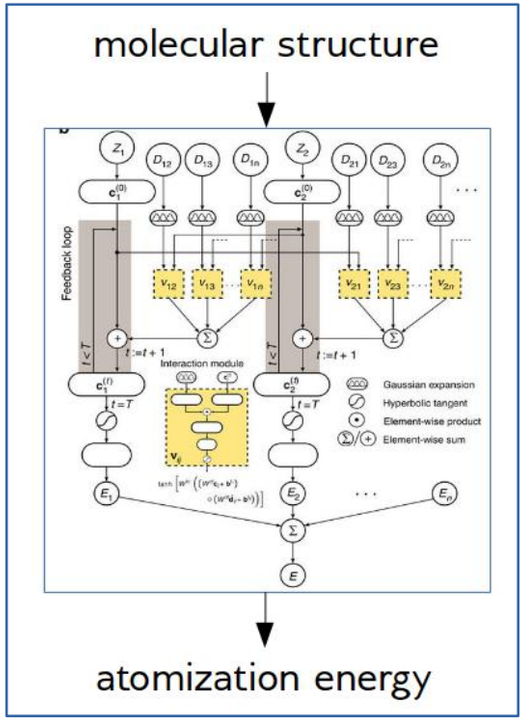
Example: Understanding physical systems at the quantum level.

time-independent
Schrödinger Equation

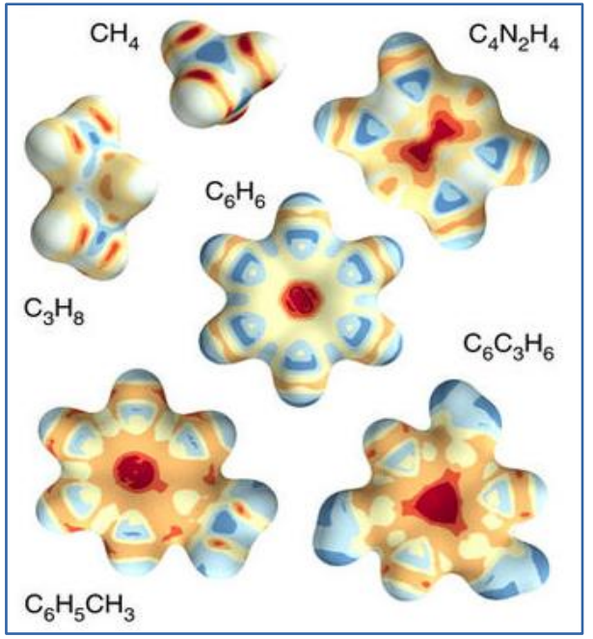
$$\hat{H}\Psi = E\Psi$$

Hamiltonian energy

equation describing
general physical systems



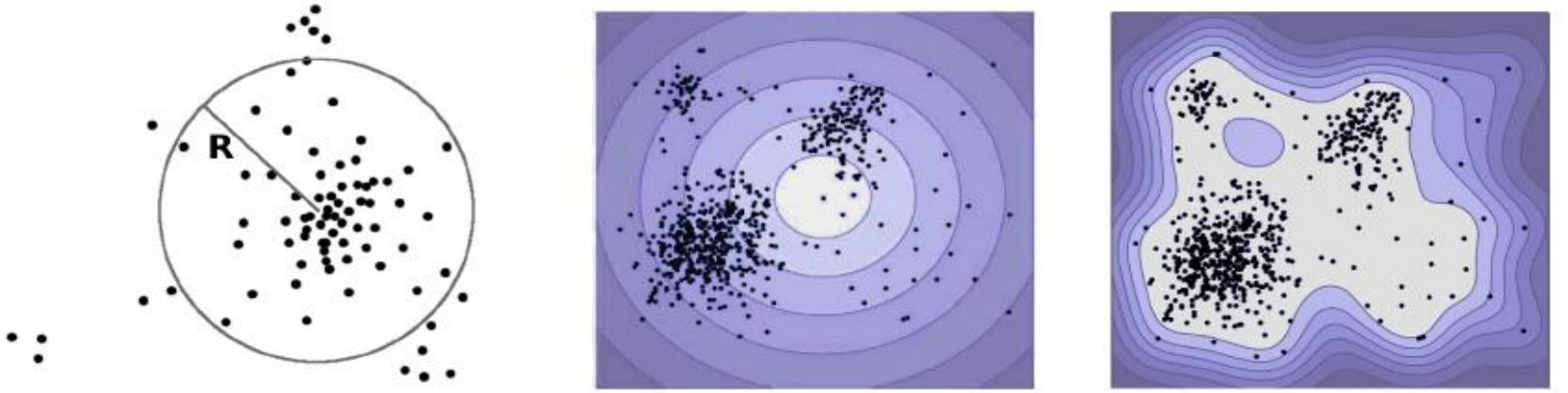
DNN approximation
for organic molecules



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

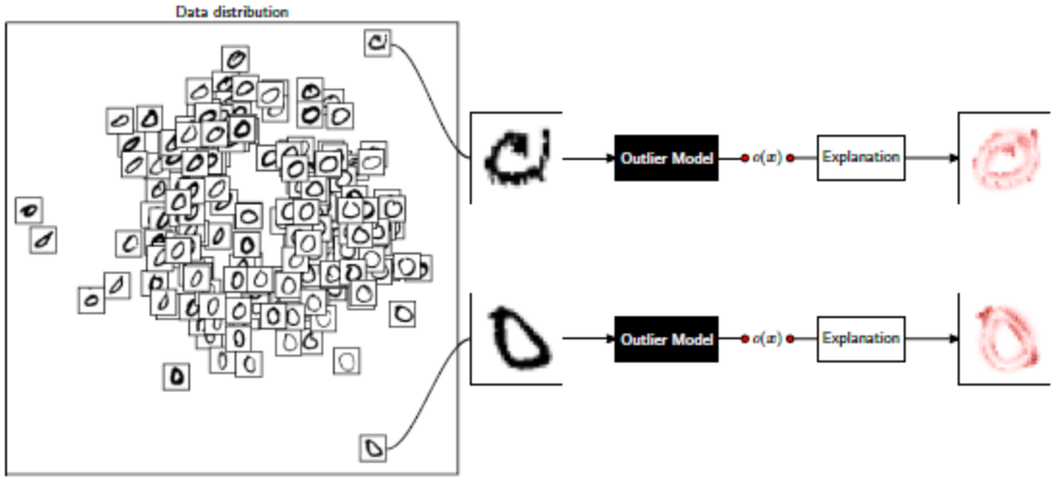


Figure 1: Illustration of the outlier detection and explanation setting. *Left:* Data is generated from an unknown distribution, we are for example interested in potential outliers; *Middle:* Unsupervised machine learning techniques estimate the data generating distribution and assign an outlier score $o(x)$ 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.

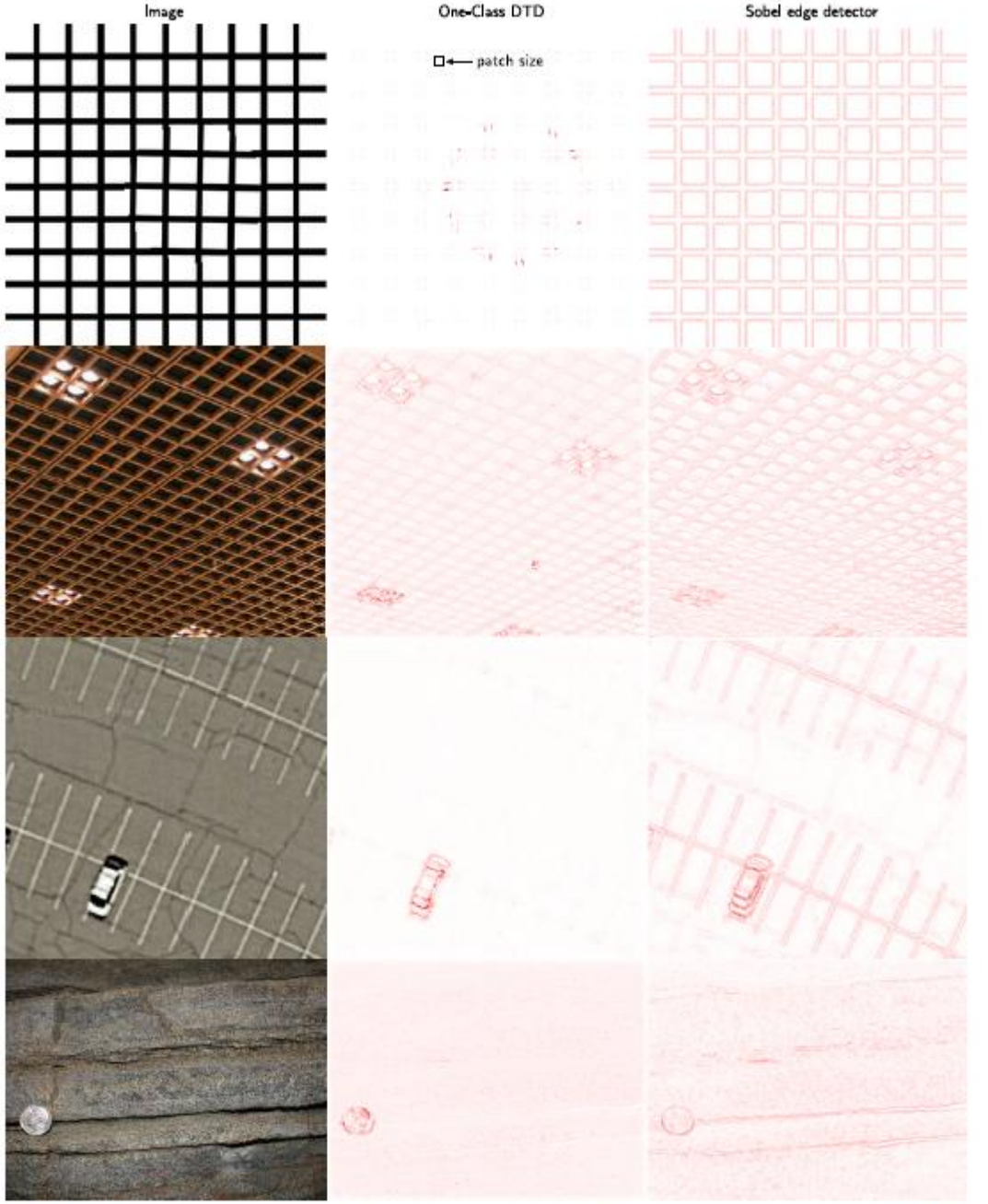


Figure 5: A One-Class SVM is trained on small 7×7 patches of the very image itself. Parameter $\nu = 0.1$ is set to allow at most 10% outliers. Images from a texture data set [11] (row one, two and four) and PatternNet [61]; top image is altered by us. For every image, we show *Left:* input image; *Middle*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

[Kaufmann, Müller, Montavon 2018]

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

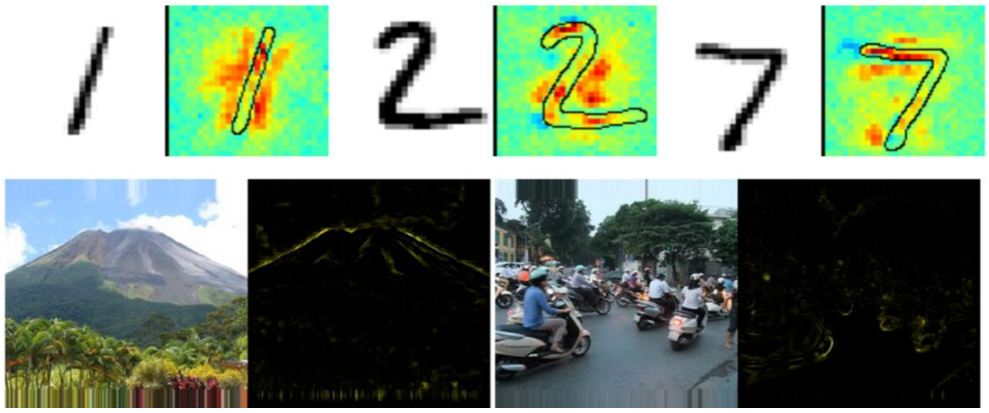
www.heatmapping.org

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

New Book: Samek, Montavon, Vedaldi, Hansen, Müller (eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNAI 11700, Springer (2019) (**coming up in 1 month**)

Keras Explanation Toolbox
https://github.com/albermax/innvestigate

Wojciech Samek · Grégoire Montavon · Andrea Vedaldi · Lars Kai Hansen · Klaus-Robert Müller (Eds.)

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning

Springer

State-of-the-Art Survey

LNAI 11700

Samek et al. (Eds.)

LNAI 11700

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning

BBDC

BERLIN BIG DATA CENTER

Further Reading I

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.R., Dähne, S. and Kindermans, P.J., 2019. iNNvestigate neural networks!. Journal of Machine Learning Research, 20(93), pp.1-8.

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).

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. The Journal of Machine Learning Research, 11, 1803-1831.

Binder et al. Machine Learning for morpho-molecular Integration, *arXiv:1805.11178* (2018)

Blankertz, B., Curio, G. and Müller, K.R., 2002. Classifying single trial EEG: Towards brain computer interfacing. In Advances in neural information processing systems (pp. 157-164).

Blankertz, B., Dornhege, G., Krauledat, M., Müller, K.R. and Curio, G., 2007. The non-invasive Berlin brain–computer interface: fast acquisition of effective performance in untrained subjects. NeuroImage, 37(2), pp.539-550.

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M. and Muller, K.R., 2007. Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal processing magazine, 25(1), pp.41-56.

Blankertz, B., Lemm, S., Treder, M., Haufe, S. and Müller, K.R., 2011. Single-trial analysis and classification of ERP components—a tutorial. NeuroImage, 56(2), pp.814-825.

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.

Brockherde, F., Vogt, L., Li, L., Tuckerman, M.E., Burke, K. and Müller, K.R., 2017. Bypassing the Kohn-Sham equations with machine learning. Nature communications, 8(1), p.872.

Further Reading II

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.

Chmiela, S., Sauceda, H.E., Müller, K.R. and Tkatchenko, A., 2018. Towards exact molecular dynamics simulations with machine-learned force fields. Nature communications, 9(1), p.3887.

Dornhege, G., Millan, J.D.R., Hinterberger, T., McFarland, D.J. and Müller, K.R. eds., 2007. Toward brain-computer interfacing. MIT press.

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.

Horst, F., Lapuschkin, S., Samek, W., Müller, K.R. and Schöllhorn, W.I., 2019. Explaining the unique nature of individual gait patterns with deep learning. Scientific reports, 9(1), p.2391.

Kauffmann, J., Müller, K.R. and Montavon, G., 2018. Towards explaining anomalies: A deep taylor decomposition of one-class models. arXiv preprint arXiv:1805.06230.

Further Reading III

Klauschen, F., Müller, K.R., Binder, A., Bockmayr, M., Hägele, M., Seegerer, P., Wienert, S., Pruneri, G., de Maria, S., Badve, S. and Michiels, S., 2018, October. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. Seminars in cancer biology (Vol. 52, pp. 151-157).

Lemm, S., Blankertz, B., Dickhaus, T. and Müller, K.R., 2011. Introduction to machine learning for brain imaging. Neuroimage, 56(2), pp.387-399.

Lapuschkin, 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).

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10(1), p.1096.

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.

Muller, K.R., Anderson, C.W. and Birch, G.E., 2003. Linear and nonlinear methods for brain-computer interfaces. IEEE transactions on neural systems and rehabilitation engineering, 11(2), pp.165-169.

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., Orr, G. & Müller, K. R. (2012). Neural Networks: Tricks of the Trade, Springer LNCS 7700. Berlin Heidelberg.

Further Reading IV

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).

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)

Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.R. (eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. LNAI 11700, Springer (2019)

Sturm, I., Lapuschkin, S., Samek, W. and Müller, K.R., 2016. Interpretable deep neural networks for single-trial EEG classification. Journal of neuroscience methods, 274, pp.141-145.