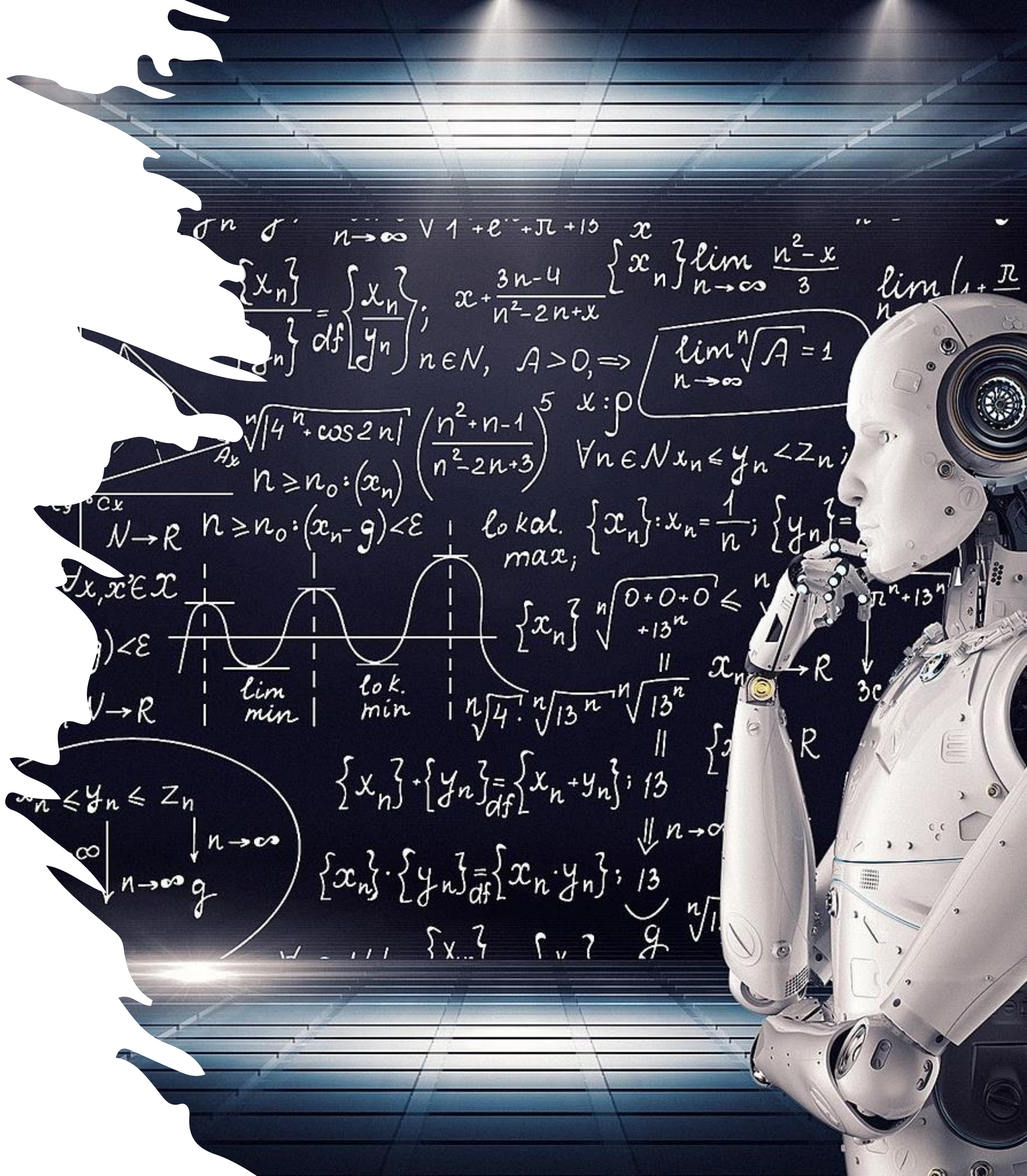
UNIVERSITY OF WARSAW

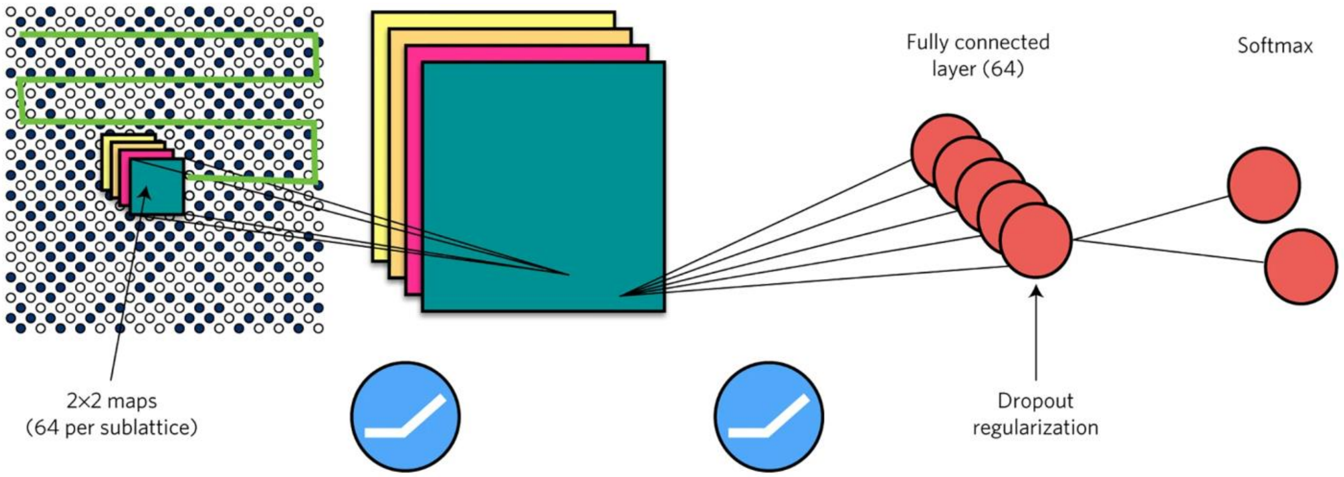
ICFO

Towards interpretable and reliable machines learning physics

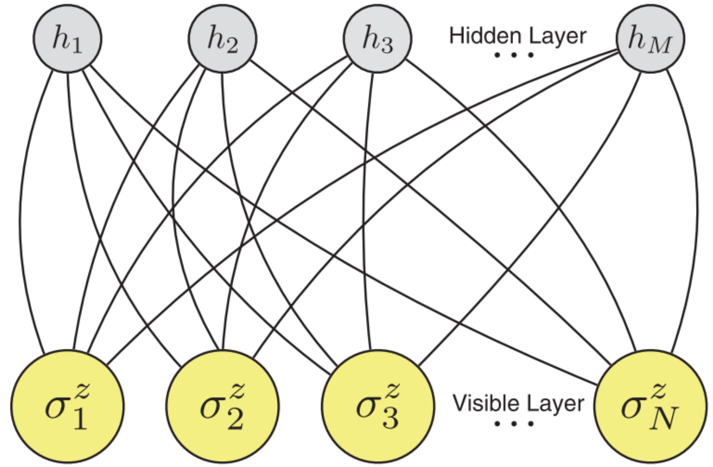
Anna Dawid

Mach. Learn.: Sci. Technol. 3, 015002 (2022)

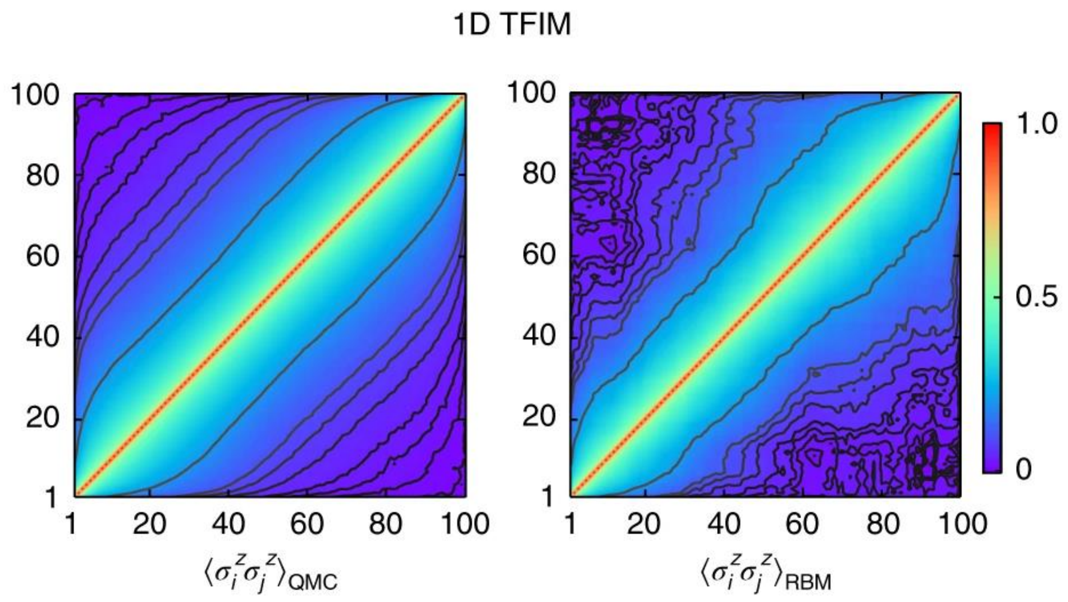




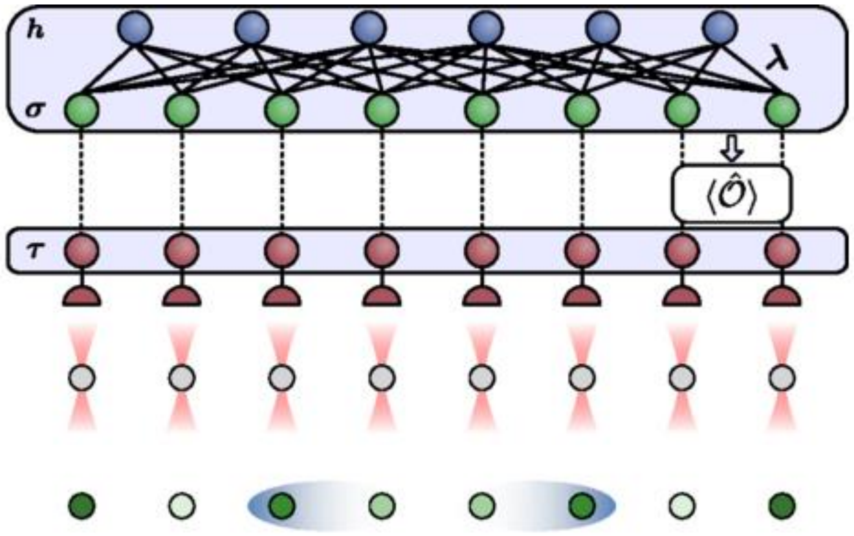
Carrasquilla & Melko, Nat. Phys. **13**, 431-434 (2017)



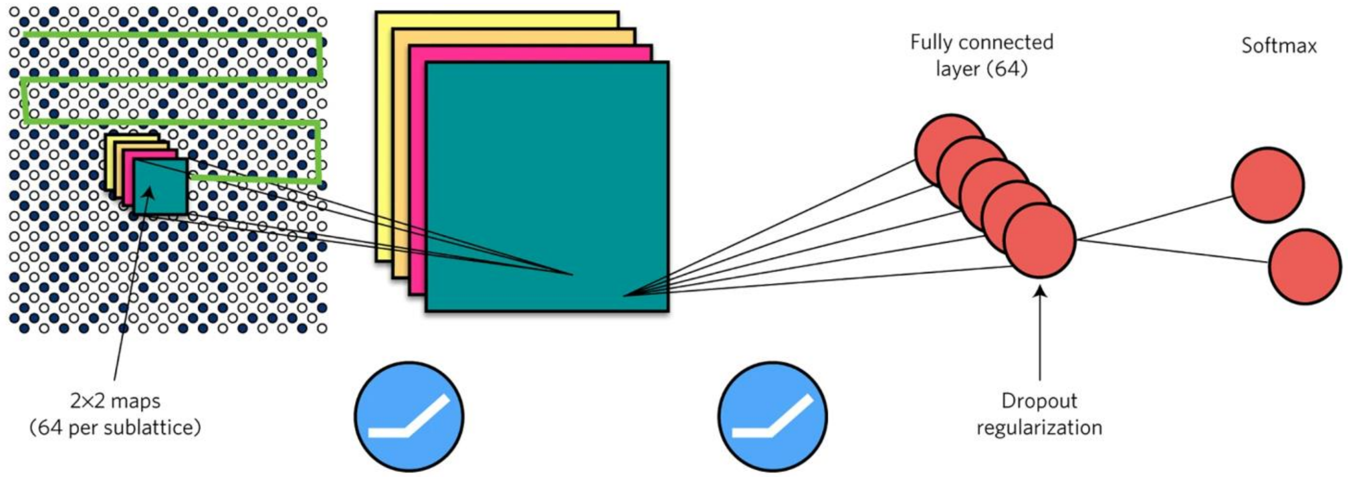
Carleo & Troyer, Science **355**, 602-606, (2017)



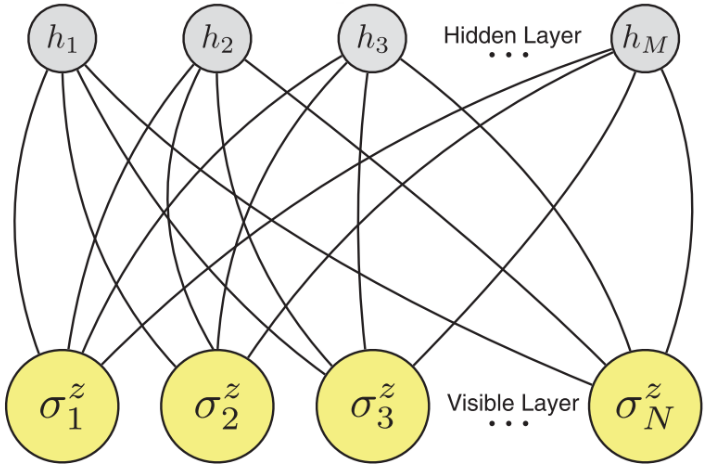
Torlai et al., Nat. Phys. **14**, 447–450 (2018)



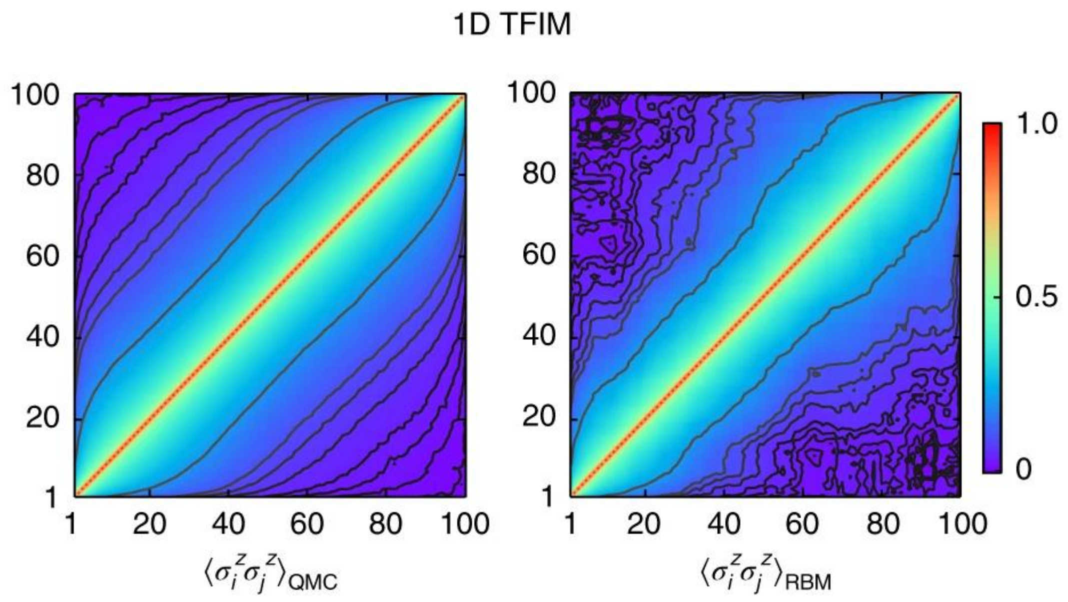
Torlai et. al., Phys. Rev. Lett. **123**, 230504 (2019)



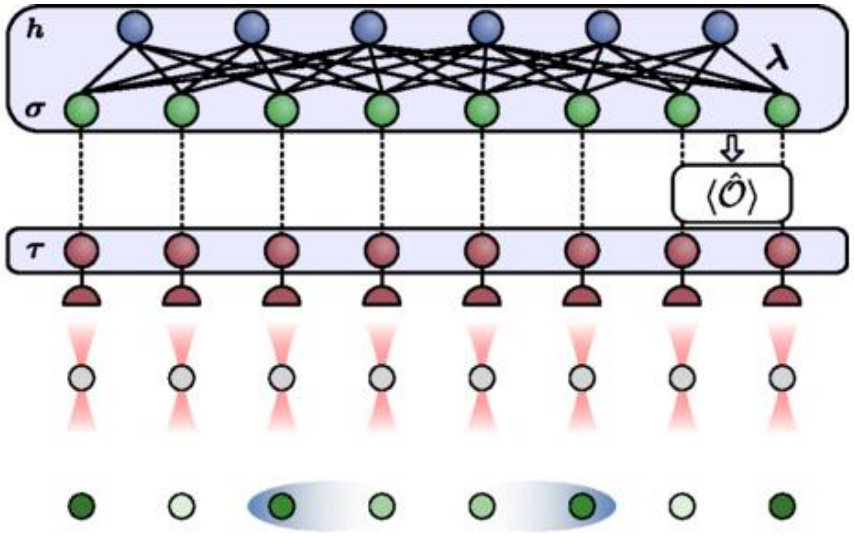
Carrasquilla & Melko, Nat. Phys. **13**, 431-434 (2017)



Carleo & Troyer, Science **355**, 602-606, (2017)



Torlai et al., Nat. Phys. **14**, 447–450 (2018)



Torlai et. al., Phys. Rev. Lett. **123**, 230504 (2019)

Machines in phase classification – open problems

quantum many-body localization

- disagreement of predicted critical exponents
- high sensitivity to hyperparameters describing the training process

topological phases of matter

- learning schemes trained on raw Monte Carlo configurations were found to be not effective
- pre-engineered features are often needed

mainly recovery of **known** results, but much cheaper

general problems with ML like…

Taken from: xkcd, A Webcomic of Romance, Sarcasm, Math, And Language, https://xkcd.com/1838/

People worry that the computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.

Pedro Domingos "The Master Algorithm"

! even small invisible changes or a different background context can completely derail predictions

high error rates for faces from minority groups **!**

! the algorithm's hiring and insurance decisions are biased towards selecting men and white people

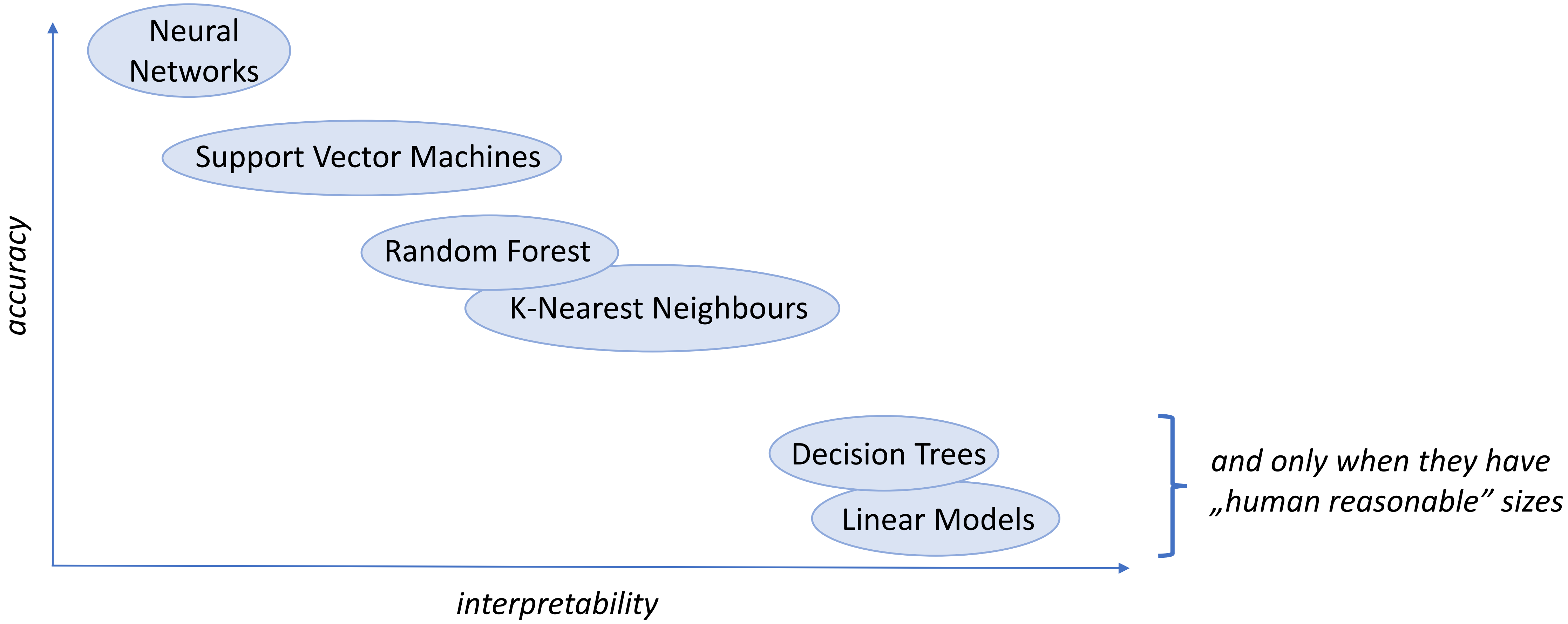
Some definitions

Interpretability
understanding what an ML model learns and how it makes its predictions

Reliability
trusting our ML model predictions (uncertainty)

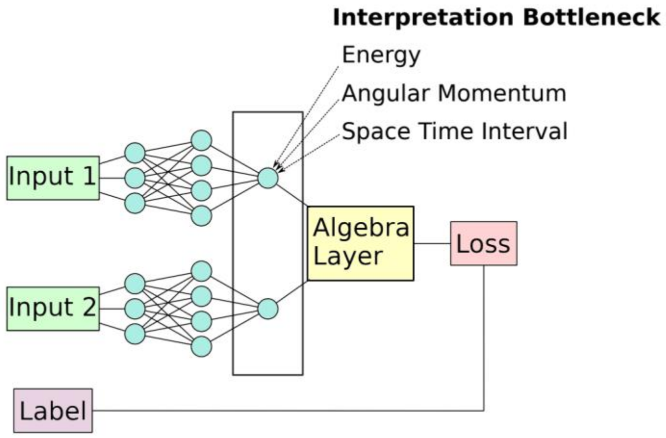
These two properties are closely intertwined.

Trade-off between complexity and interpretability

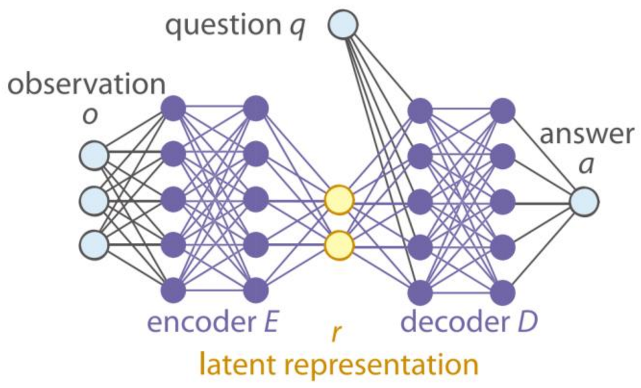


Interpretation of ML in physics so far

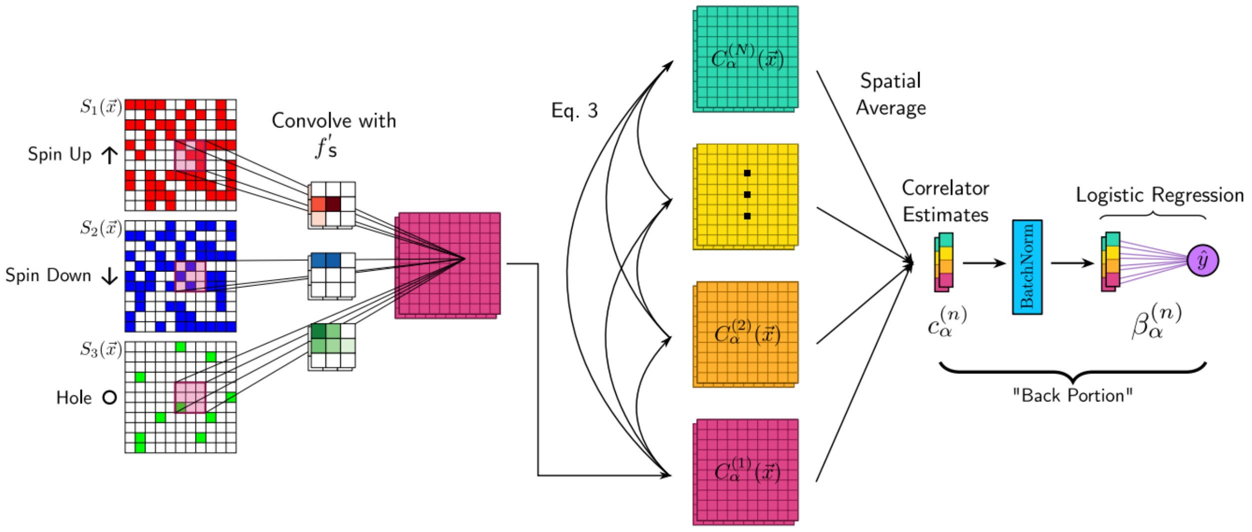
- Decision trees, kernel methods
- Bottleneck analysis



Phys. Rev. Research 2, 033499 (2020)



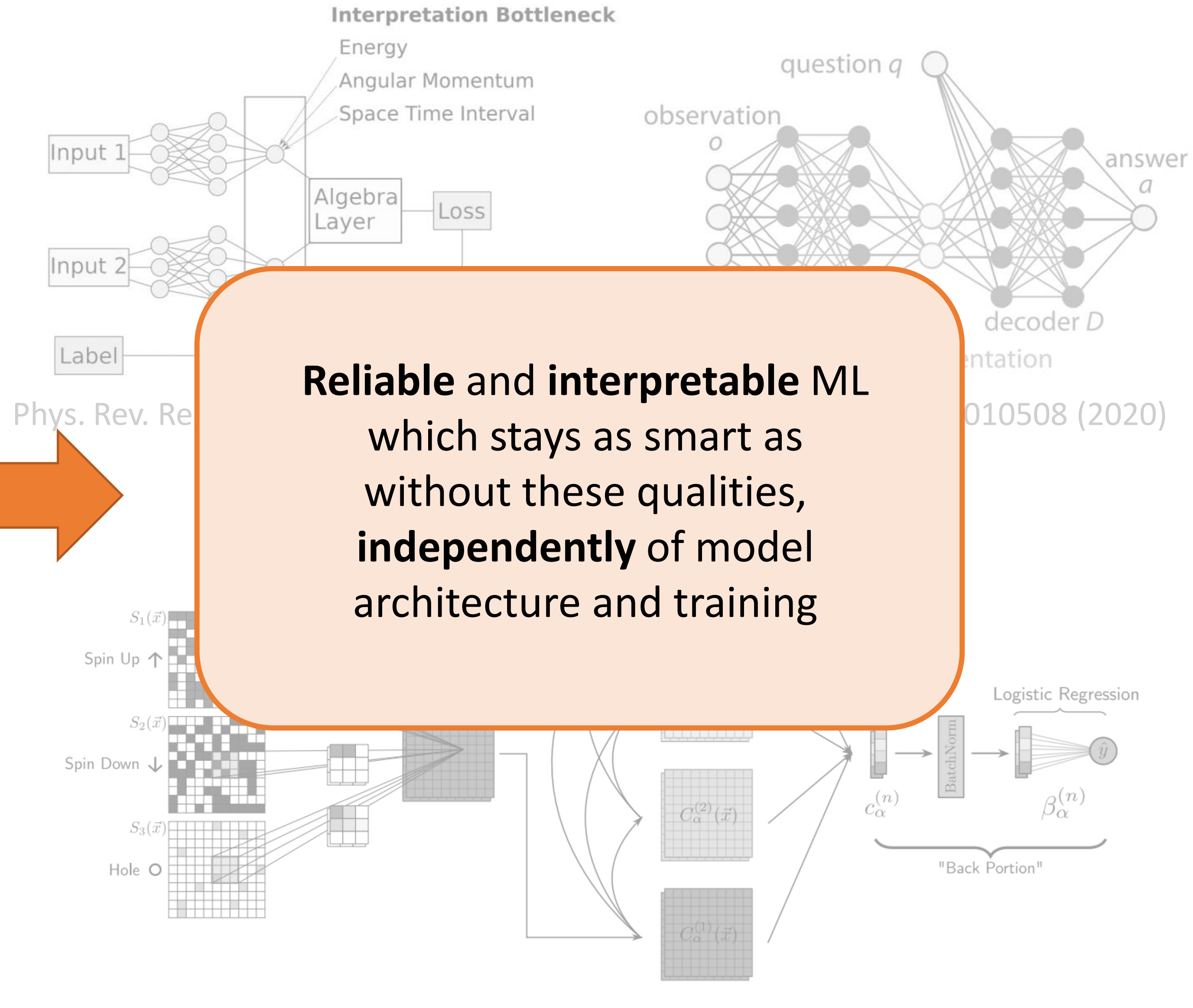
Phys. Rev. Lett. 124, 010508 (2020)



Nat. Commun. 12, 3905 (2021)

Interpretation of ML in physics so far

- Decision trees, kernel methods
- Bottleneck analysis

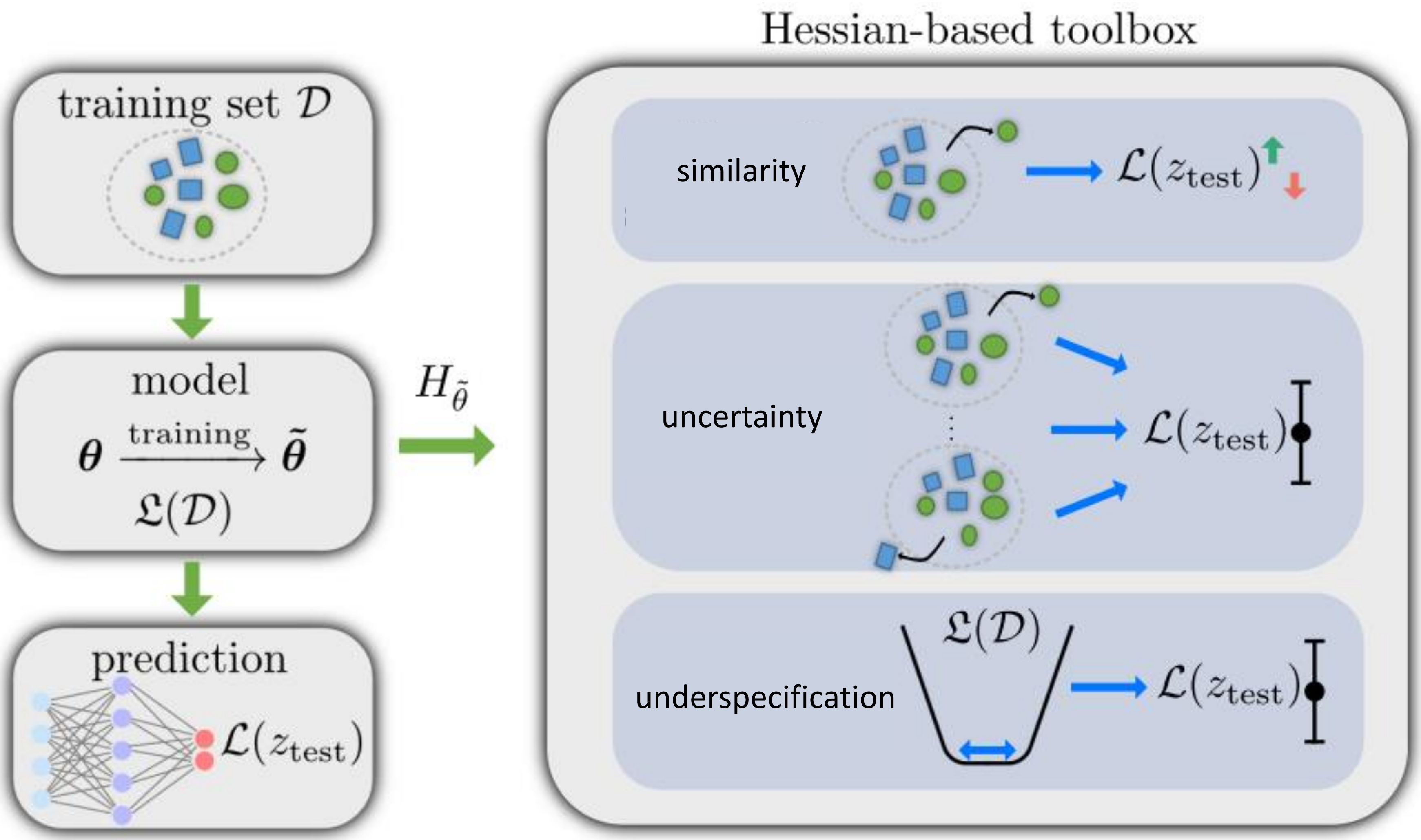


Phys. Rev. Re… …010508 (2020)

Nat. Commun. 12, 3905 (2021)

Reliable and **interpretable** ML which stays as smart as without these qualities, **independently** of model architecture and training

Hessian-based toolbox



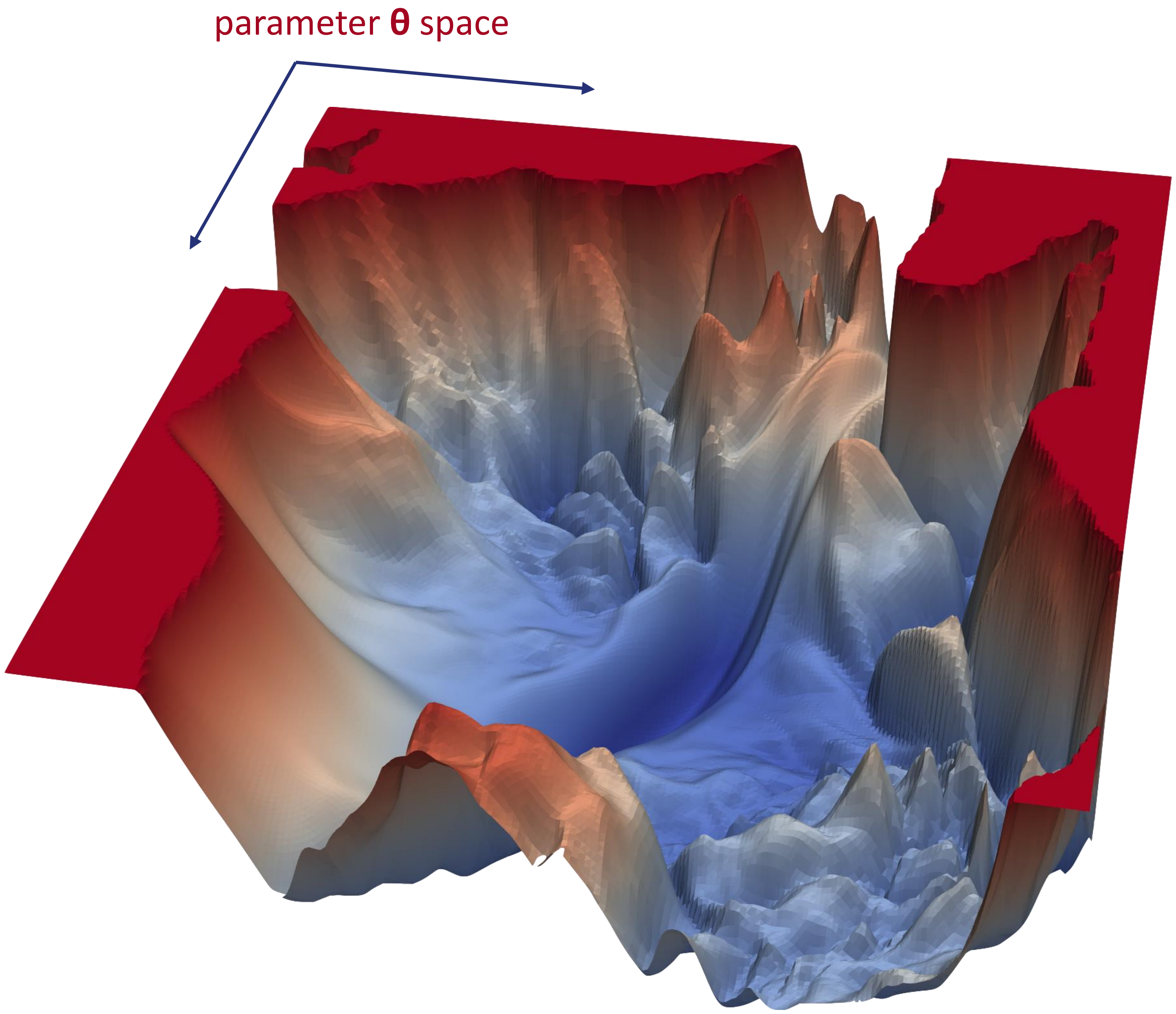
„Minimum” of ML loss landscape

(# of classes – 1): $\lambda_i > 0$

majority: $\lambda_i \approx 0$

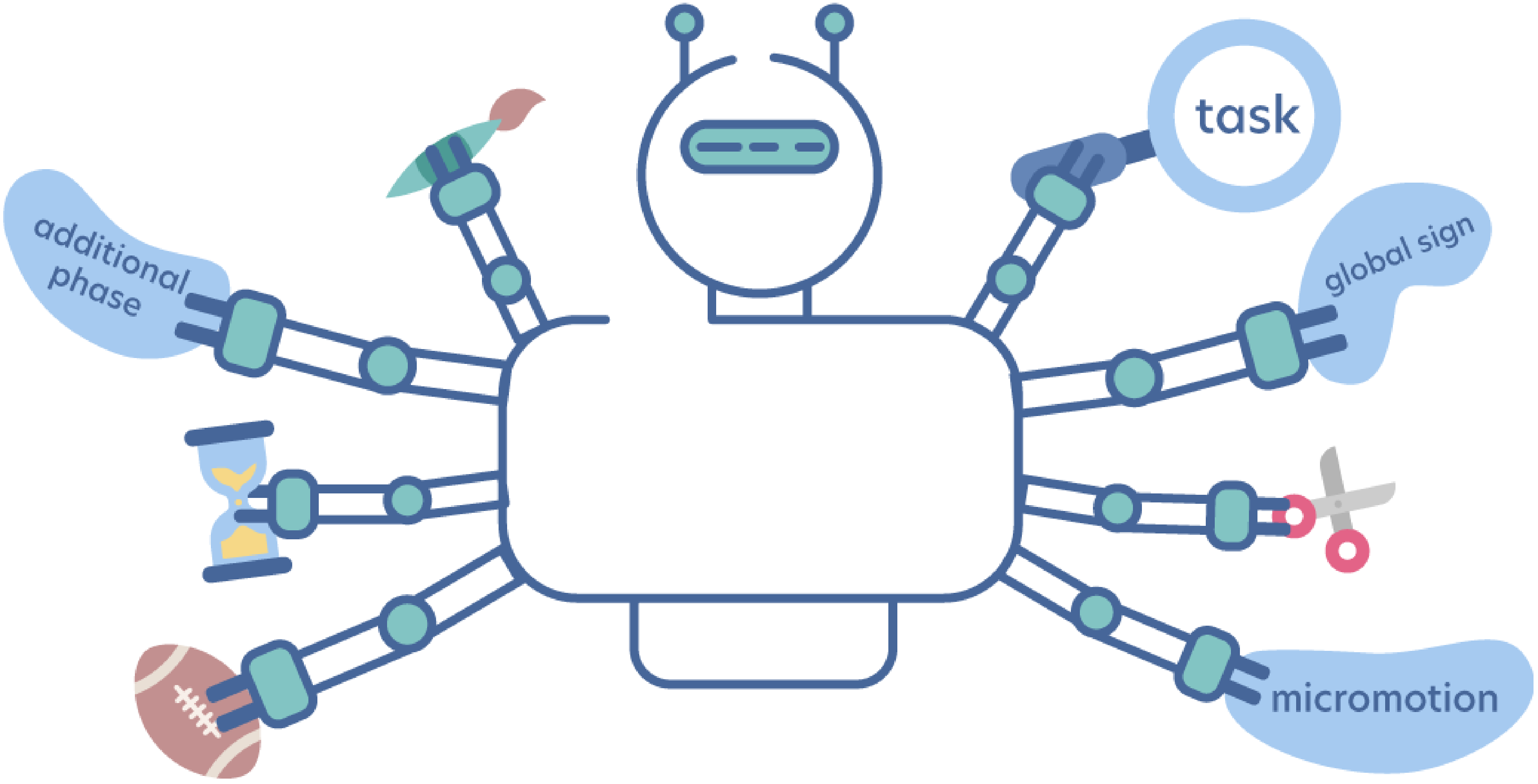
few: $\lambda_i < 0$

$$H_{ij} = \left. \frac{\partial^2 \mathcal{L}(D_{\text{train}}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}$$



Outline

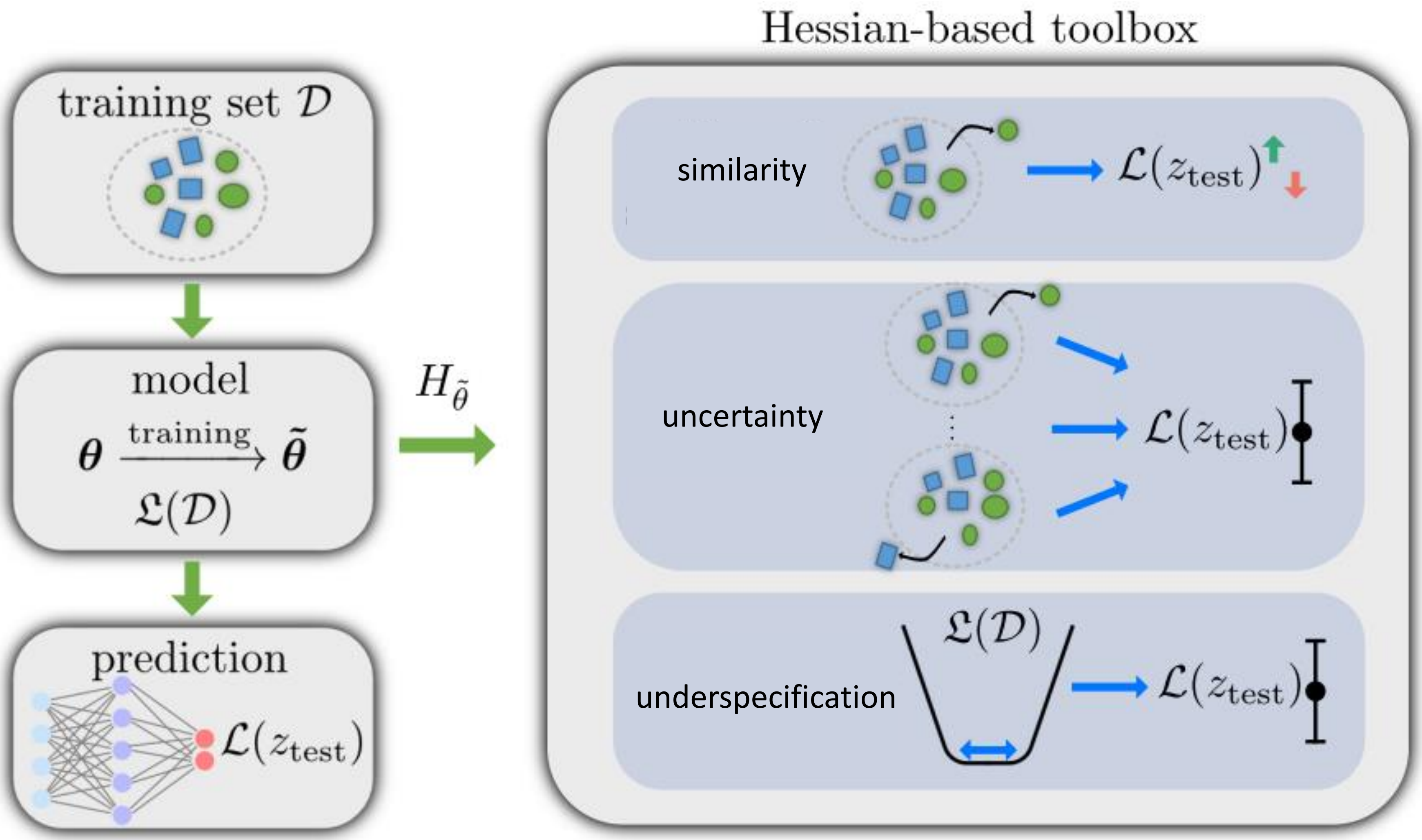
1. Interpreting an ML model
2. Reliability methods


additional phase
task
global sign
micromotion

Outline

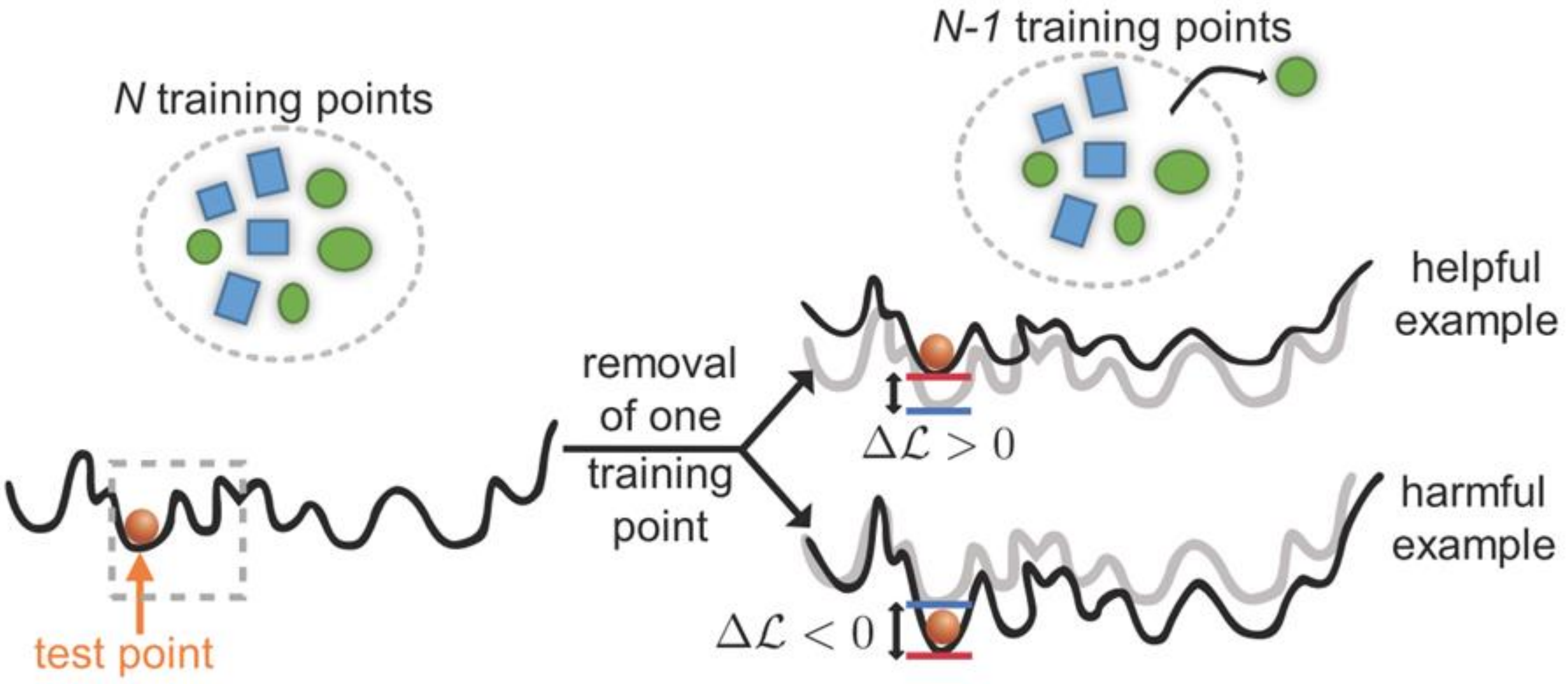
1. Interpreting an ML model
2. Reliability methods

Hessian-based toolbox

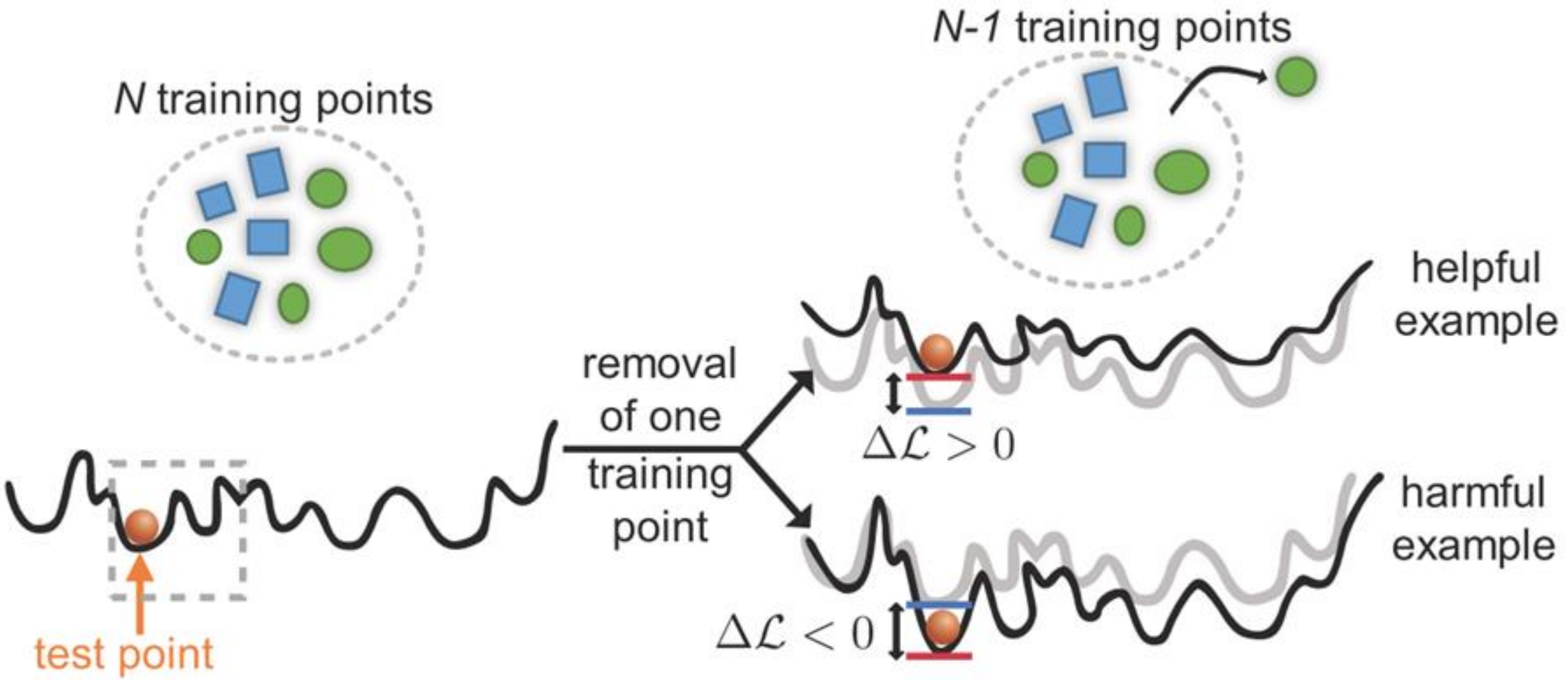


- Influence functions
  Koh & Liang
  arXiv:1703.04730
- Resampling Uncertainty Estimation (RUE)
  Schulam & Saria
  arXiv:1901.00403
- Local Ensembles (LEs)
  Madras, Atwood, D'Amour
  arXiv:1910.09573

Leave-one-out training

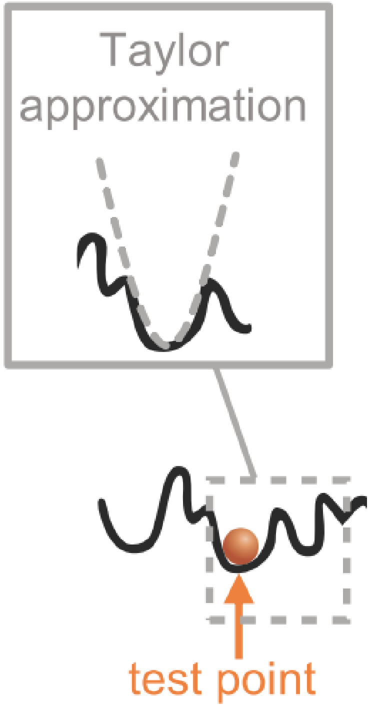


Leave-one-out training



prohibitively expensive!

Influence functions



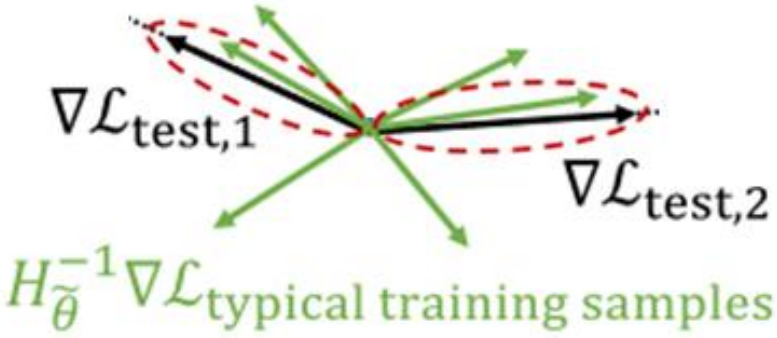
Analytical approximation for leave-one-out training

$$\mathcal{I}(z_{\mathrm{r}}, z_{\mathrm{test}}) = \frac{1}{n} \nabla_\theta \mathcal{L}(z_{\mathrm{test}}, \hat{\theta})^T \underbrace{H_\theta^{-1}(\hat{\theta}) \nabla_\theta \mathcal{L}(z_{\mathrm{r}}, \hat{\theta})}$$

approximated change in parameters due to removal of z_{r}

Assumption: Hessian is positive-definite.
Generalization to non-convex models was done by Koh & Liang: arXiv:1703.04730, ICML 2017's best paper

Geometrical interpretation



$$\mathcal{I}(z_{\text{r}}, z_{\text{test}}) = \frac{1}{n}\nabla_\theta\mathcal{L}(z_{\text{test}}, \hat{\theta})^T H_\theta^{-1}(\hat{\theta})\nabla_\theta\mathcal{L}(z_{\text{r}}, \hat{\theta})$$

it is a scalar product of gradient of a test point and the gradient of a training point, corrected by local curvature described by the Hessian

notion of similarity in the model internal representation!

Three messages

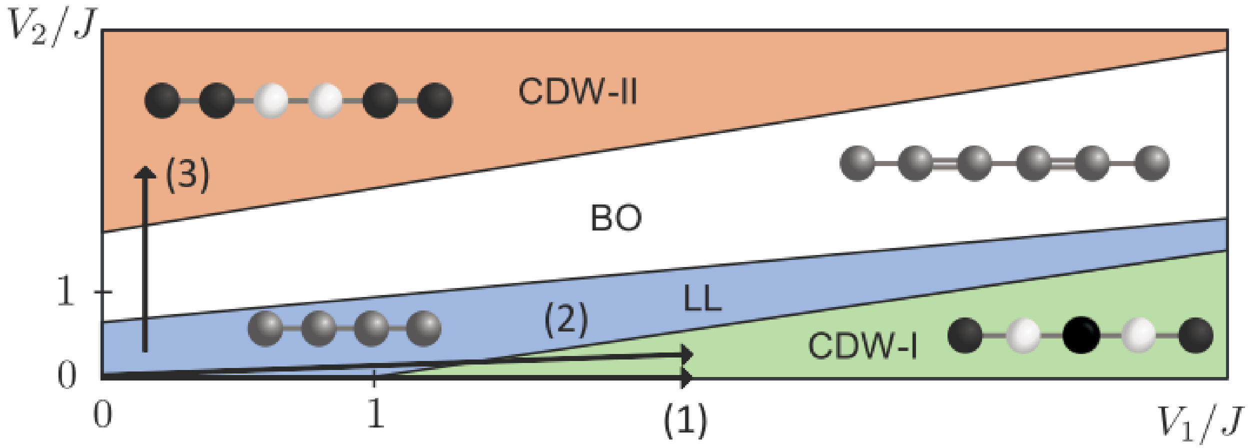
- Detection additional phases
- Detecting influential data features
- Anomaly detection with influence functions

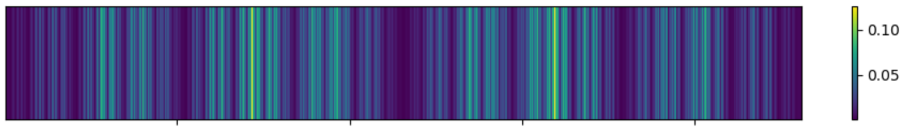
Physical input data

1) simulated

spinless 1D Fermi-Hubbard model at half-filling

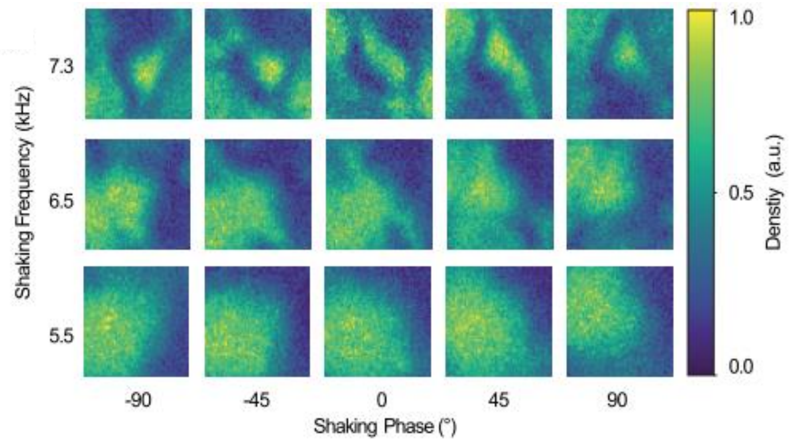
$$\hat{H} = -J \sum_{\langle i,j \rangle} c_i^\dagger c_j + V_1 \sum_{\langle i,j \rangle} n_i n_j + V_2 \sum_{\langle\langle i,j \rangle\rangle} n_i n_j$$

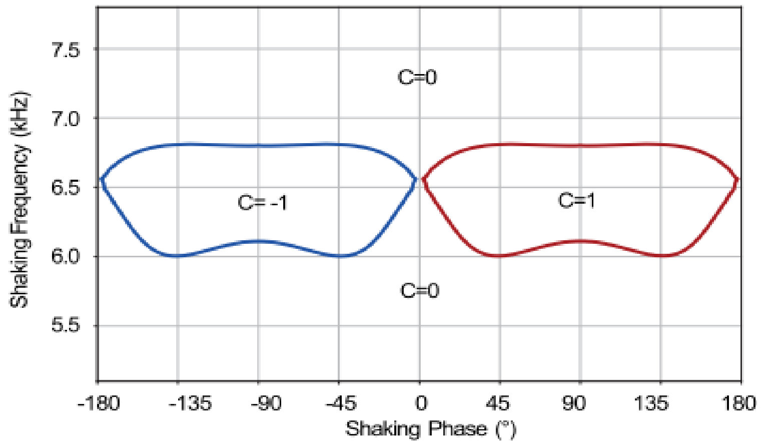




2) experimental

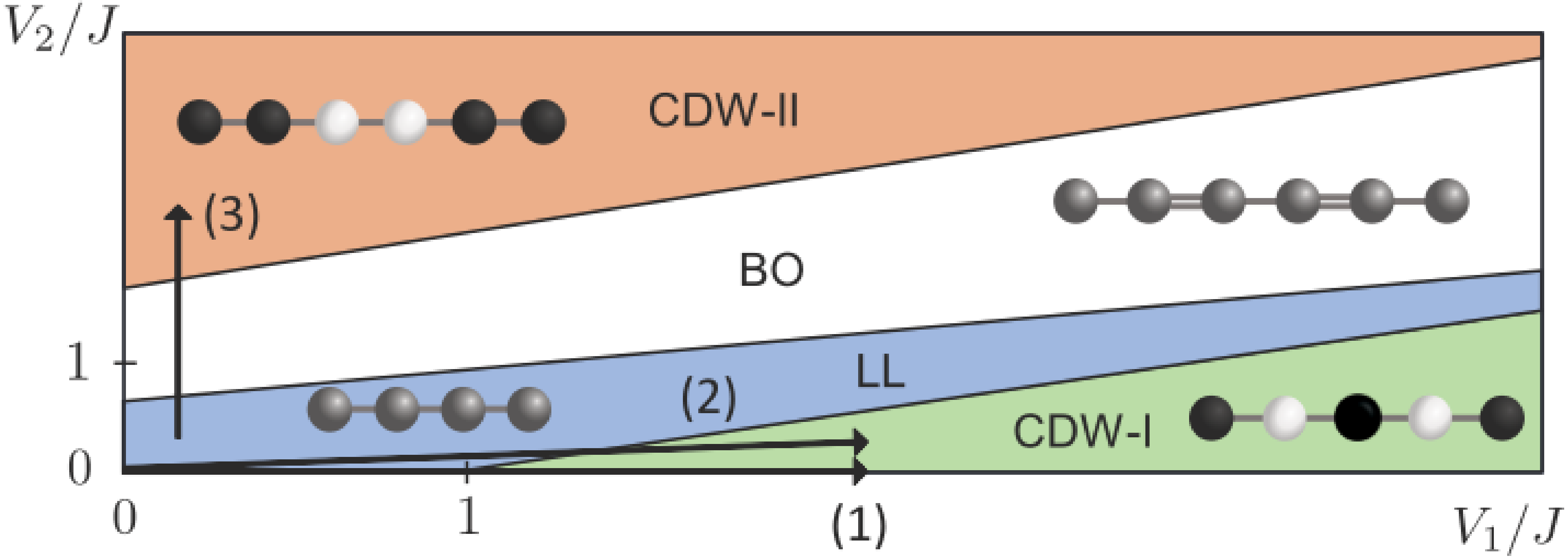
topological Haldane model

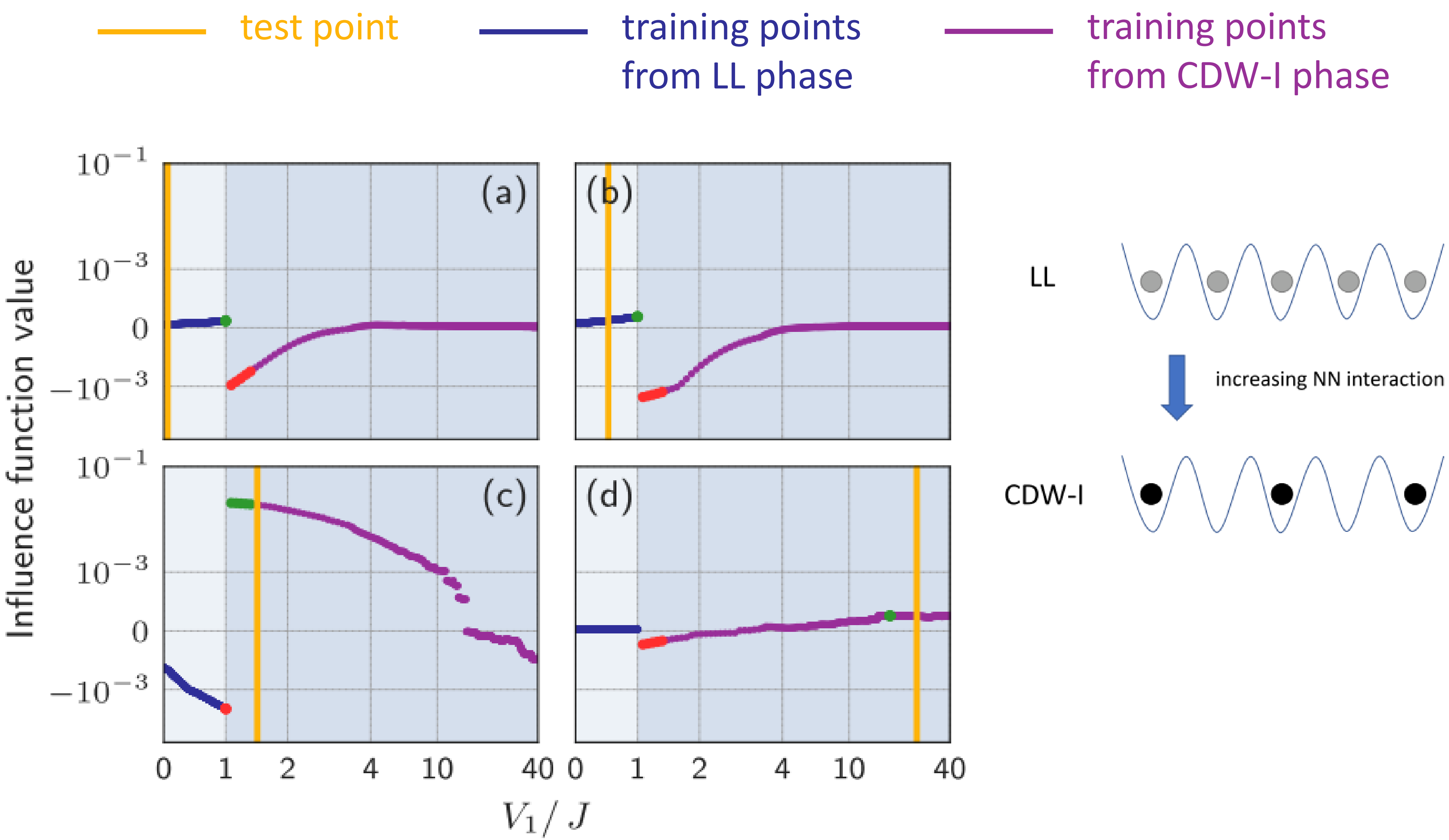




Three messages

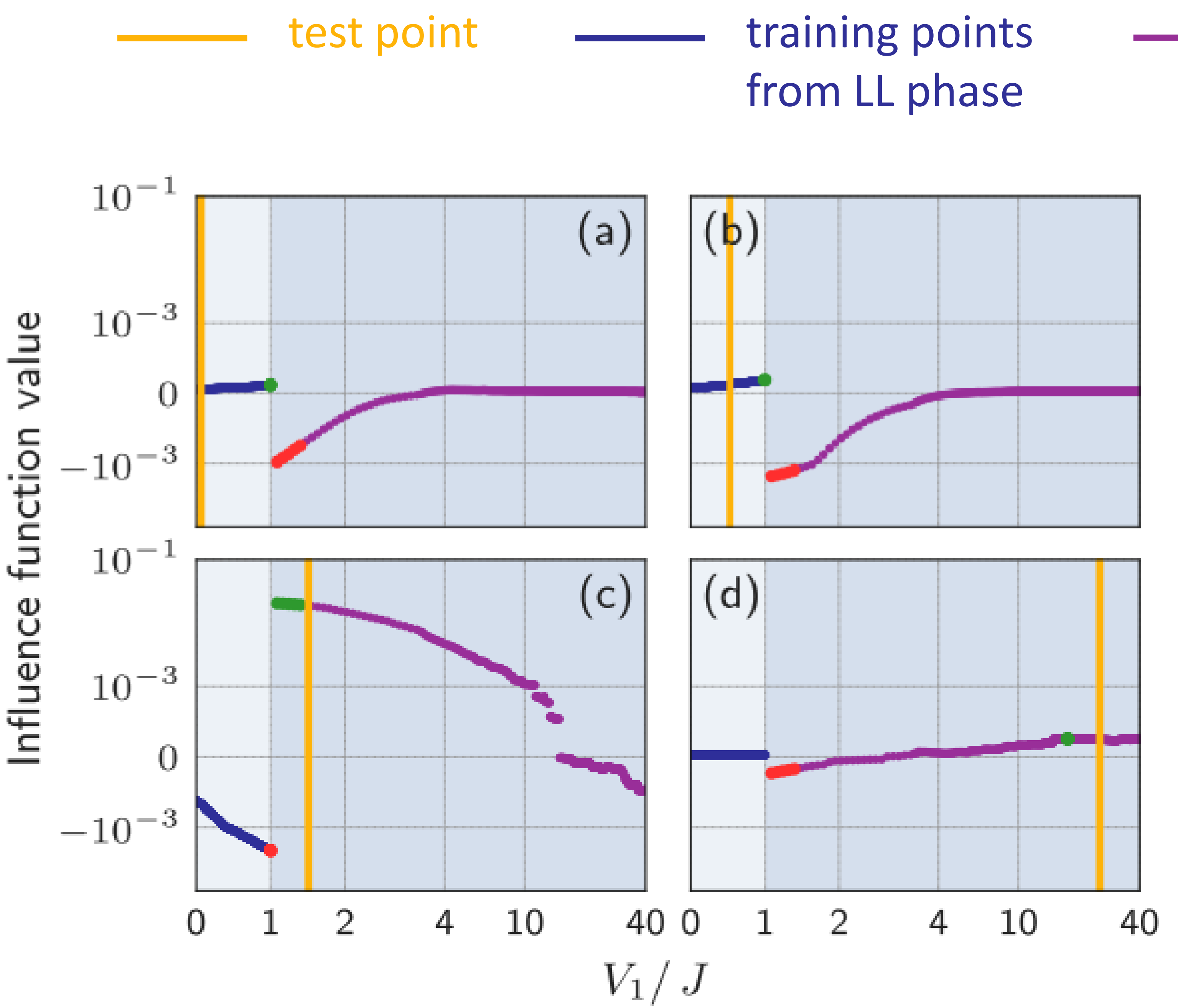
- Detection additional phases
- Detecting influential data features
- Anomaly detection with influence functions


V_2/J
CDW-II
(3)
BO
1
LL
(2)
CDW-I
0
0
1
(1)
V_1/J

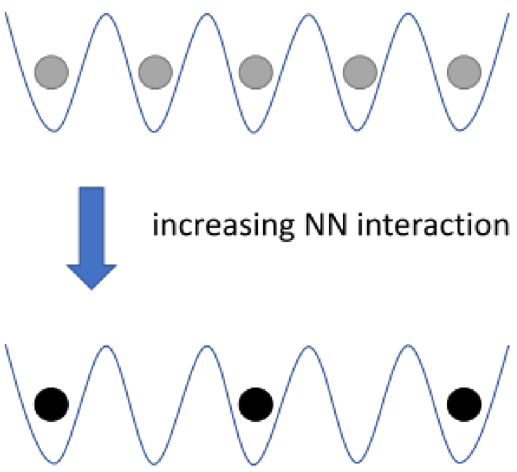


nearest neighbor interaction / hopping amplitude

A. Dawid et al, *New J. Phys.* **22** 115001 (2020)

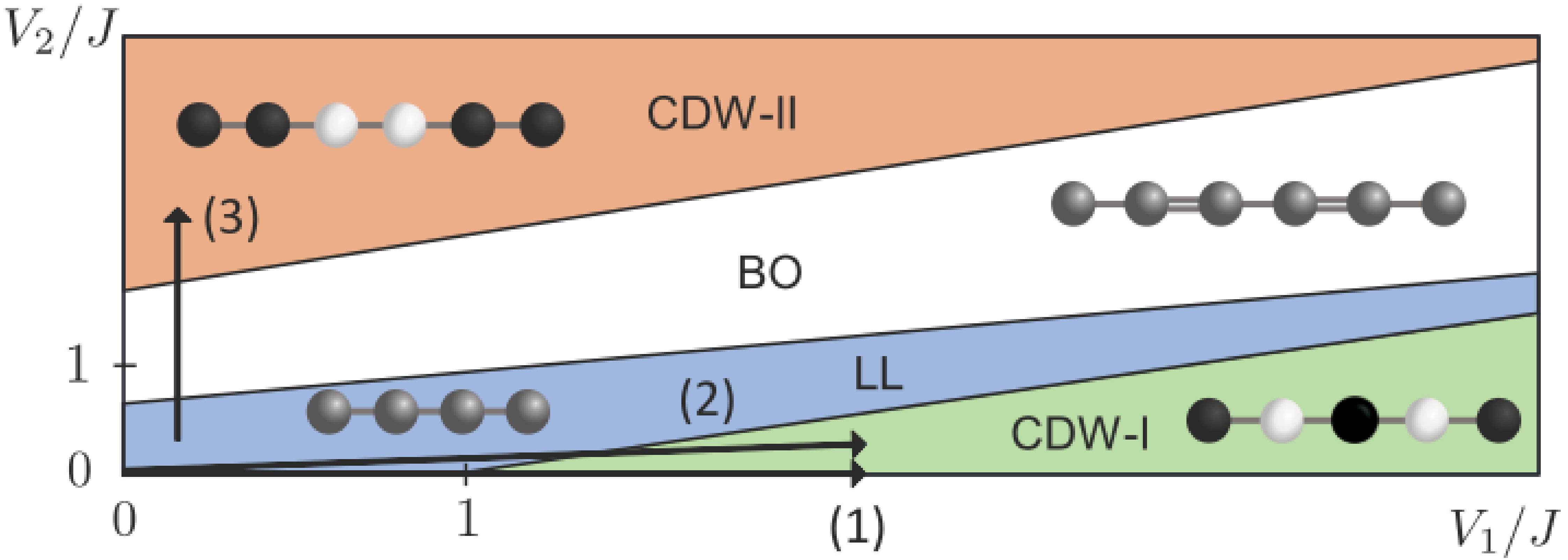


nearest neighbor interaction / hopping amplitude

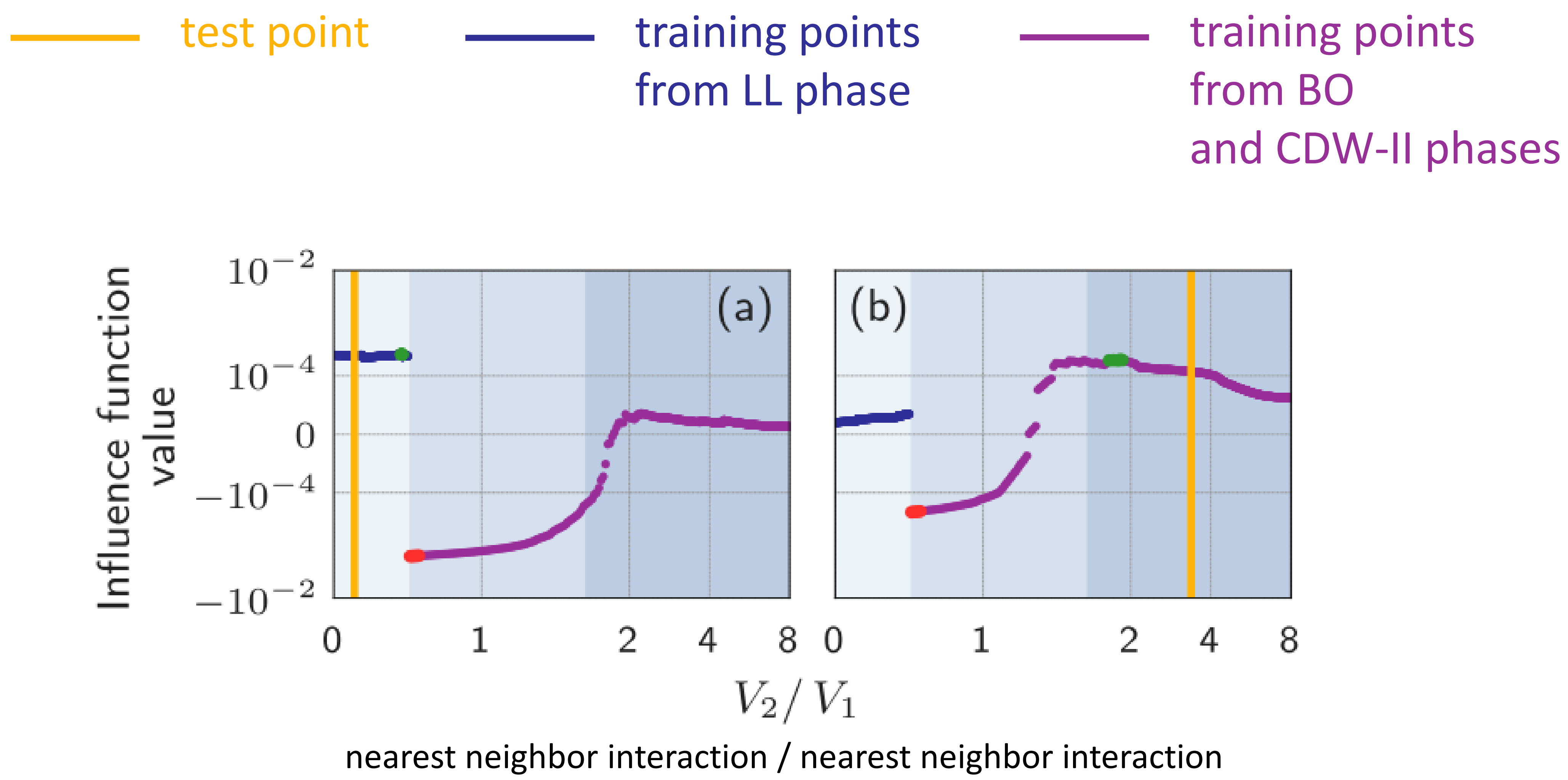


It learns sth related to order parameter!

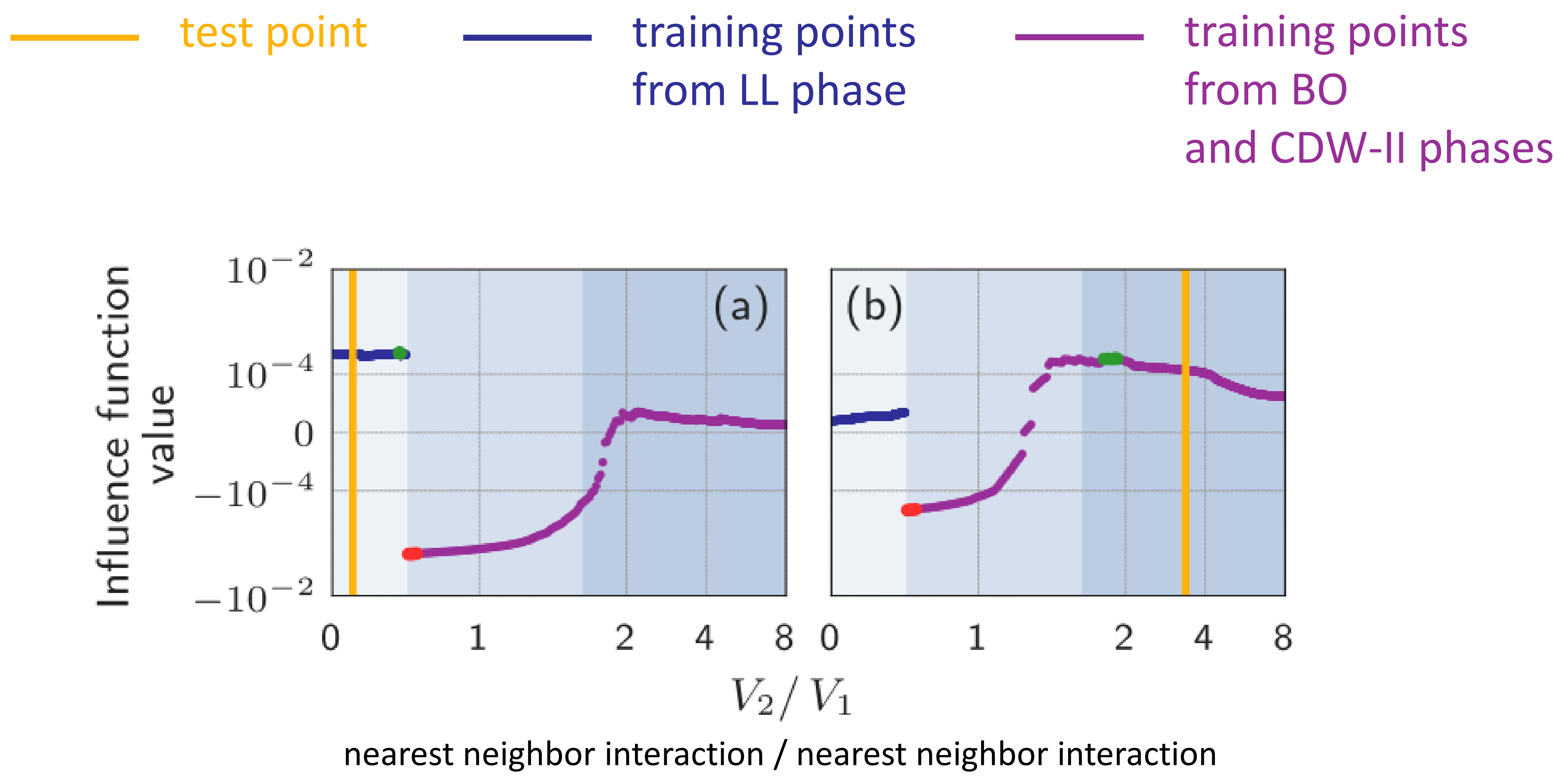
A. Dawid et al, *New J. Phys.* **22** 115001 (2020)



V_2/J
CDW-II
(3)
BO
1
LL
(2)
CDW-I
0
0
1
(1)
V_1/J



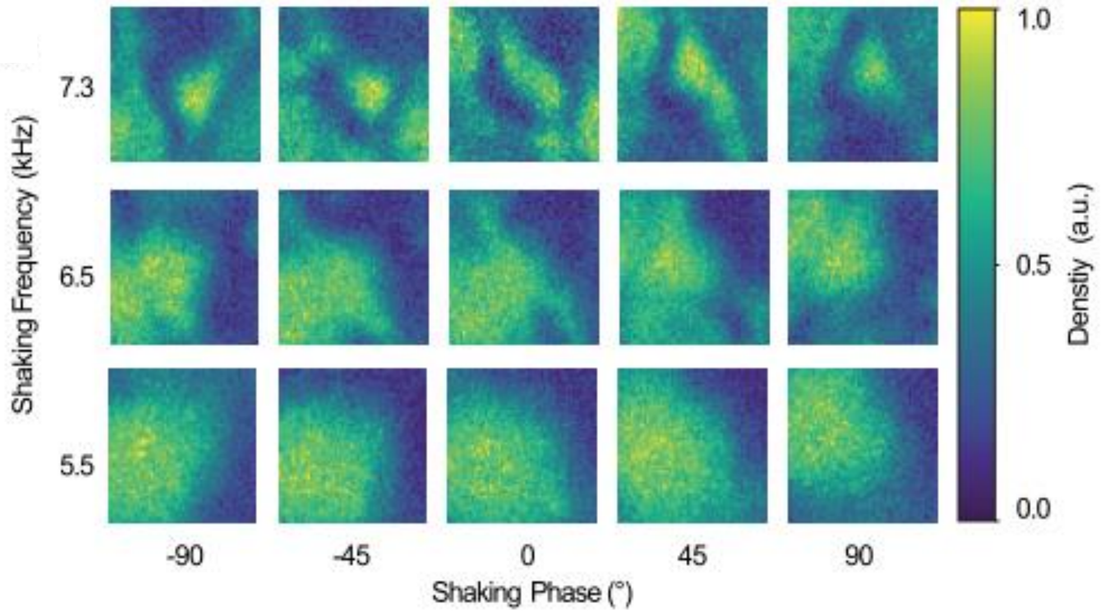
A. Dawid et al, *New J. Phys.* **22** 115001 (2020)

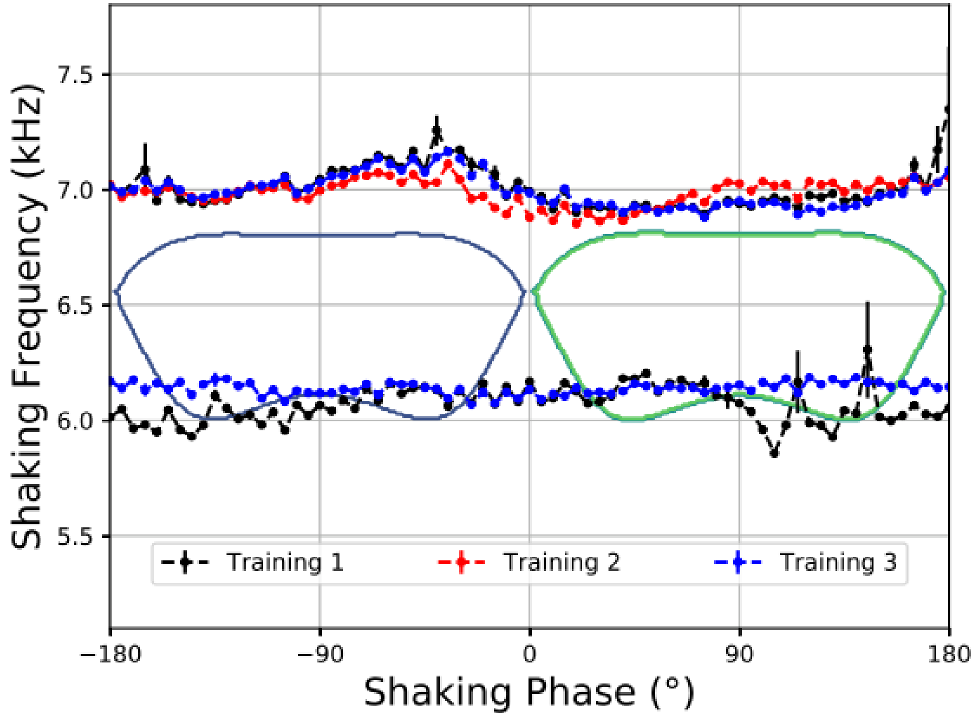


It sees additional phase!

A. Dawid et al, *New J. Phys.* **22** 115001 (2020)

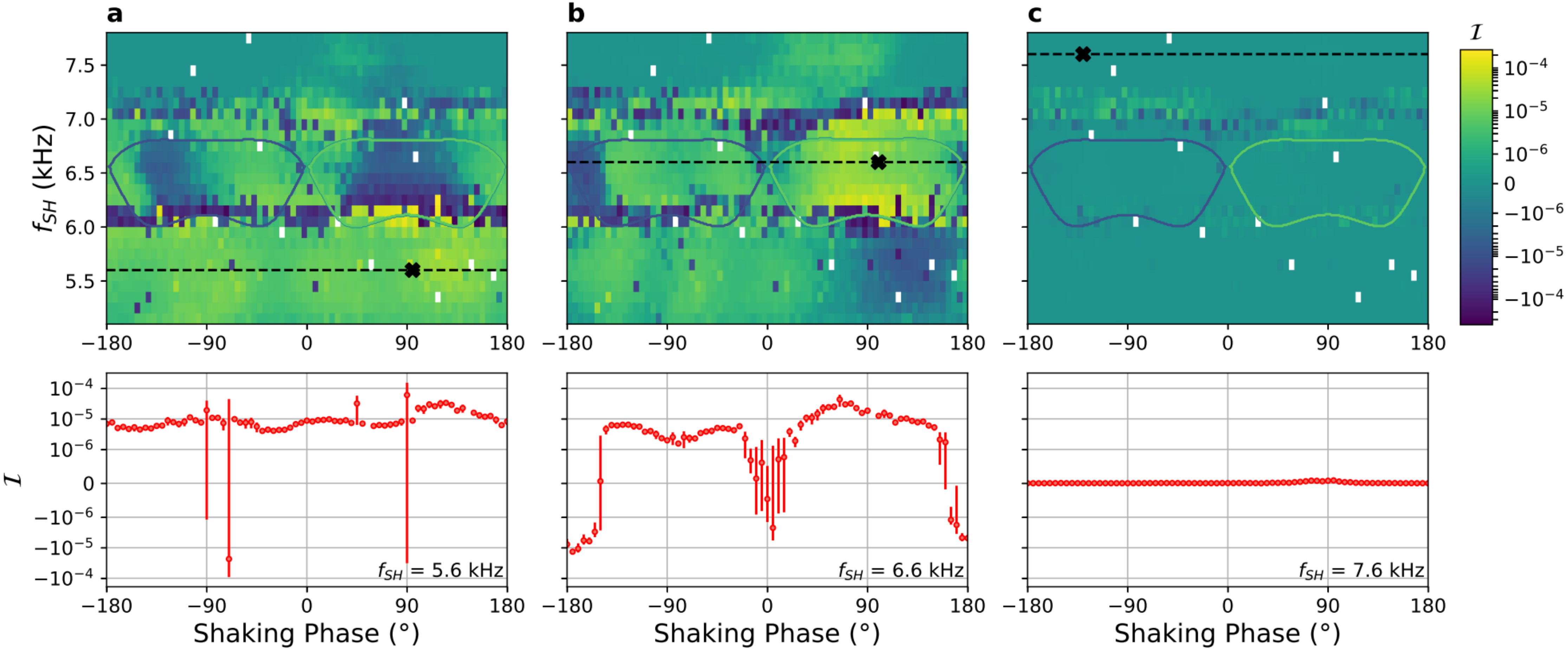
Unsupervised machine learning of topological phase transition from experimental data





unsupervised approaches had troubles with distinghuishing between two topological phases…

N. Käming, A. Dawid et al., *MLST* **2** 035037 (2021)



N. Käming, A. Dawid et al., *MLST* **2** 035037 (2021)

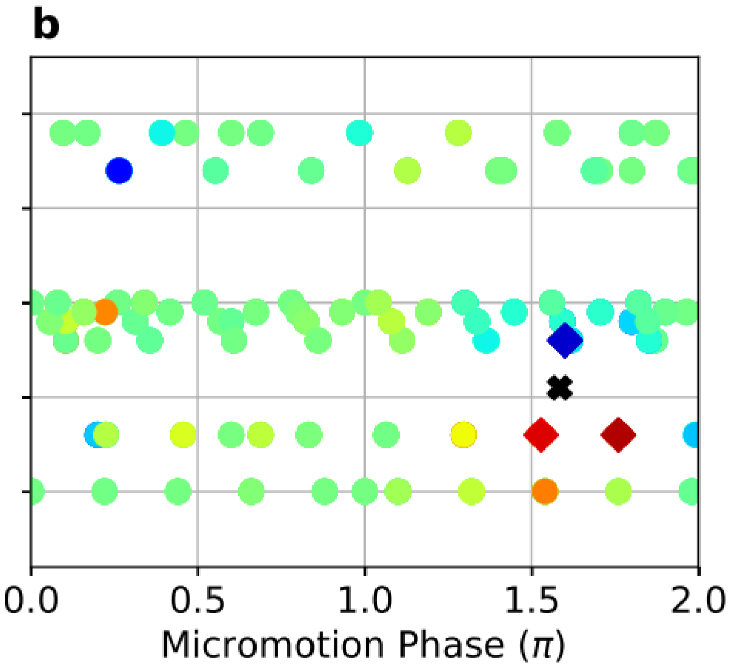
Three messages

Detection additional phases

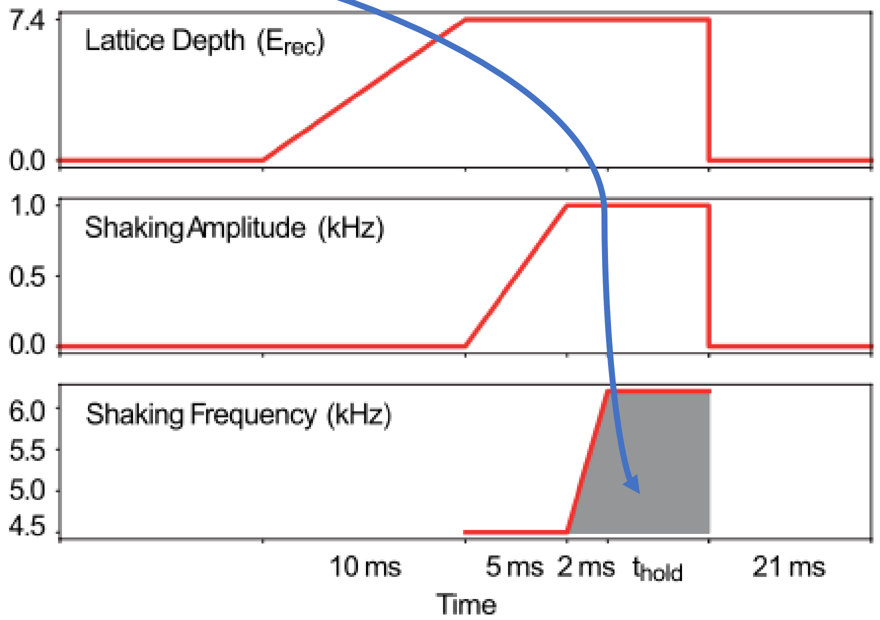
Detecting influential data features

Anomaly detection with influence functions

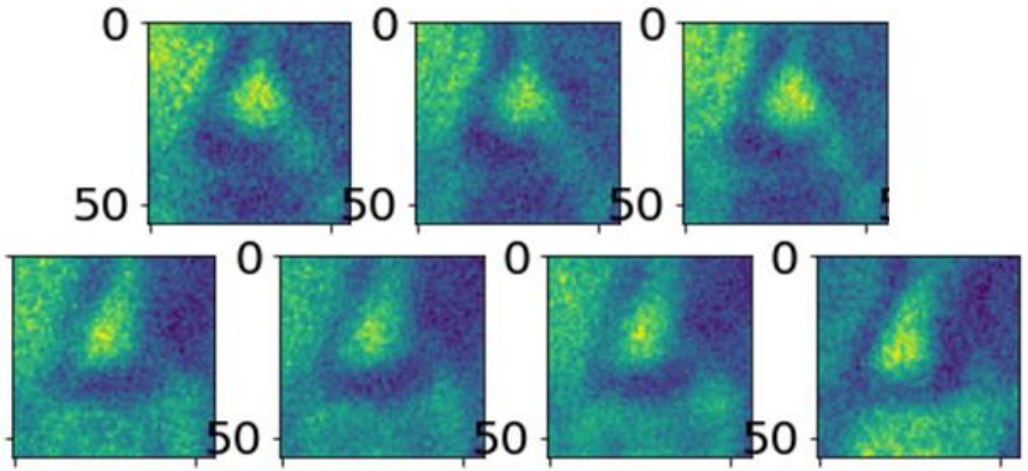
Micromotion phase



The most influential points are localized around the same micromotion phase as test point



Same shaking frequency and shaking phase different micromotion phases



N. Käming, A. Dawid et al., *MLST* **2** 035037 (2021)

Three messages

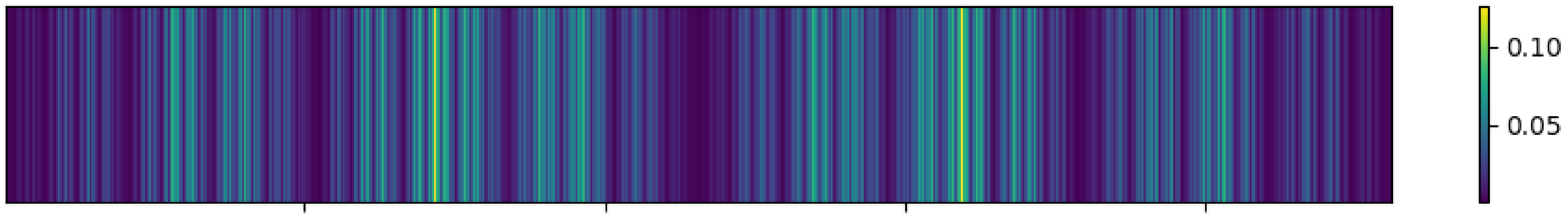
Detection additional phases

Detecting influential
data features

Anomaly detection
with influence functions

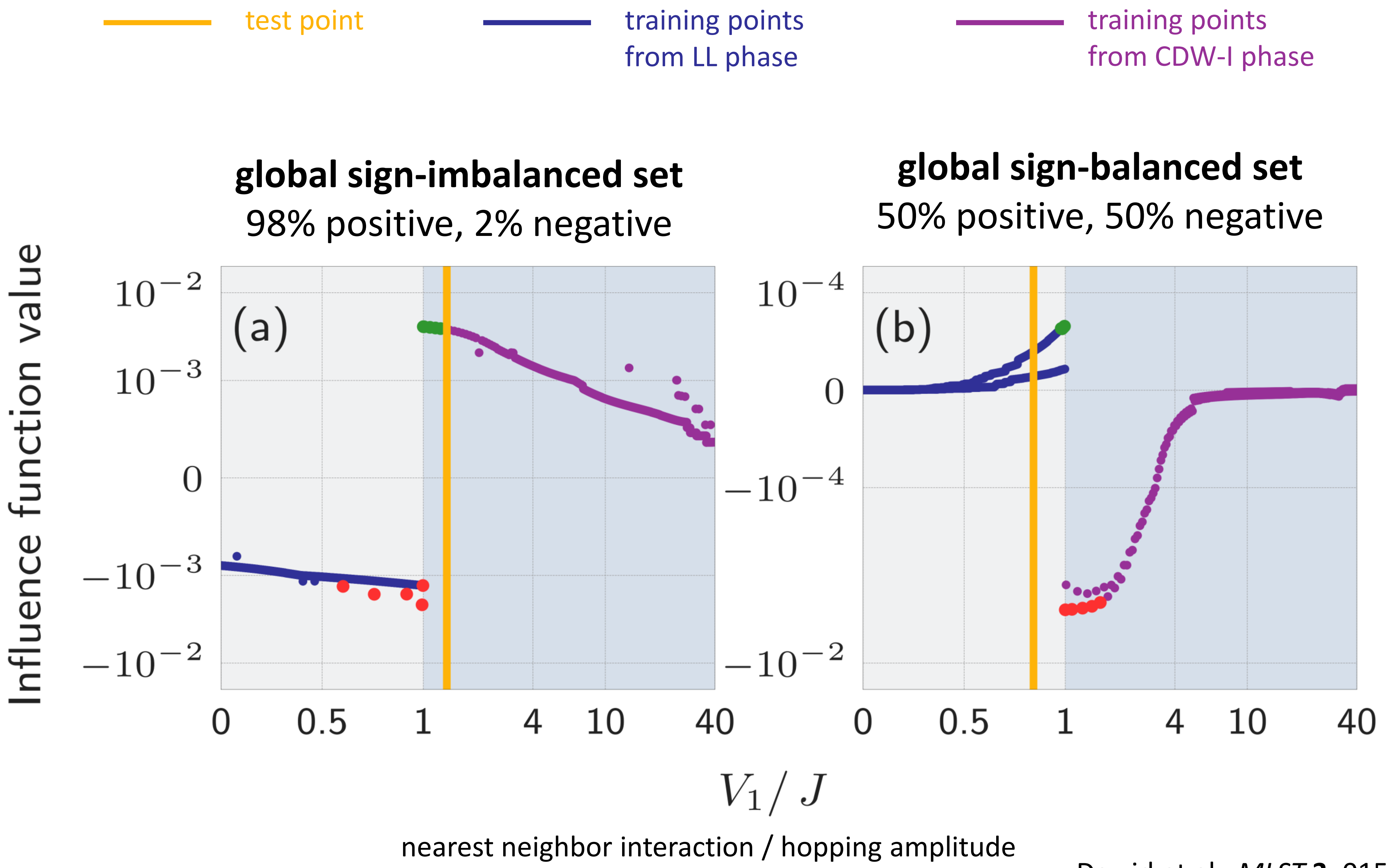
Global sign

What machine gets

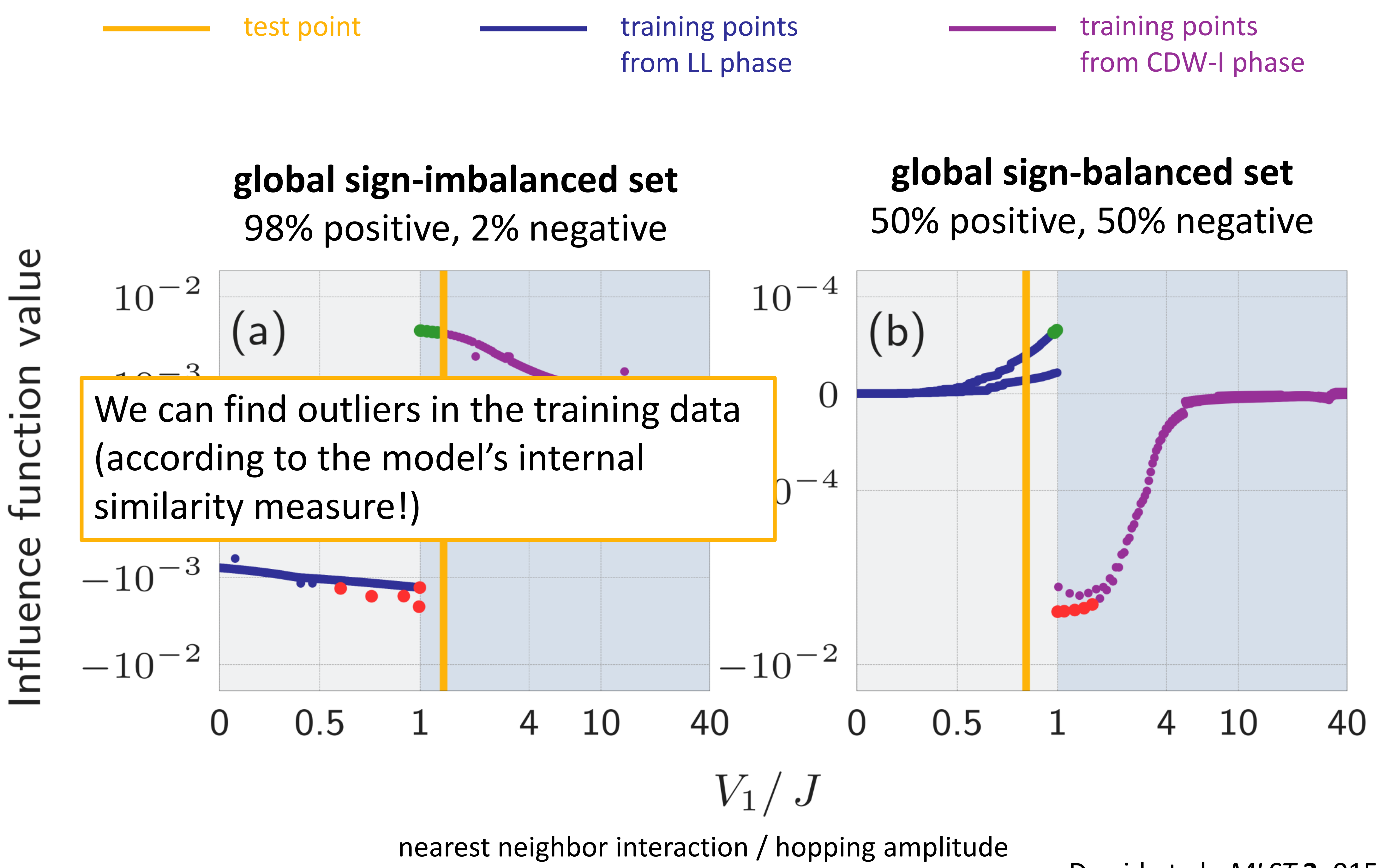


+ label: e.g., 0 for LL, 1 for CDW-I

We usually fix the global sign to +
Choice of global sign changes nothing in physics

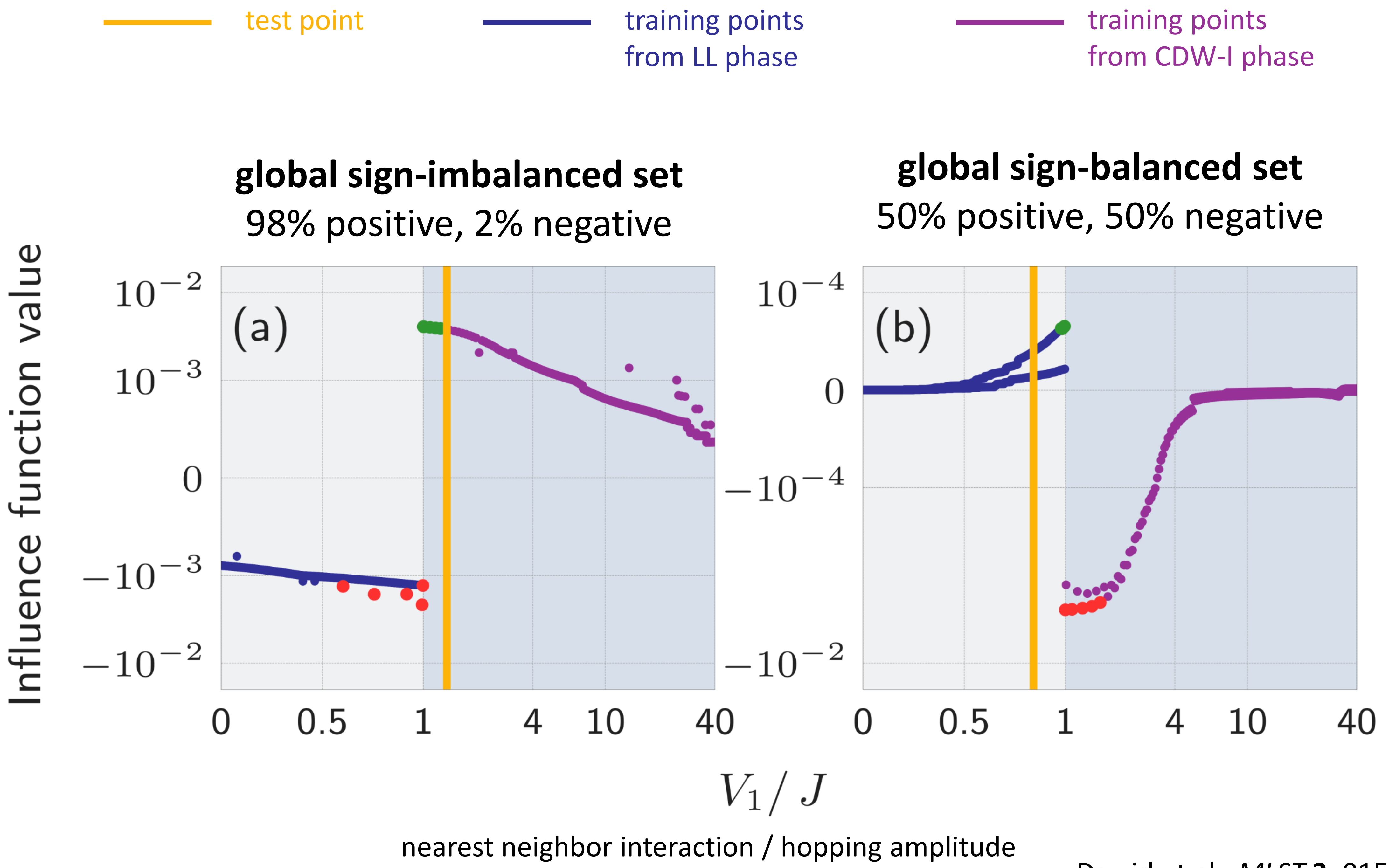


Dawid et al., *MLST* **3**, 015002 (2022)

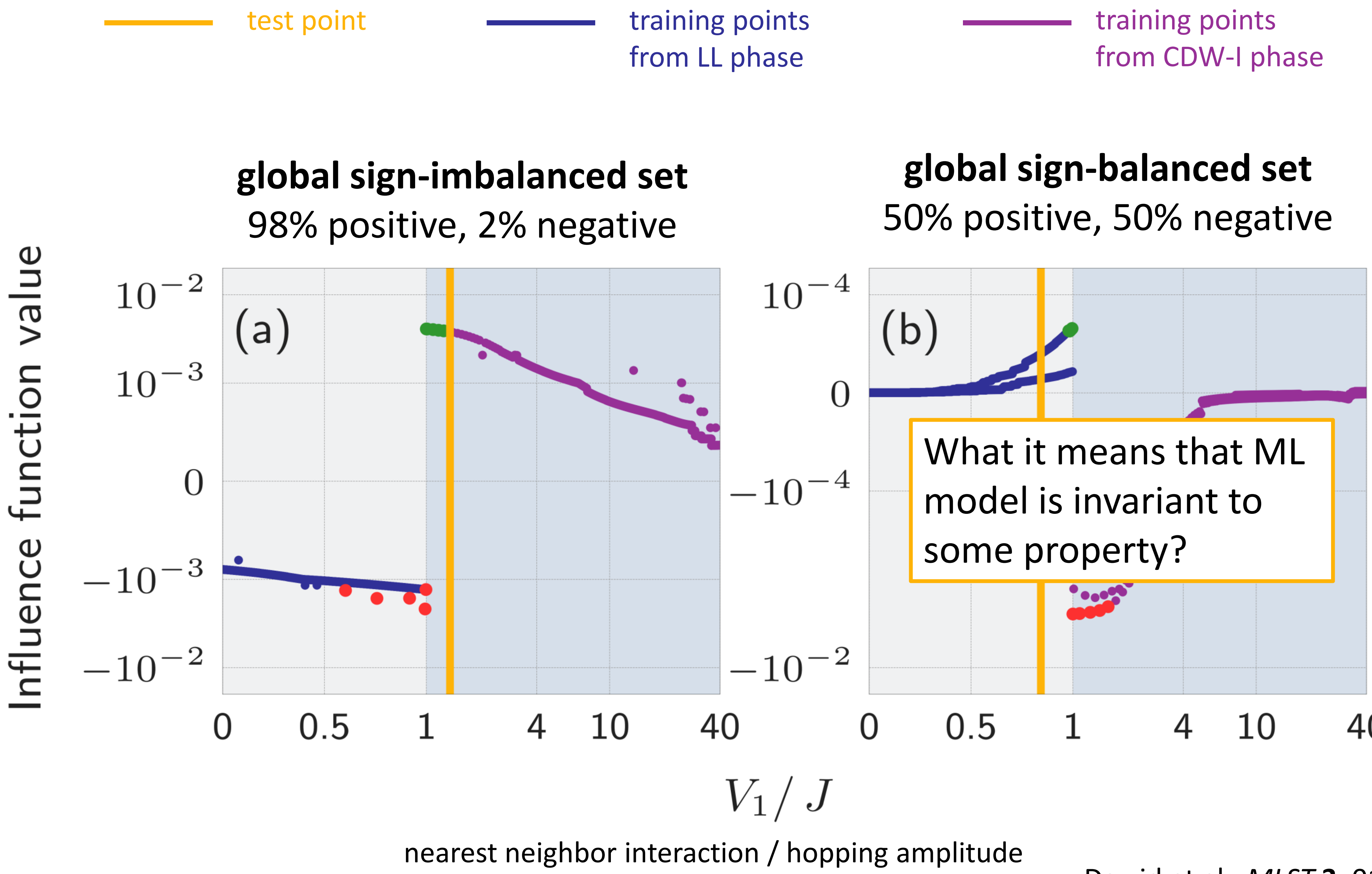


We can find outliers in the training data (according to the model's internal similarity measure!)

Dawid et al., *MLST* **3**, 015002 (2022)



Dawid et al., *MLST* **3**, 015002 (2022)

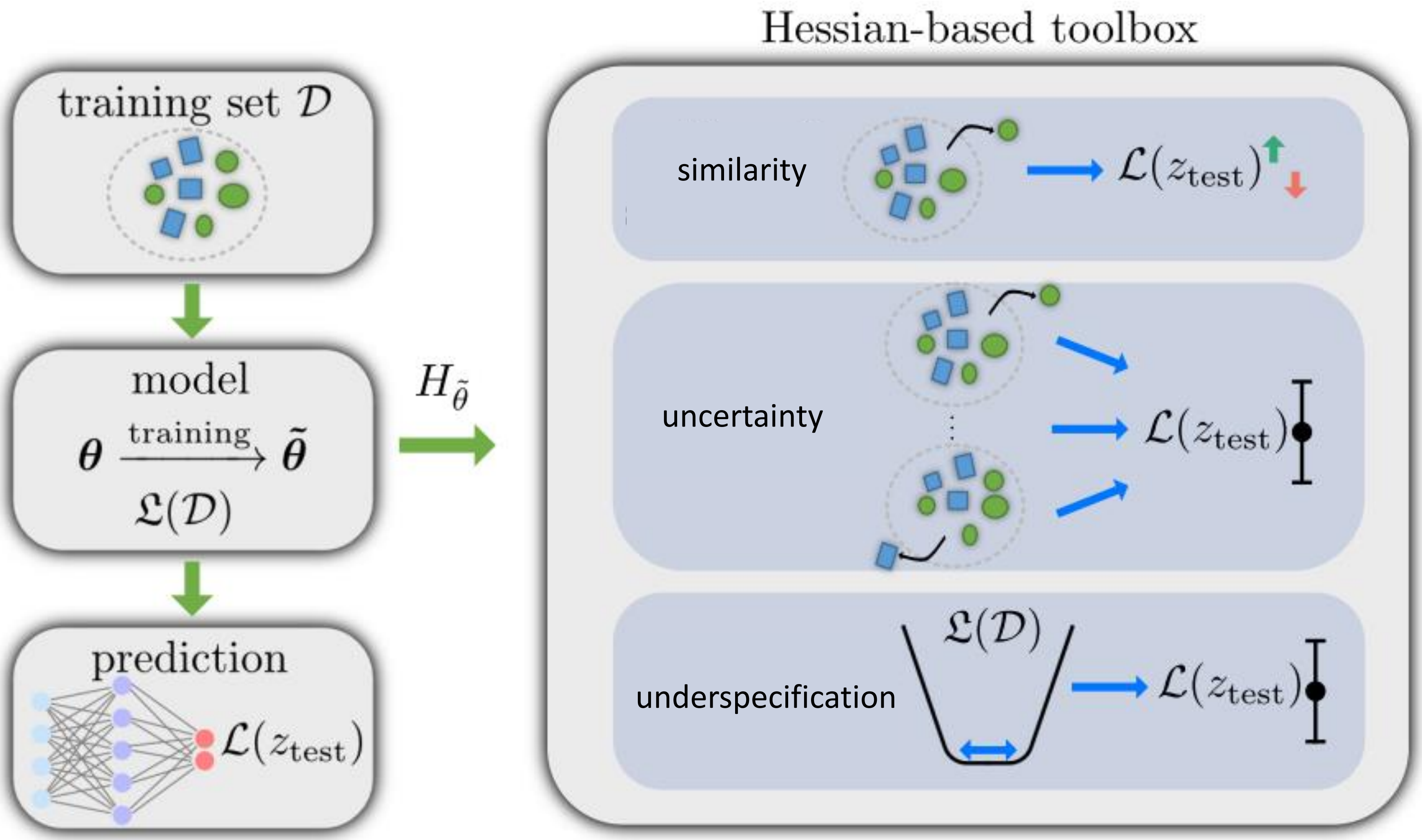


Dawid et al., *MLST* **3**, 015002 (2022)

Outline

1. Interpreting an ML model
2. Reliability methods

Hessian-based toolbox



- Influence functions
  Koh & Liang
  arXiv:1703.04730
- Resampling Uncertainty Estimation (RUE)
  Schulam & Saria
  arXiv:1901.00403
- Local Ensembles (LEs)
  Madras, Atwood, D'Amour
  arXiv:1910.09573

Resampling Uncertainty Estimation (RUE)

studied ML model

$$D[1,1,1,1,\ldots,1,1]$$

training dataset, every
data point is taken once

b bootstrap samples

$$D_1[0,1,3,\ldots,1,2]$$

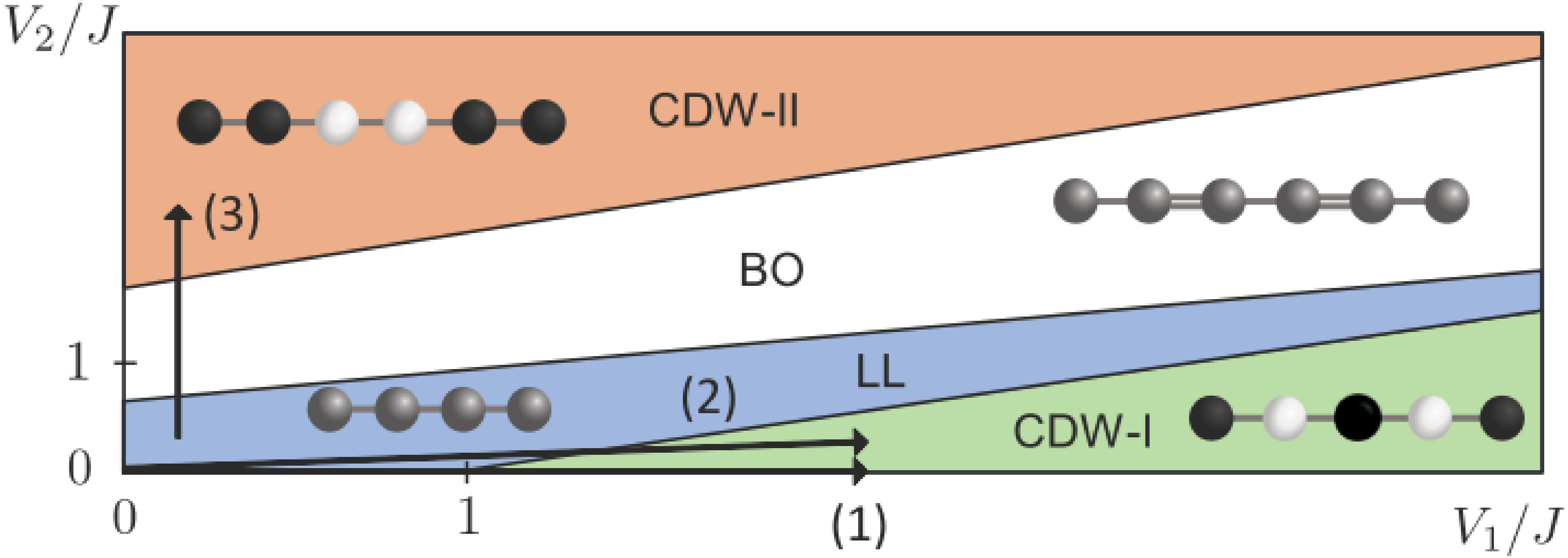
$$D_2[1,4,0,\ldots,0,1]$$

$$D_b[2,0,3,\ldots,1,0]$$

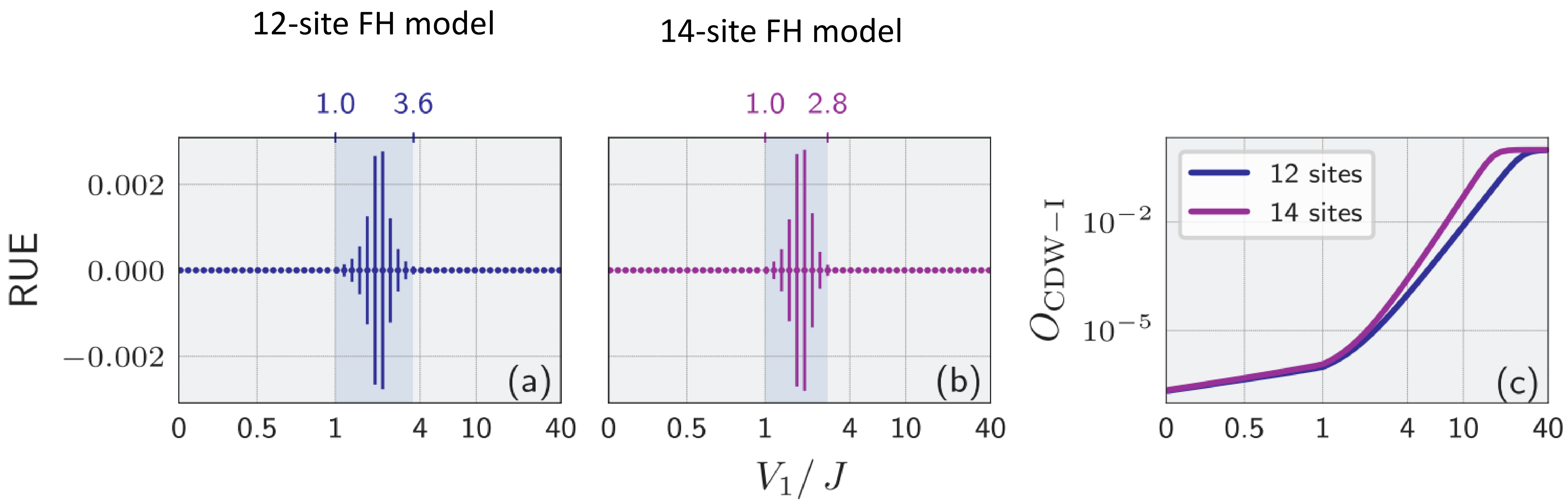
b predictions

variance

P. Schulam and S. Saria, *Can you trust this prediction? Auditing pointwise reliability after learning*, AISTATS 2019 - 22nd Int. Conf. Artif. Intell. Stat. 89 (2020), arXiv:1901.00403v2.



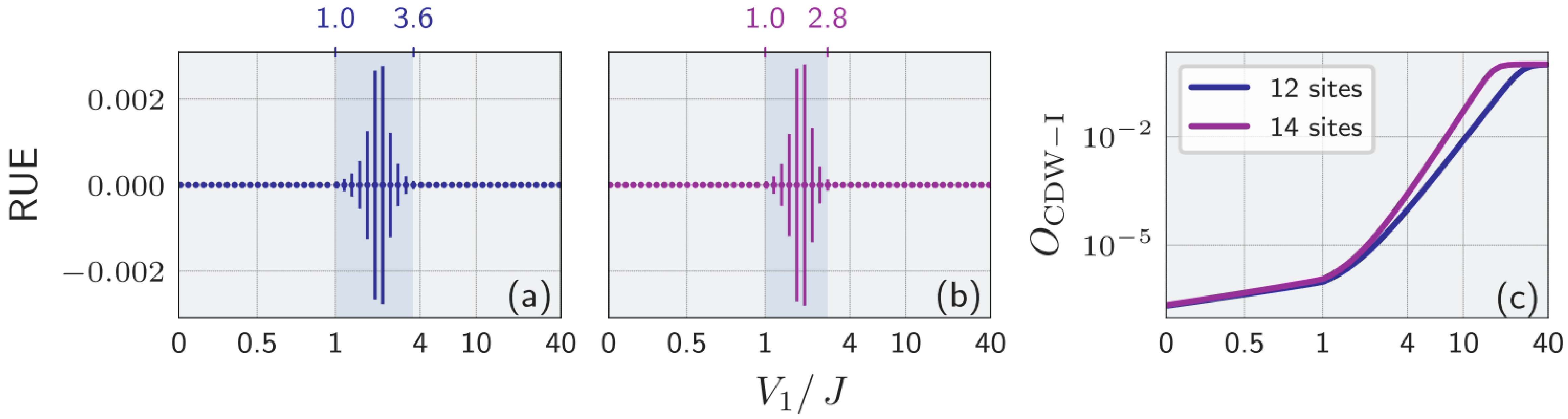
V_2/J
CDW-II
(3)
BO
1
LL
(2)
CDW-I
0
0
1
(1)
V_1/J



Dawid et al., *MLST* **3**, 015002 (2022)

12-site FH model

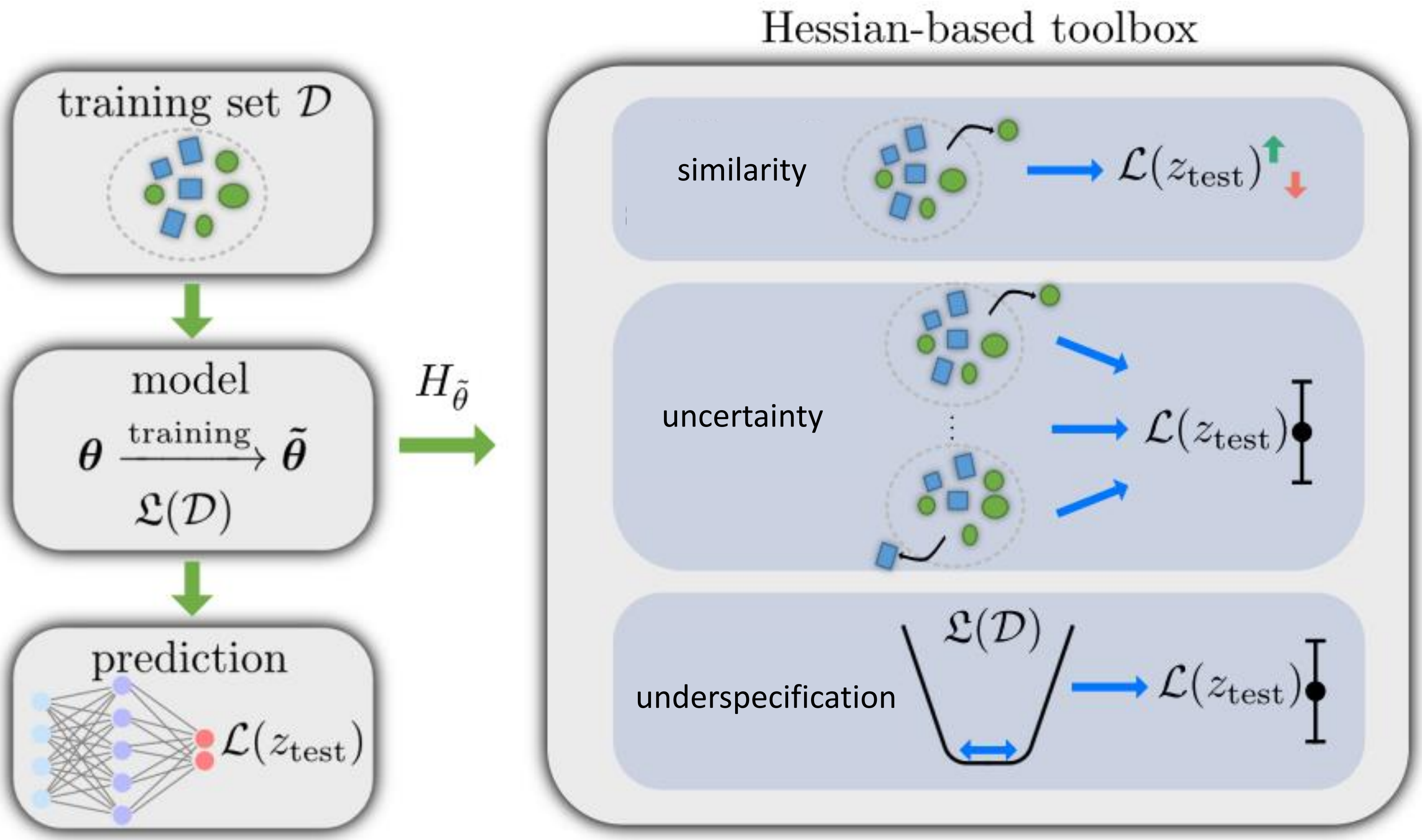
14-site FH model



RUE indicates sharpness of the transition

Dawid et al., *MLST* **3**, 015002 (2022)

Hessian-based toolbox

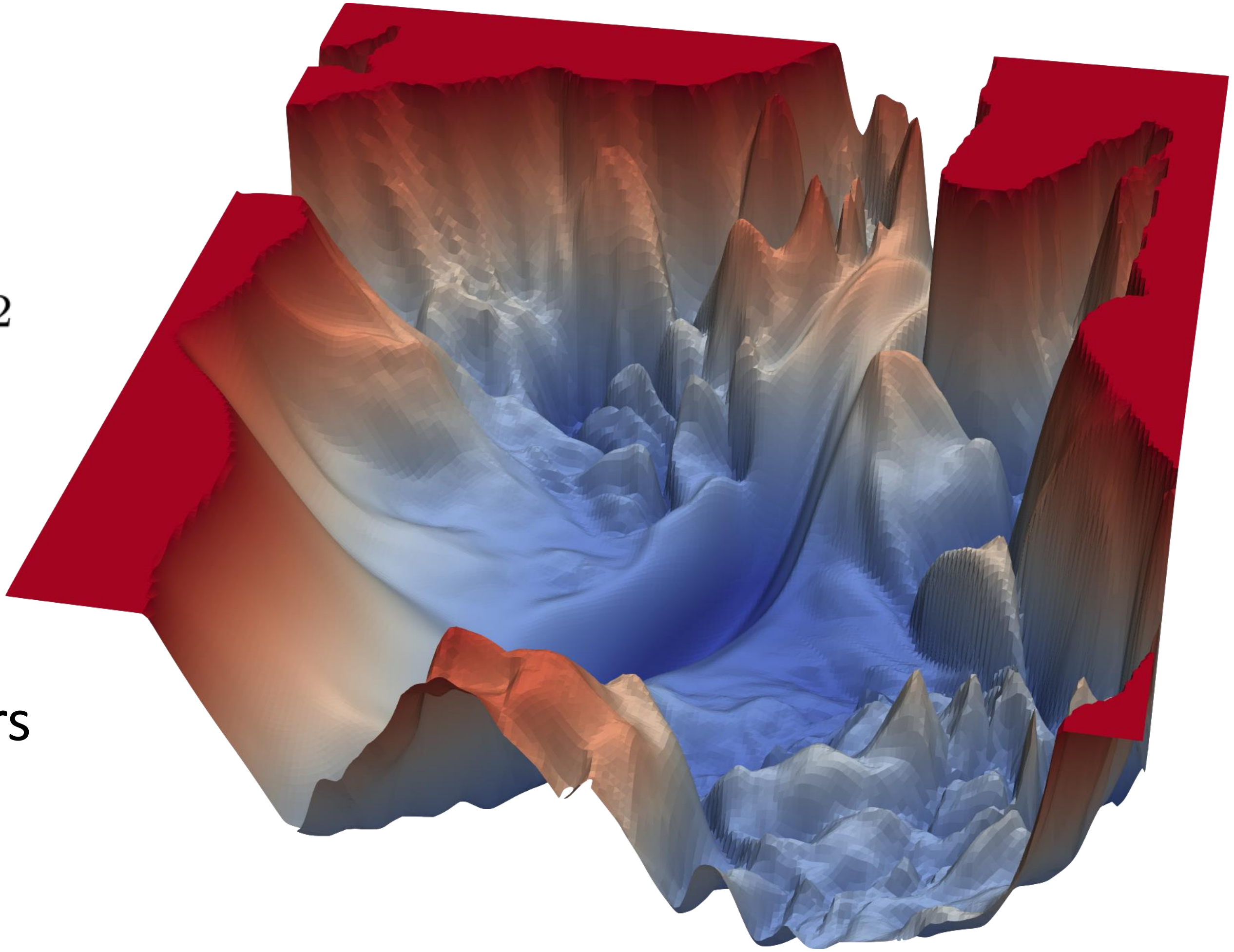


- Influence functions
  Koh & Liang
  arXiv:1703.04730
- Resampling Uncertainty Estimation (RUE)
  Schulam & Saria
  arXiv:1901.00403
- Local Ensembles (LEs)
  Madras, Atwood, D'Amour
  arXiv:1910.09573

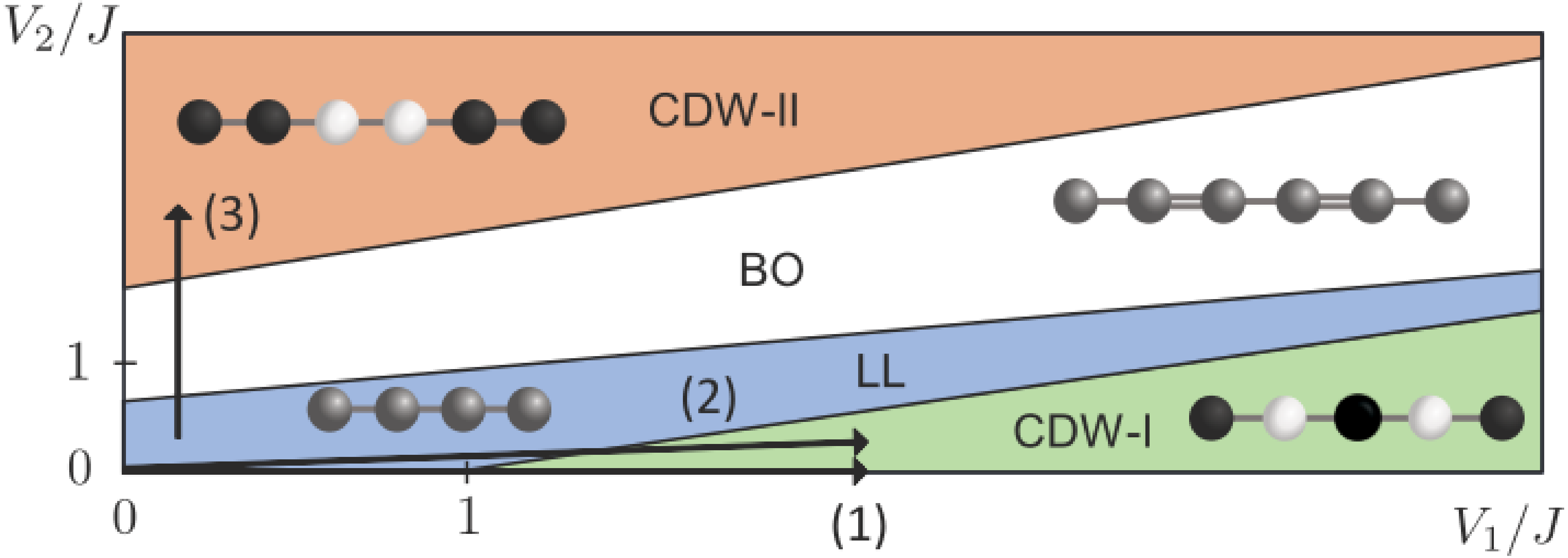
Local-Ensemble-based Extrapolation Score (LEES)

$$\mathcal{E}_m(x_{\text{test}}) = ||U_m^\top \nabla_\theta \mathcal{L}(z_{\text{test}}, \tilde{\theta})||_2$$

matrix of (M − m) Hessian eigenvectors spanning a subspace of low curvature

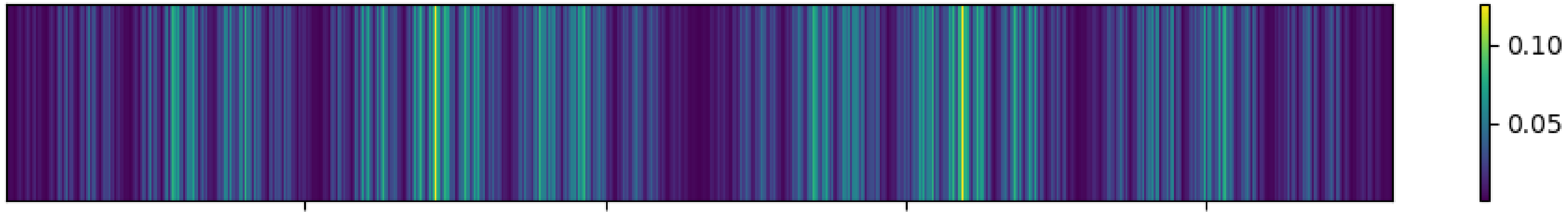


D. Madras, J. Atwood, and A. D'Amour, *Detecting Extrapolation with Local Ensembles*. (2019) arXiv:1910.09573.



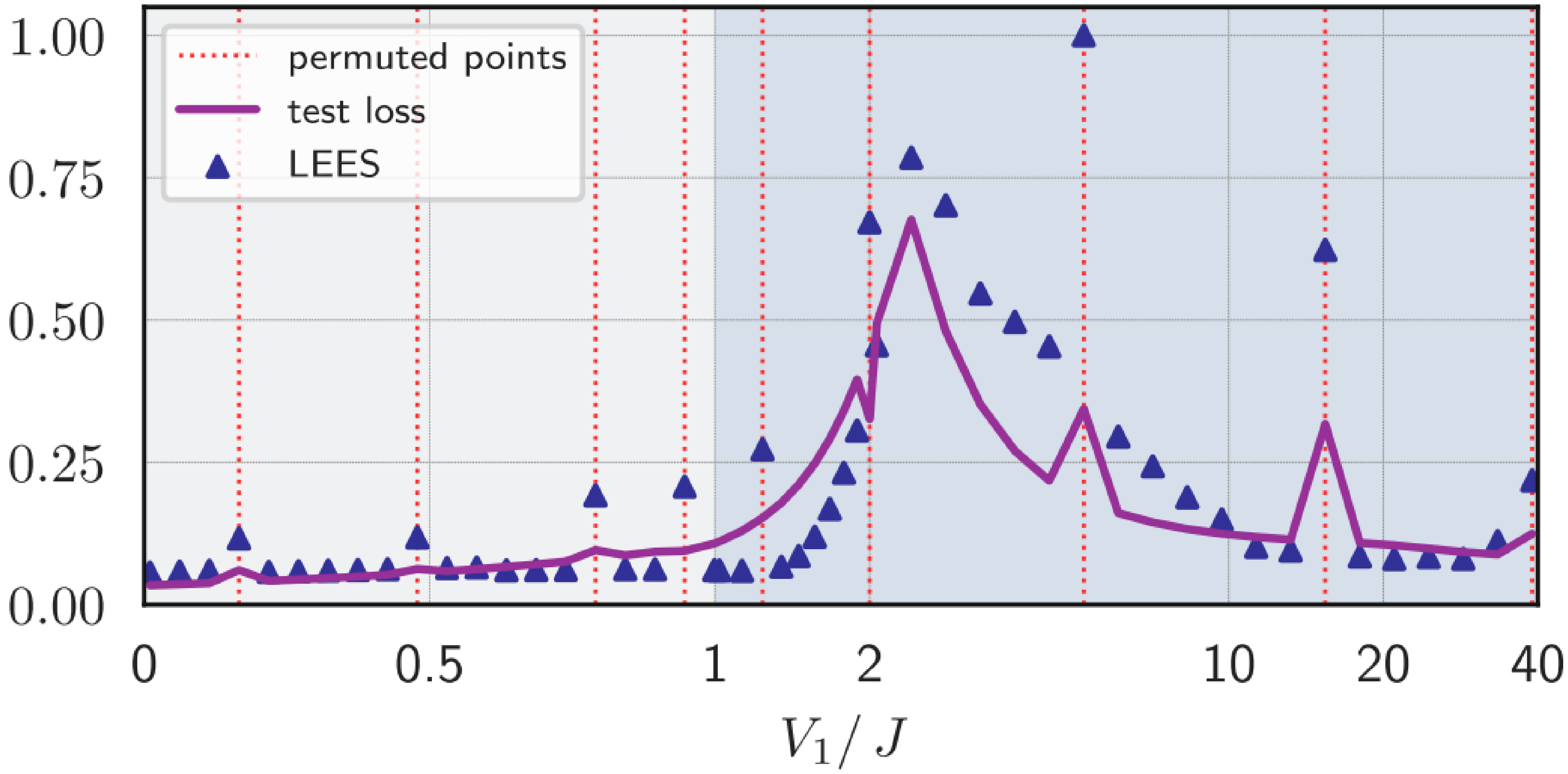
V_2/J
CDW-II
(3)
BO
1
LL
(2)
CDW-I
0
0
1
(1)
V_1/J

Out-Of-Distribution (OOD) test points



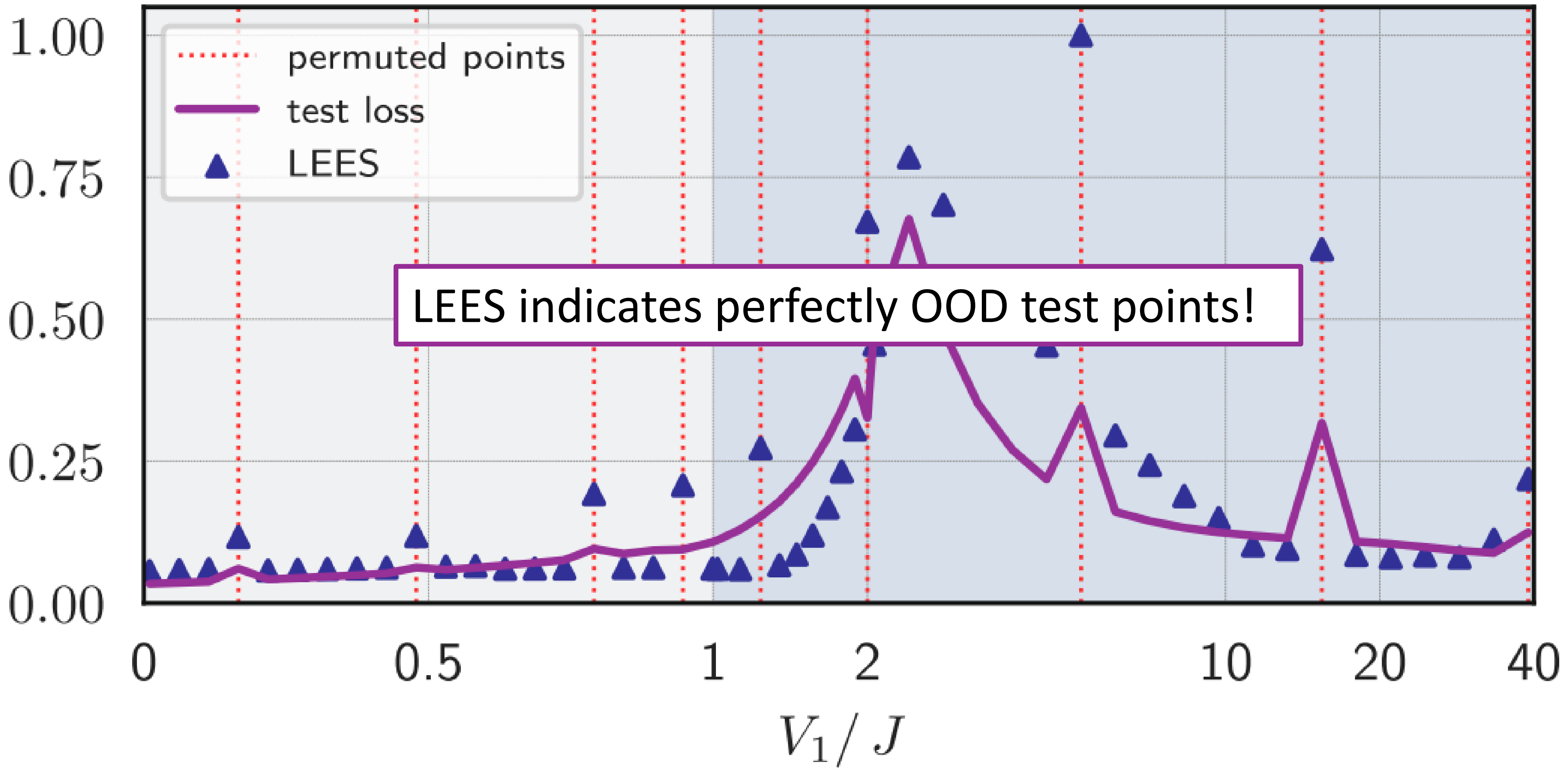
random permutation of eigenvector elements

Local Ensemble-based Extrapolation Score



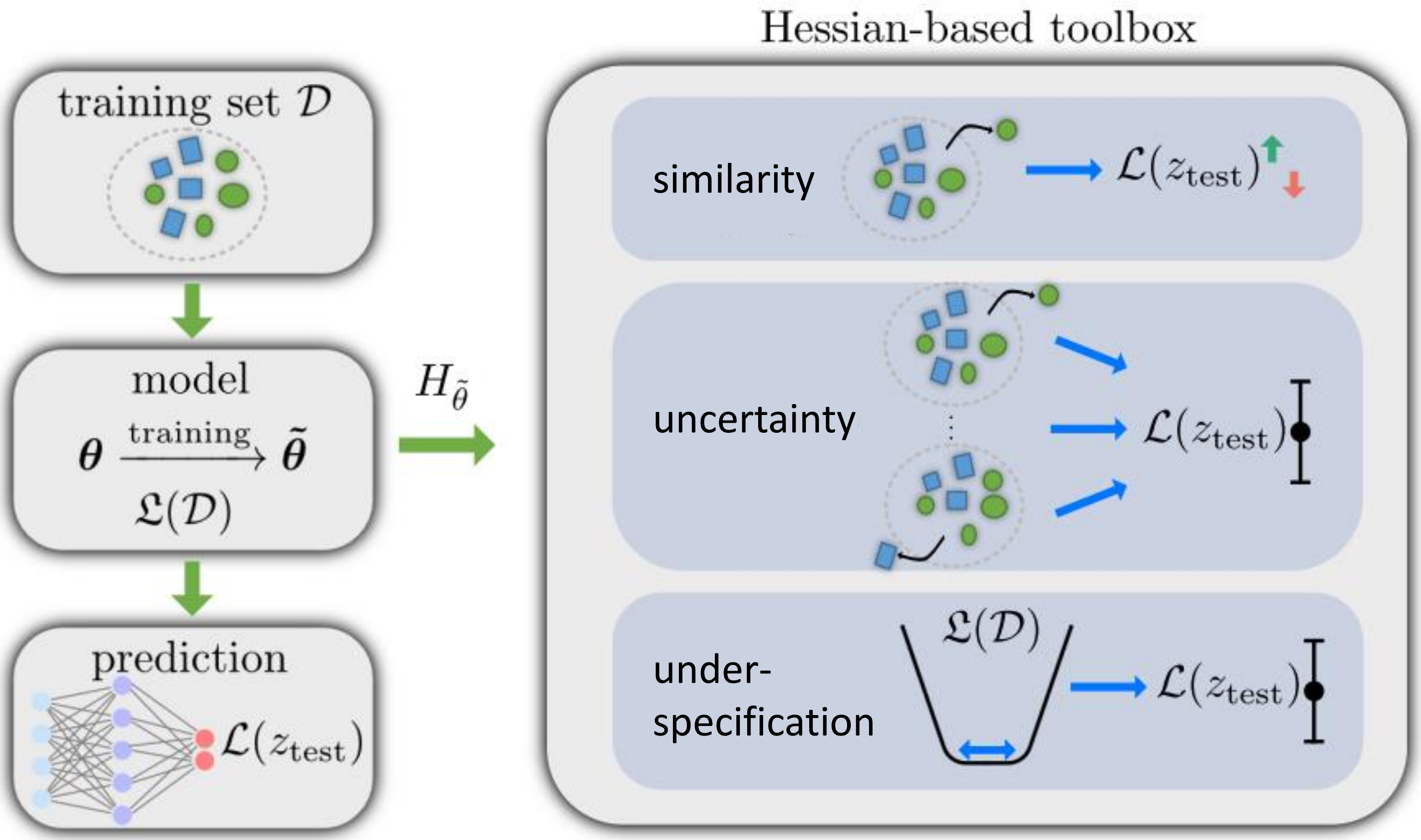
Dawid et al., *MLST* **3**, 015002 (2022)

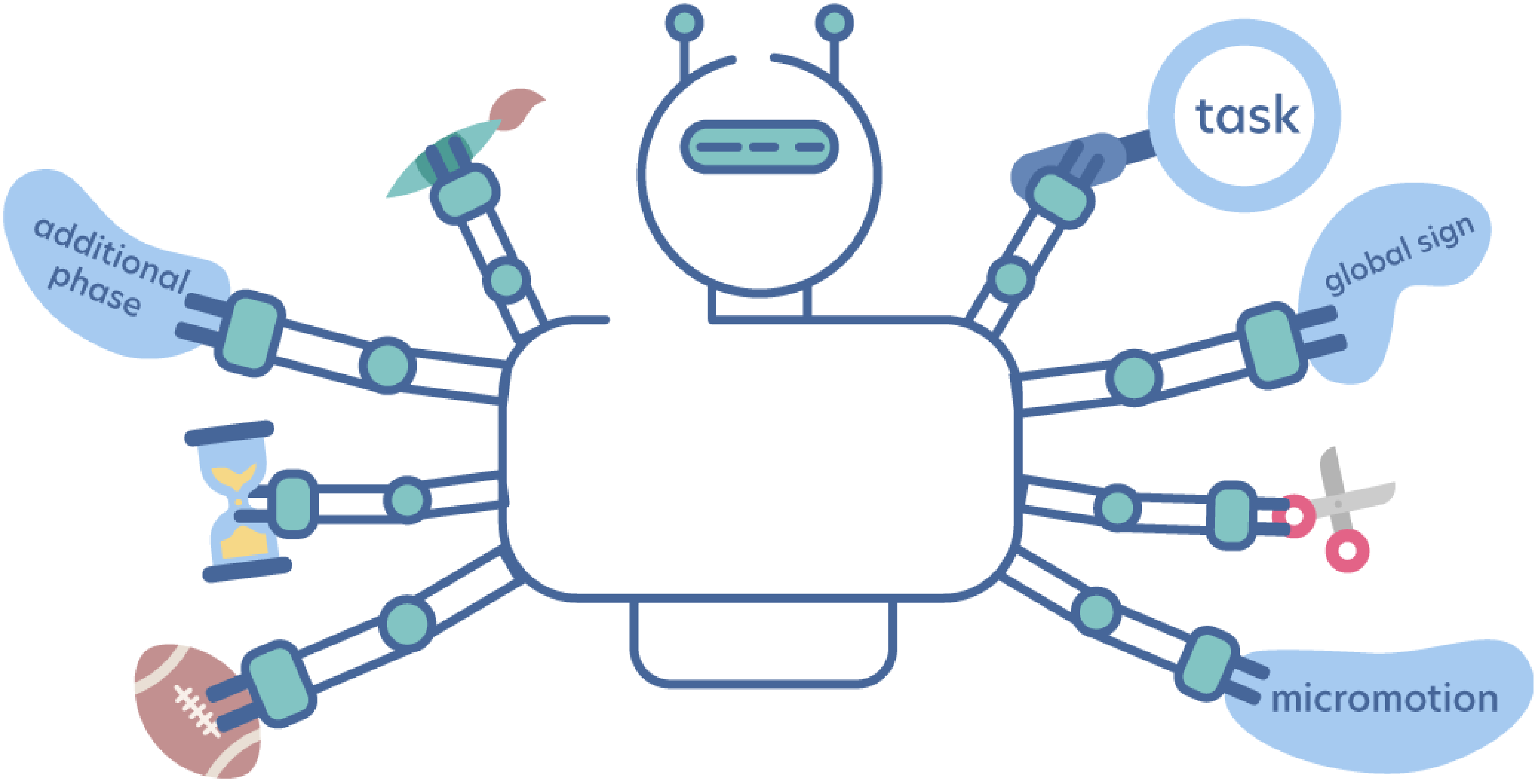
Local Ensemble-based Extrapolation Score



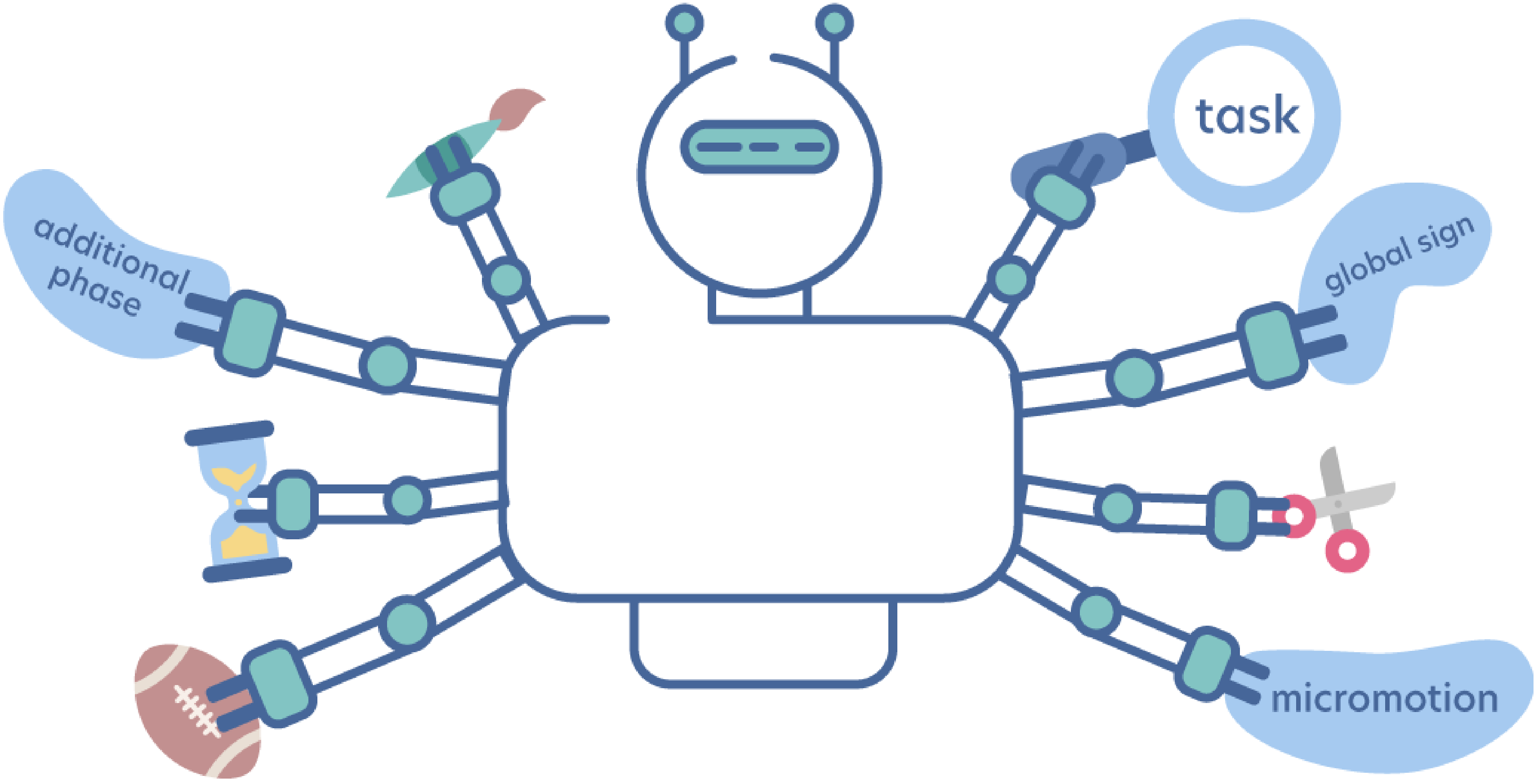
Dawid et al., *MLST* **3**, 015002 (2022)

Conclusions

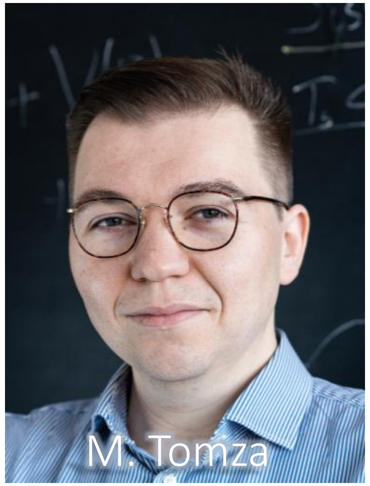



additional phase
task
global sign
micromotion

More likely to learn physical features than spurious correlations?



UNIVERSITY OF WARSAW



ICFO
The Institute of Photonic Sciences

Universität Hamburg

If you want to know more...

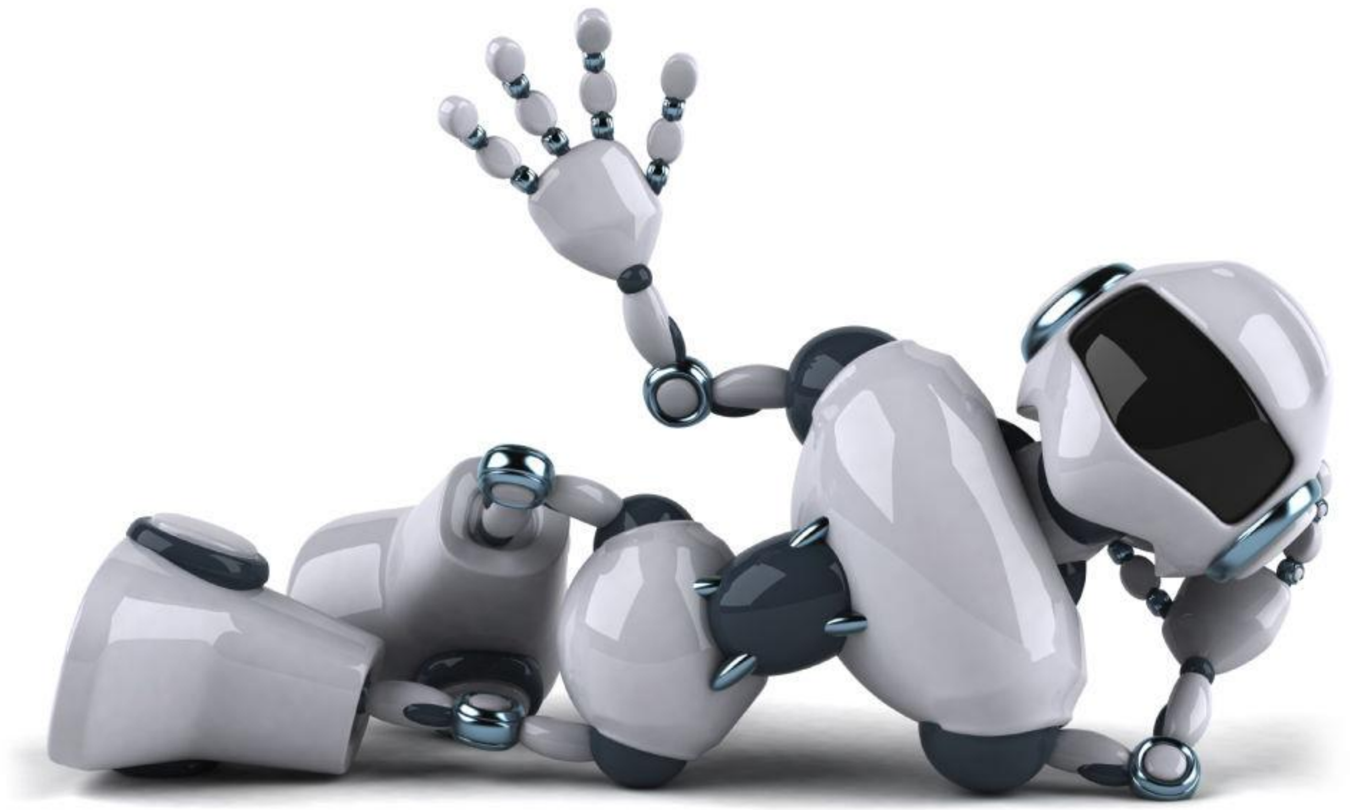
/Shmoo137

A. Dawid et al (2022)
Mach. Learn.: Sci. Technol. **3** 015002 (2022)

N. Käming, A. Dawid et al (2021)
Mach. Learn.: Sci. Technol. **2** 035037

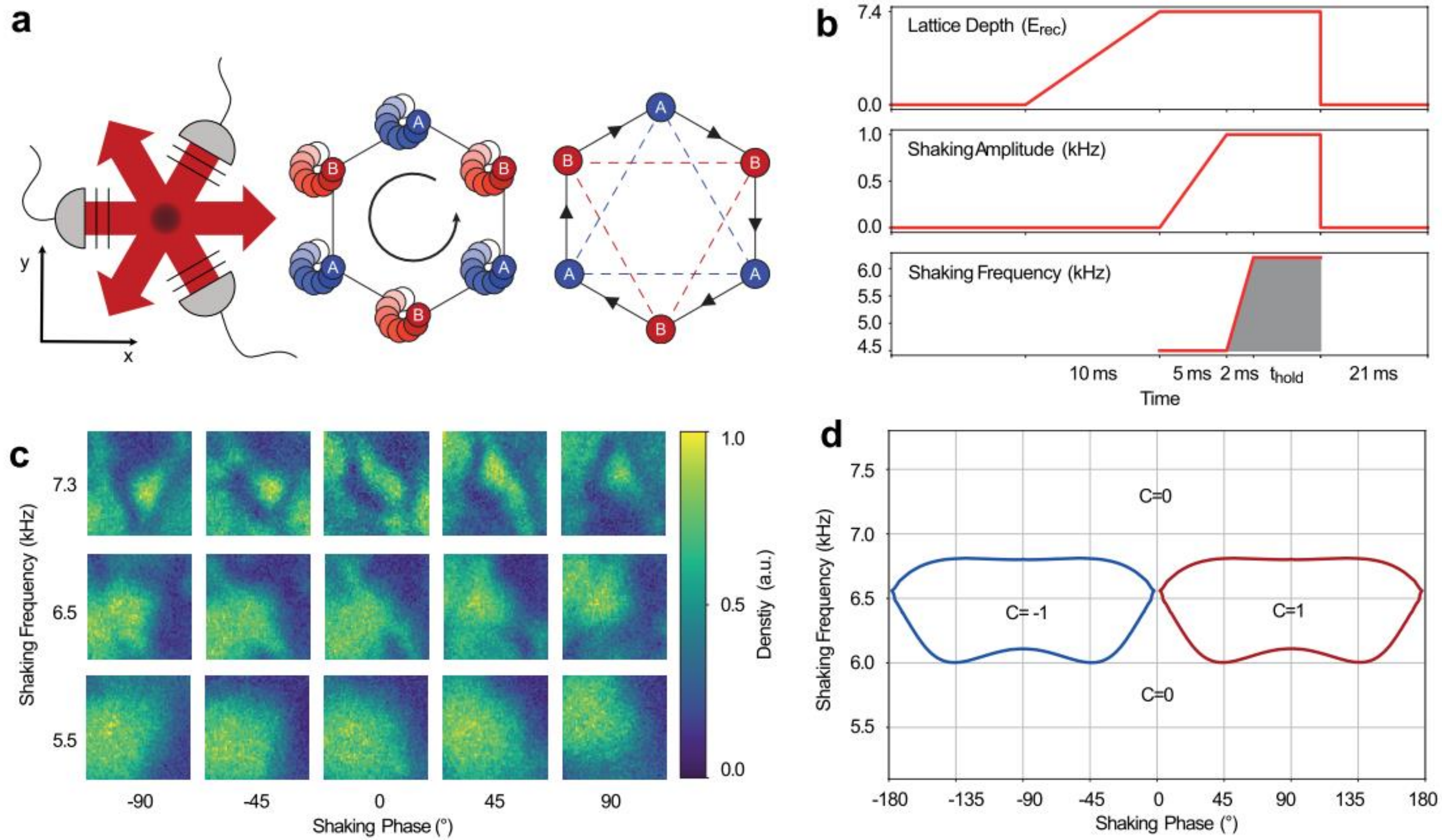
A. Dawid et al (2020)
New J. Phys. **22** 115001

Thank you for your attention!



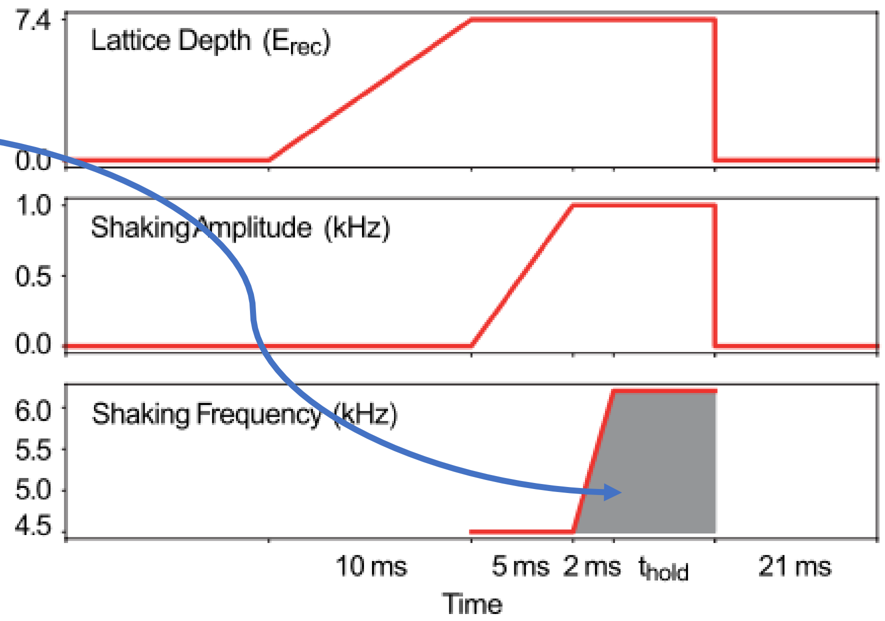
Topological Haldane model

realized via Floquet-driving of ultracold fermions (^{40}K) in a honeycomb lattice

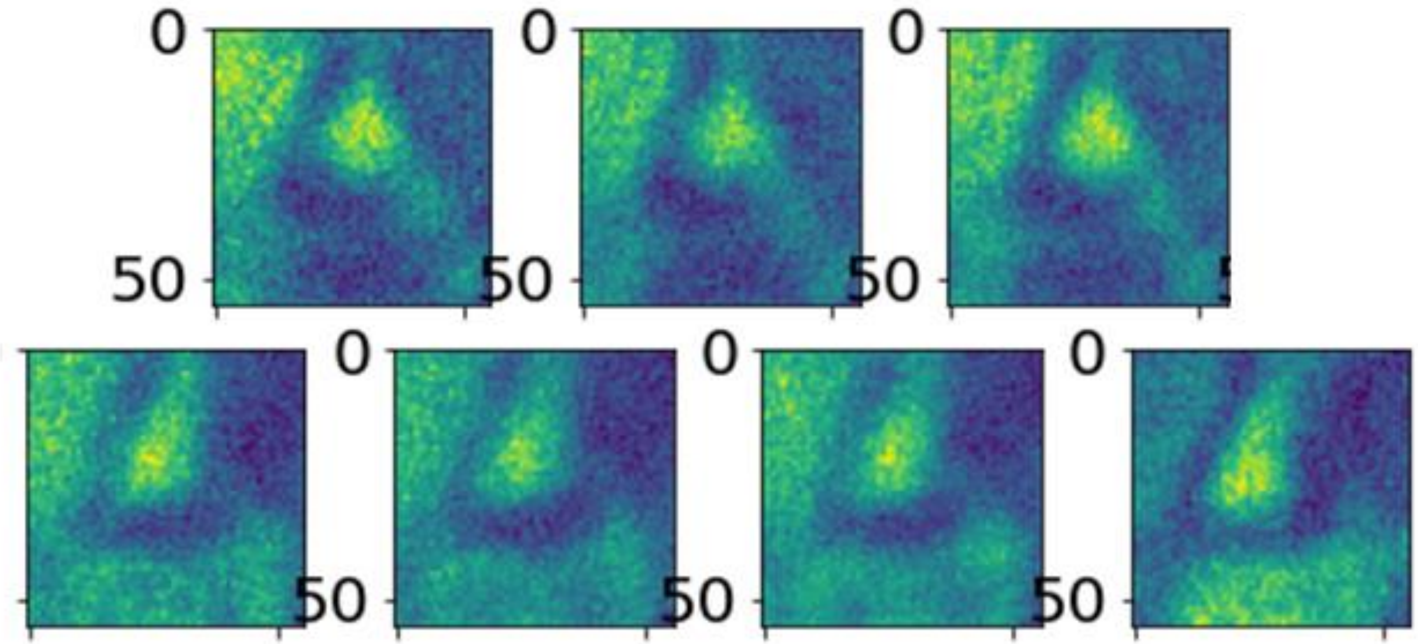


N. Käming, A. Dawid et al., *MLST* **2** 035037 (2021)

Micromotion phase

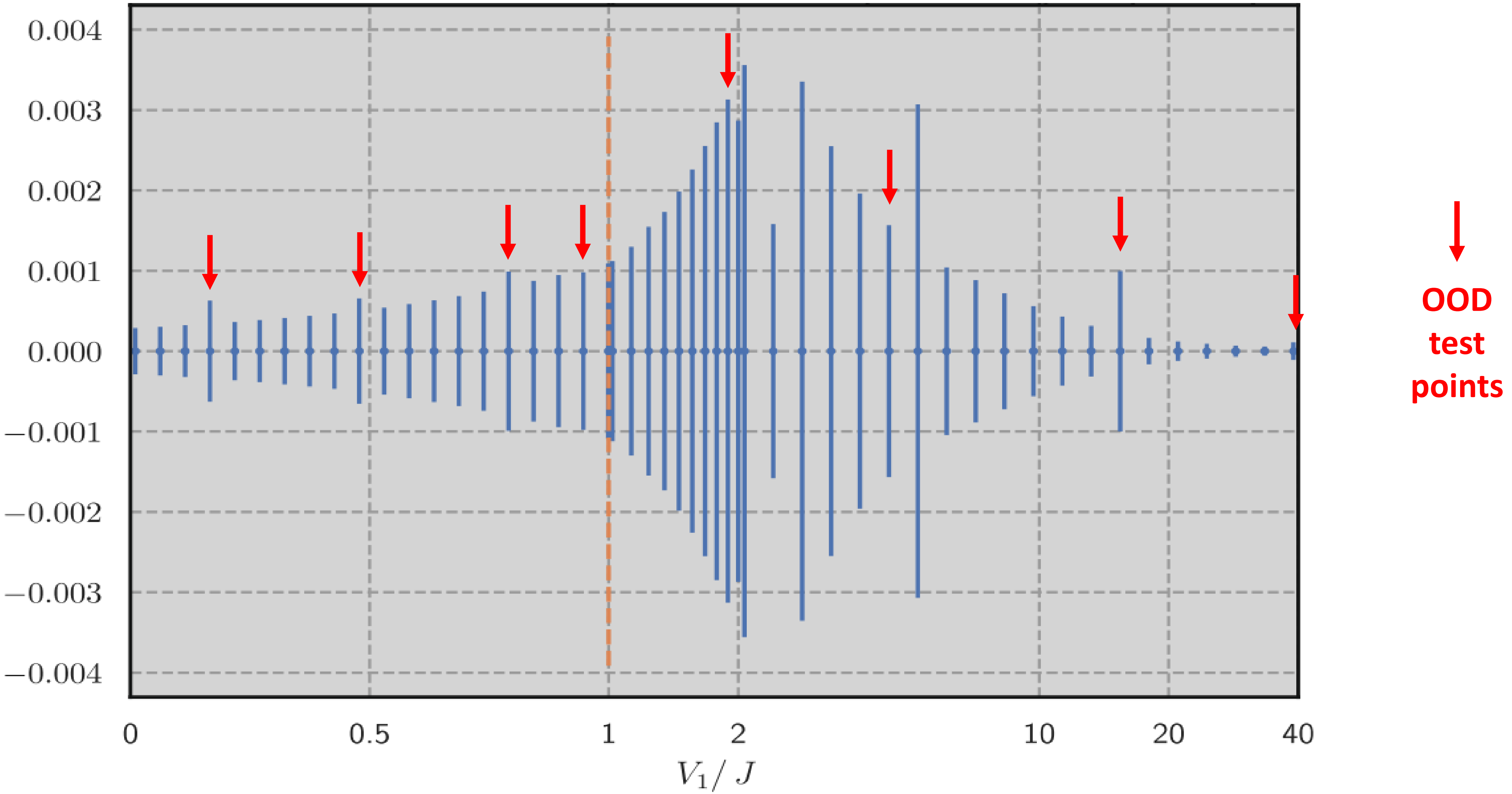


shaking frequency = 7.4 Hz
shaking phase = 90˚
different micromotion phases



N. Käming, A. Dawid et al., *MLST* **2** 035037 (2021)

RUE vs LEES



0.004
0.003
0.002
0.001
0.000
−0.001
−0.002
−0.003
−0.004
0
0.5
1
2
10
20
40
V_1/J
OOD test points