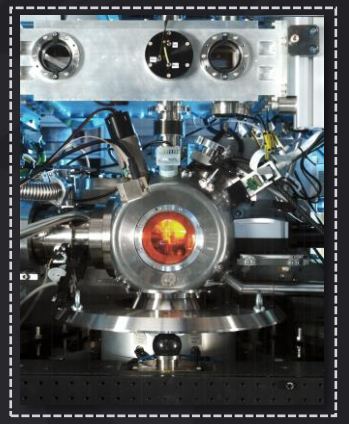
Teaching Core-Hole Spectroscopy to a Deep Neural Network

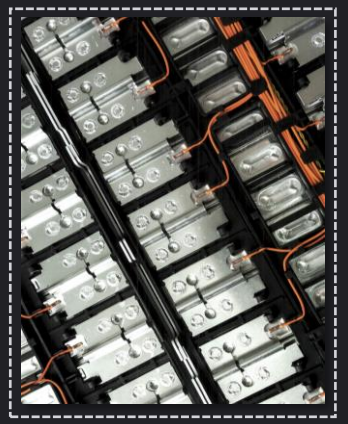
Dr. Conor D. Rankine

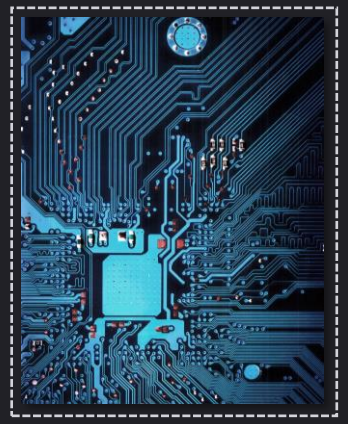
Lecturer in Machine Learning & Computational Chemistry

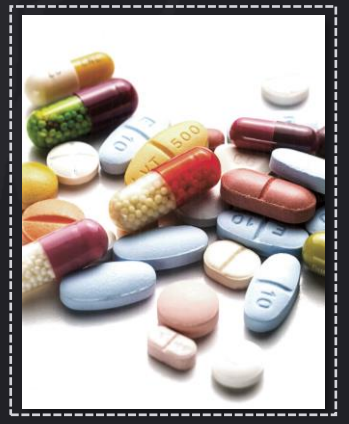
X-RAY SPECTROSCOPY


crystallography

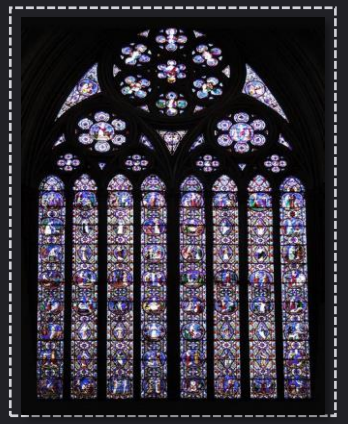

cryo-imaging

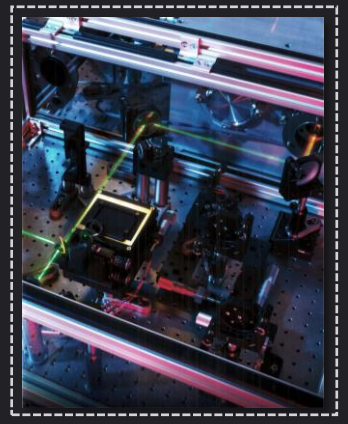

surface science


magnetic materials


drug design

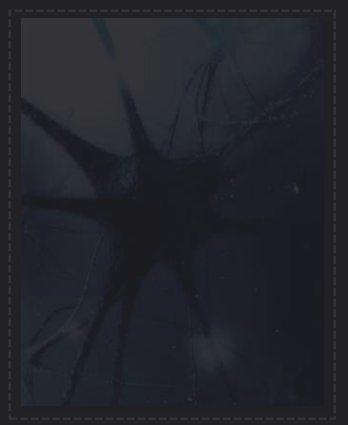

agriculture


conservation


dynamics

Teaching Core-Hole Spectroscopy to a Deep Neural Network

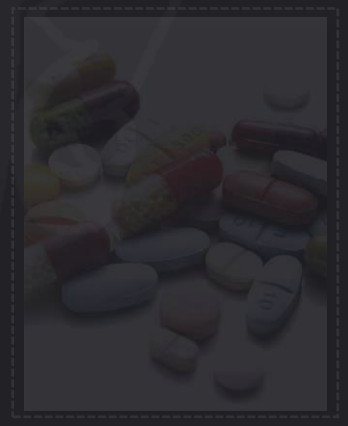
crystallography



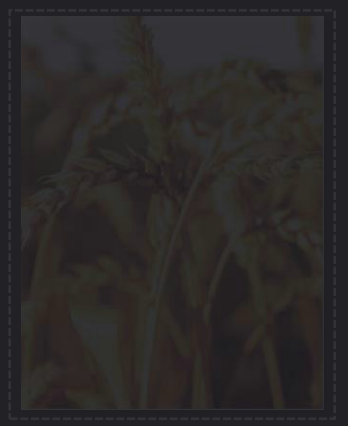
cryo-imaging

surface science

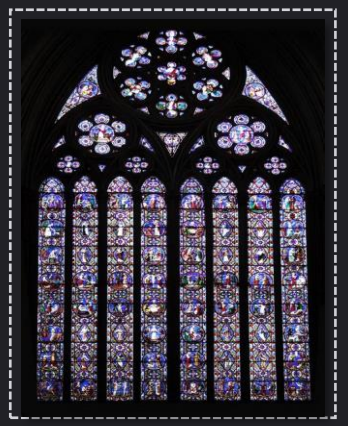
magnetic materials



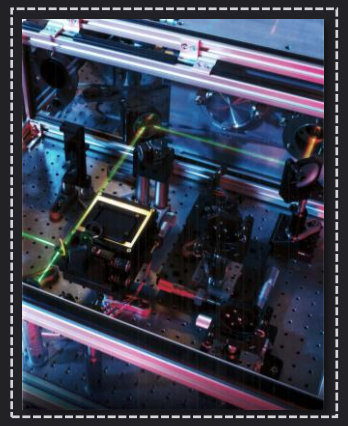
drug design



agriculture



conservation



dynamics

simulating core-hole spectroscopy with theory is the best way to link the experimental observables to the chemistry and physics…

…but **theory is hard!** – all too often, simulations are:

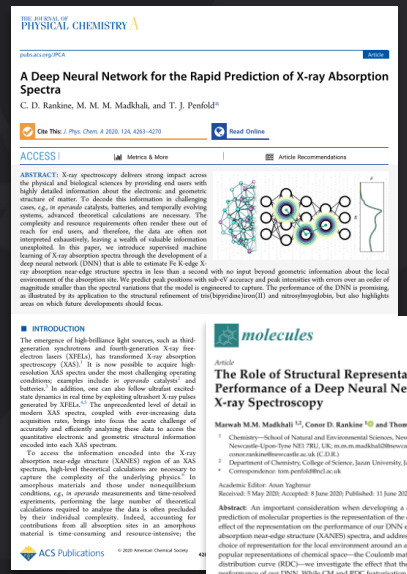
- inefficient; time- and resource intensive
- reliant on legacy/unmaintained codes
- the preserve of expert theoreticians

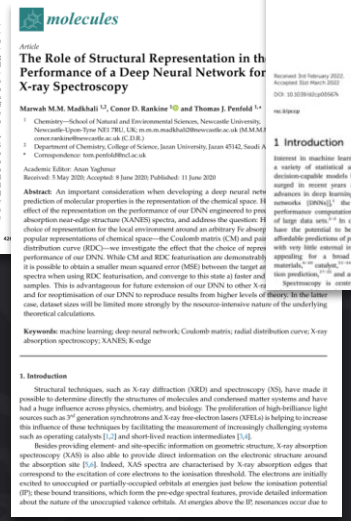
…and **it's easier to forget about it!**

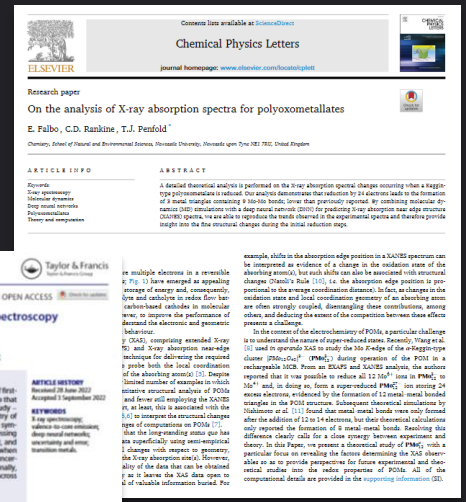
Teaching Core-Hole Spectroscopy to a Deep Neural Network

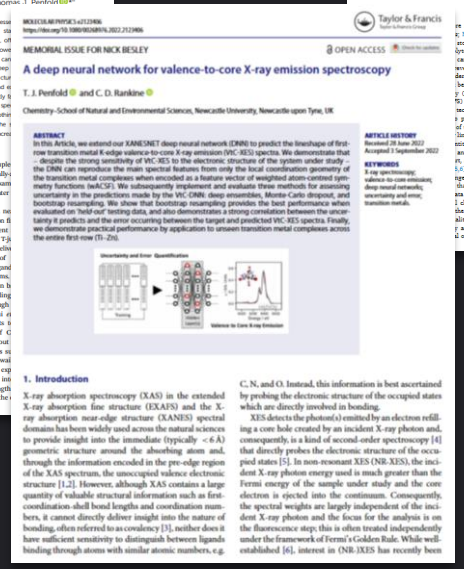
X-ray Spectroscopy | Results | Dataset | Model | Lifting The Lid | Outlook

deep neural nets can make simulating and interpreting core-hole spectra **accurate**, **affordable**, **fast**, and **easy**!



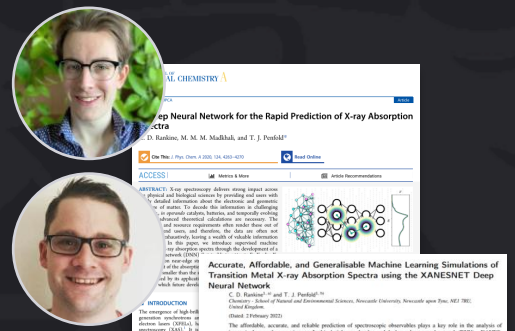






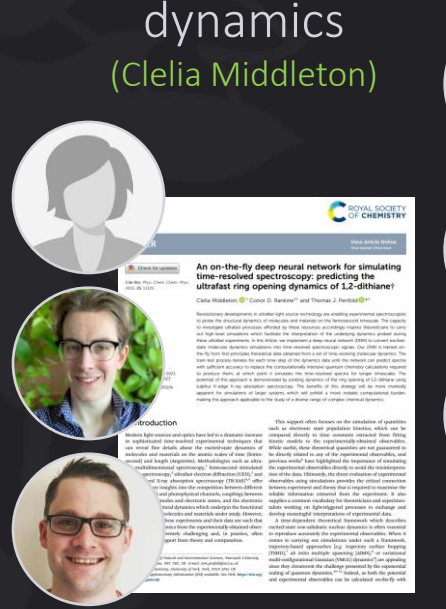
Teaching Core-Hole Spectroscopy to a Deep Neural Network

X-ray Spectroscopy | Results | Dataset | Model | Lifting The Lid | Outlook



neural net design, development, and optimisation
(Dr. Conor Rankine)

simulating chemical dynamics
(Clelia Middleton)

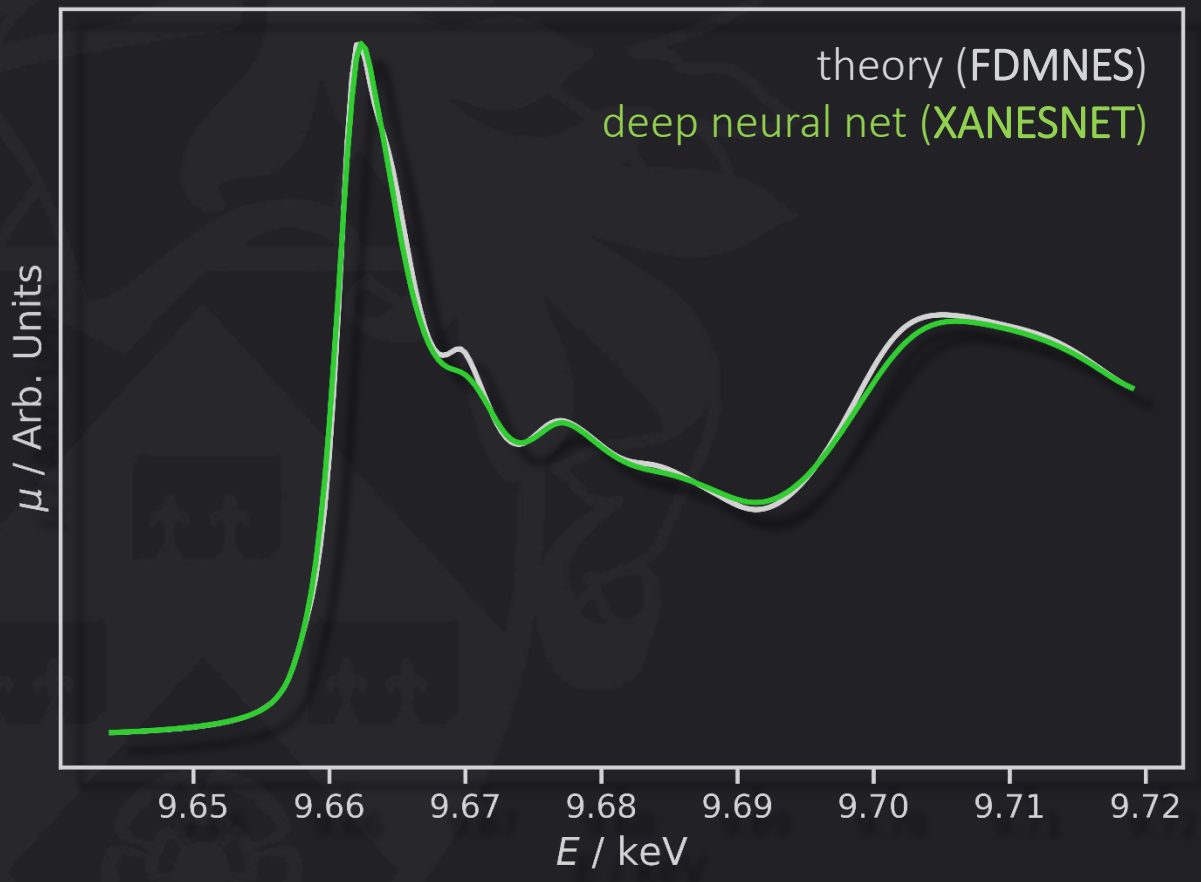


simulating *T*-jump spectroscopy
(Dr. Marwah Madkhali)

new X-ray spectroscopies
(Luke Watson)

modelling polyoxometalates
(Dr. Emanuele Falbo)

Teaching Core-Hole Spectroscopy to a Deep Neural Network



theoretical K-edge X-ray spectrum
popular codes | hours | many CPUs

deep neural net prediction
XANESNET | a second | one CPU

metrics‡

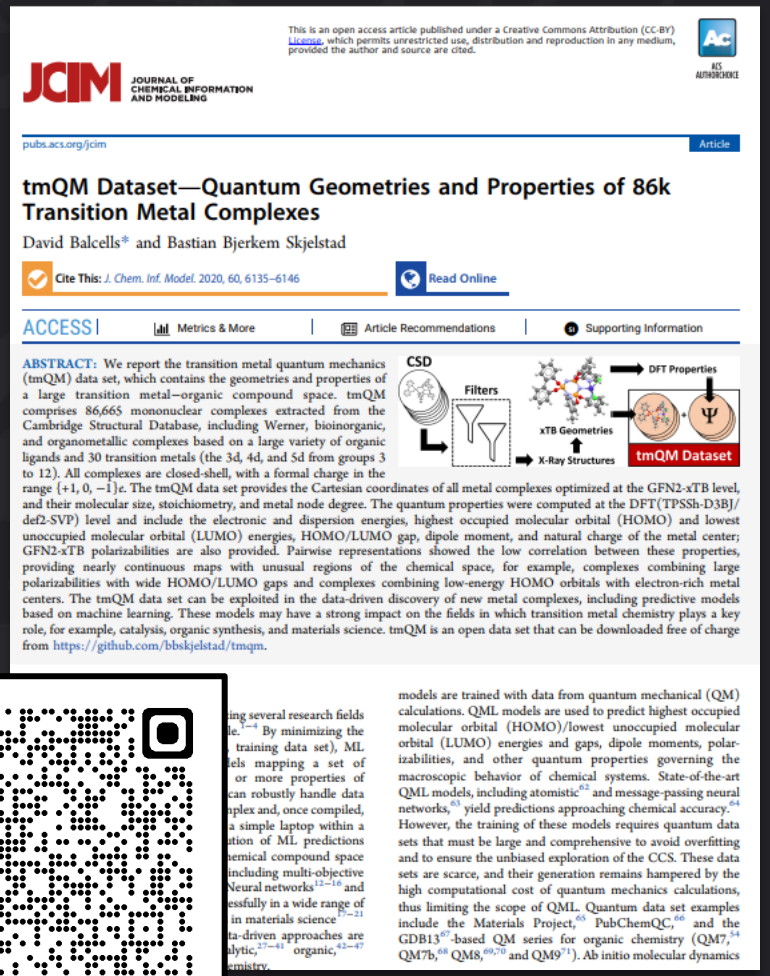
average error on peak intensities (%)	2.9
average accuracy on peak positions (eV)	0.8

‡ evaluated on 2,500 first-row transition metal (Ti–Zn) complexes

Teaching Core-Hole Spectroscopy to a Deep Neural Network

Teaching Core-Hole Spectroscopy to a Deep Neural Network

X-ray Spectroscopy | Results | Dataset | Model | Lifting The Lid | Outlook



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

tmQM Dataset—Quantum Geometries and Properties of 86k Transition Metal Complexes

David Balcells* and Bastian Bjerkem Skjelstad

Cite This: J. Chem. Inf. Model. 2020, 60, 6135−6146 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

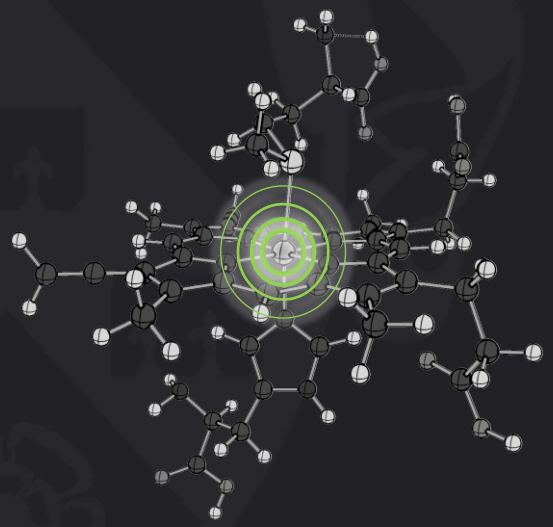
ABSTRACT: We report the transition metal quantum mechanics (tmQM) data set, which contains the geometries and properties of a large transition metal−organic compound space. tmQM comprises 86,665 mononuclear complexes extracted from the Cambridge Structural Database, including Werner, bioinorganic, and organometallic complexes based on a large variety of organic ligands and 30 transition metals (the 3d, 4d, and 5d from groups 3 to 12). All complexes are closed-shell, with a formal charge in the range {+1, 0, −1}e. The tmQM data set provides the Cartesian coordinates of all metal complexes optimized at the GFN2-xTB level, and their molecular size, stoichiometry, and metal node degree. The quantum properties were computed at the DFT(TPSSh-D3BJ/def2-SVP) level and include the electronic and dispersion energies, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies, HOMO/LUMO gap, dipole moment, and natural charge of the metal center; GFN2-xTB polarizabilities are also provided. Pairwise representations showed the low correlation between these properties, providing nearly continuous maps with unusual regions of the chemical space, for example, complexes combining large polarizabilities with wide HOMO/LUMO gaps and complexes combining low-energy HOMO orbitals with electron-rich metal centers. The tmQM data set can be exploited in the data-driven discovery of new metal complexes, including predictive models based on machine learning. These models may have a strong impact on the fields in which transition metal chemistry plays a key role, for example, catalysis, organic synthesis, and materials science. tmQM is an open data set that can be downloaded free of charge from https://github.com/bbskjelstad/tmqm.

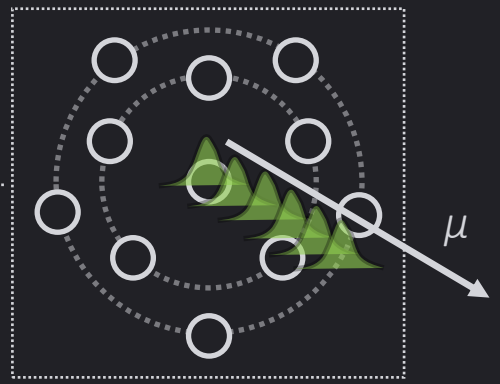
models are trained with data from quantum mechanical (QM) calculations. QML models are used to predict highest occupied molecular orbital (HOMO)/lowest unoccupied molecular orbital (LUMO) energies and gaps, dipole moments, polarizabilities, and other quantum properties governing the macroscopic behavior of chemical systems. State-of-the-art QML models, including atomistic[62] and message-passing neural networks,[63] yield predictions approaching chemical accuracy.[64] However, the training of these models requires quantum data sets that must be large and comprehensive to avoid overfitting and to ensure the unbiased exploration of the CCS. These data sets are scarce, and their generation remains hampered by the high computational cost of quantum mechanics calculations, thus limiting the scope of QML. Quantum data set examples include the Materials Project,[65] PubChemQC,[66] and the GDB13[67]-based QM series for organic chemistry (QM7,[54] QM7b,[68] QM8,[69,70] and QM9[71]). Ab initio molecular dynamics

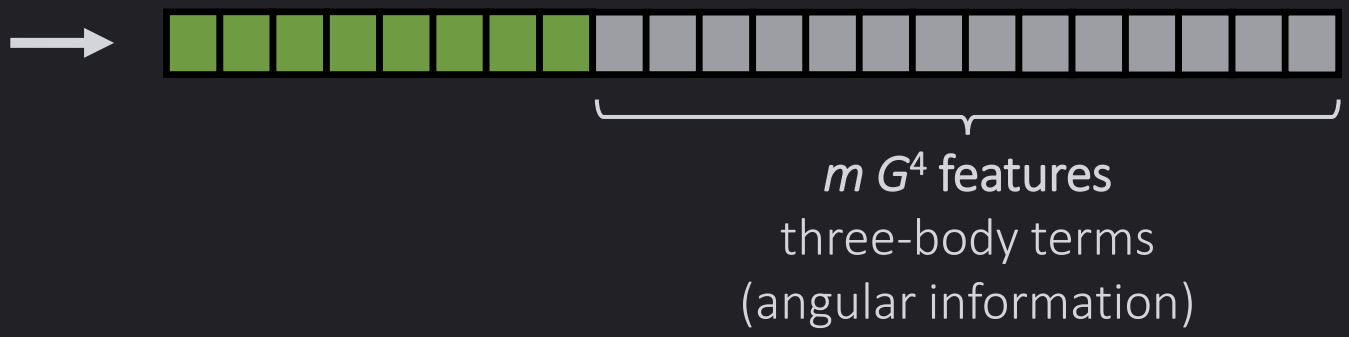
- the (GFN2-xTB optimised) Cartesian coordinates of 30k transition metal (Ti–Zn) complexes were selected as a subsample of the tmQM dataset
  - **XTB**: *github.com/grimme-lab/xtb*
- XAS were calculated using LSDA DFT and multiple scattering theory (MST) under the ‘muffin-tin’ approximation in FDMNES
  - **FDMNES**: *fdmnes.neel.cnrs.fr*

- the local environment around each X-ray absorption site is encoded as a feature vector of weighted atom-centred symmetry functions (WACSFs)

$$\boldsymbol{G}_i = \{G^2_{i,1}, \ldots, G^2_{i,n}, G^4_{i,1}, \ldots, G^4_{i,m}\}$$







G_0

cutoff sum $G_i^1 = \sum_{j \neq i} f_c(r_{ij})$

G_1

G_2

n radial functions
(two-body terms) $G_i^2 = \sum_{j \neq i} Z_j \cdot f_c(r_{ij}) \cdot e^{-\eta(r_{ij}-\mu)^2}$

G_n

internuclear distance information is encoded in these terms

G_{n+1}

G_{n+2}

m angular functions
(three-body terms) $G_i^4 = 2^{1-\zeta} \sum_{j \neq i} \sum_{k \neq i,j} Z_j Z_k \cdot (1 + \lambda \cos\theta_{jik})^{\zeta} \cdot f_c(r_{ij}) \cdot f_c(r_{ik}) \cdot f_c(r_{jk}) \cdot e^{-\eta(r_{ij}-\mu)^2} \cdot e^{-\eta(r_{ik}-\mu)^2} \cdot e^{-\eta(r_{jk}-\mu)^2}$

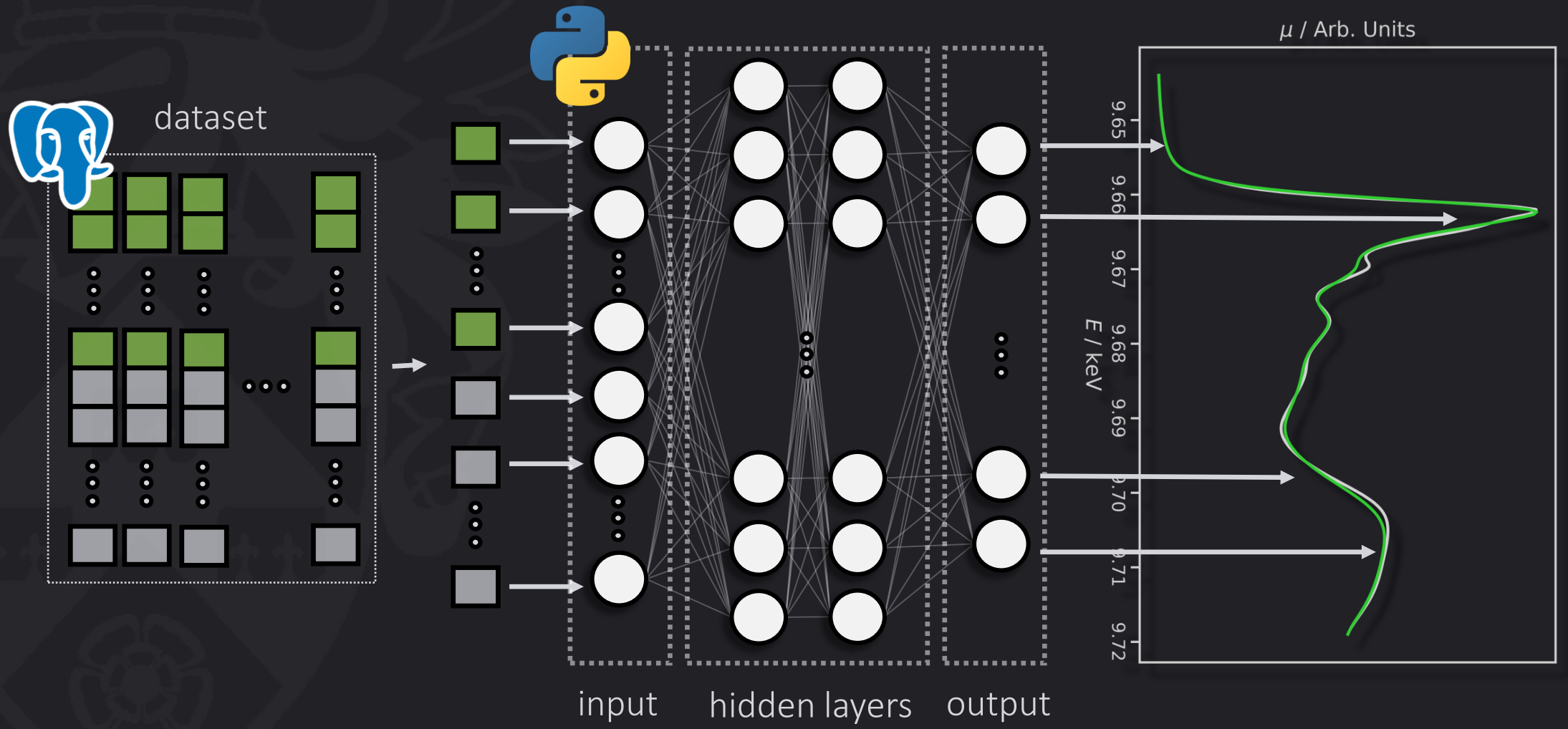
G_{n+m}

angular information is encoded in these terms

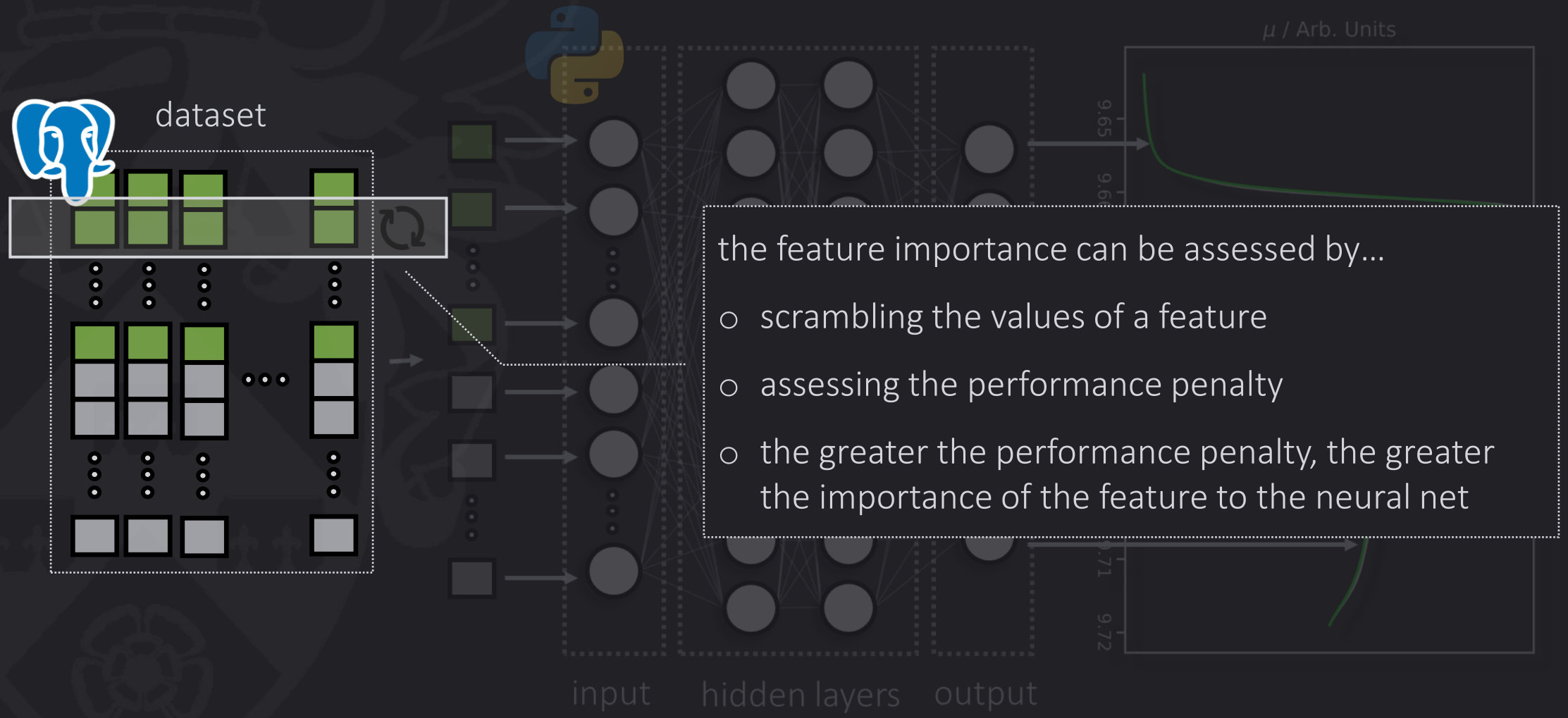
$$f_c(r_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1\right], & r_{ij} < r_c \\ 0, & r_{ij} > r_c \end{cases}$$

the cutoff function

Teaching Core-Hole Spectroscopy to a Deep Neural Network



Teaching Core-Hole Spectroscopy to a Deep Neural Network

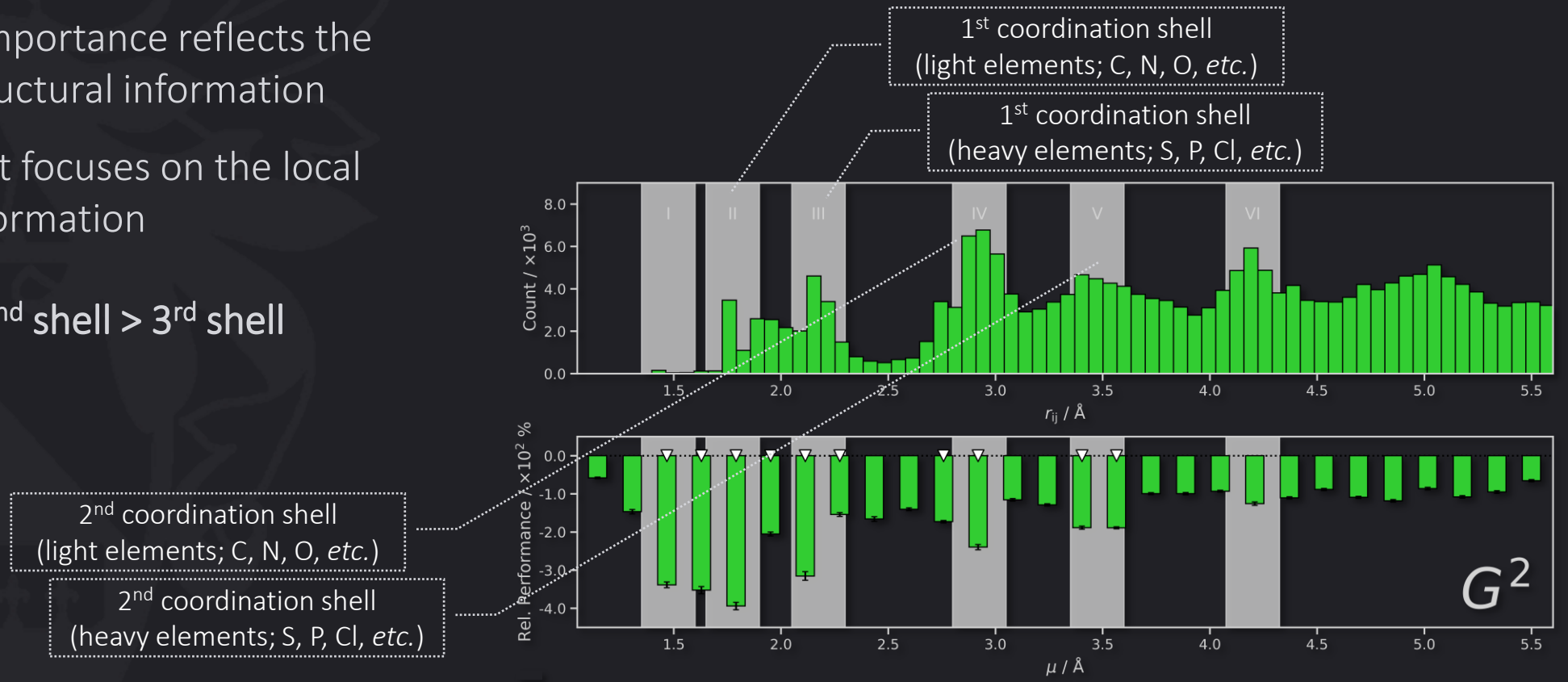


the feature importance can be assessed by...

- scrambling the values of a feature
- assessing the performance penalty
- the greater the performance penalty, the greater the importance of the feature to the neural net

- the feature importance reflects the density of structural information
- the neural net focuses on the local structural information

1st shell >> 2nd shell > 3rd shell

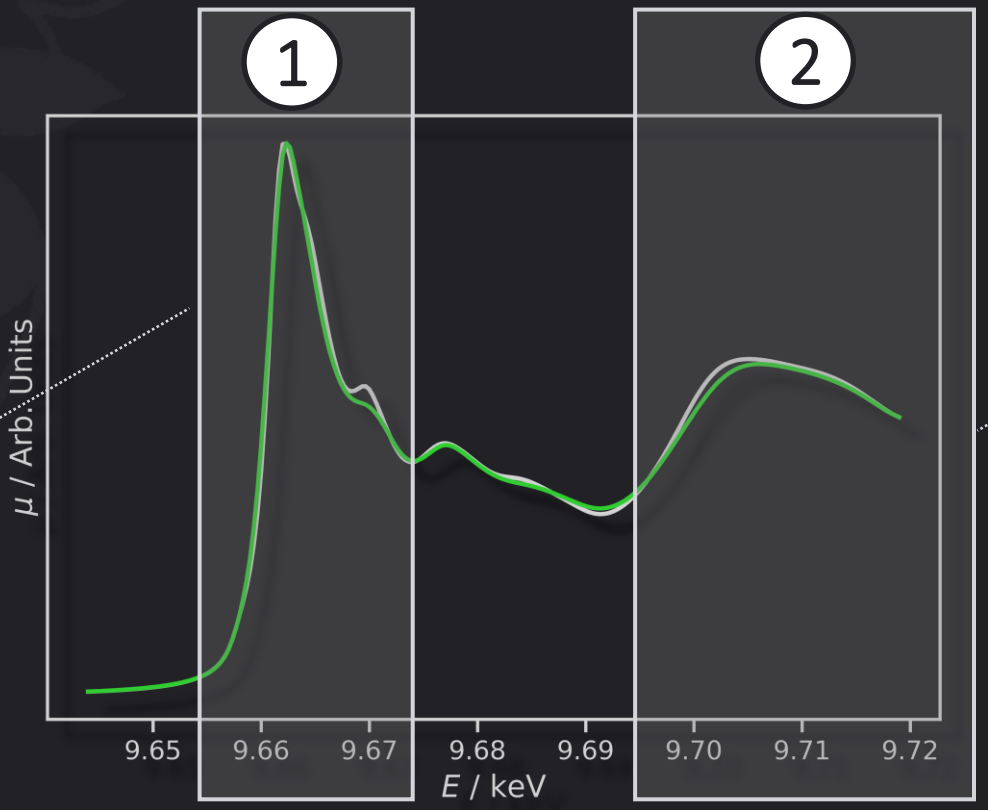


histogram (upper panel) of radial distances (r_{ij}) and bar plot (lower panel) of G^2 feature importance for G^2 features placed at a distance μ; data are for the Fe *K*-edge

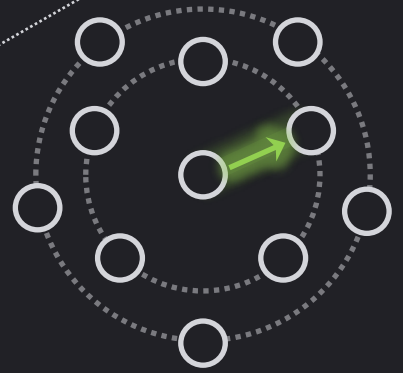
examining the feature importance over two spectral windows reveals
how the neural net reproduces more of the expected physics

low-energy photoelectrons
multiple scattering

low-energy window
−3.0 → +3.0 eV

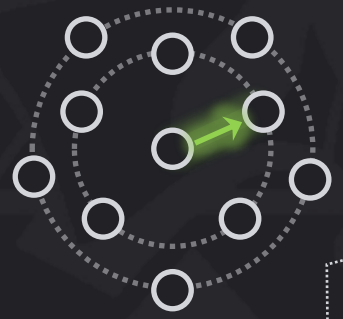


high-energy window
+50.0 → +56.0 eV



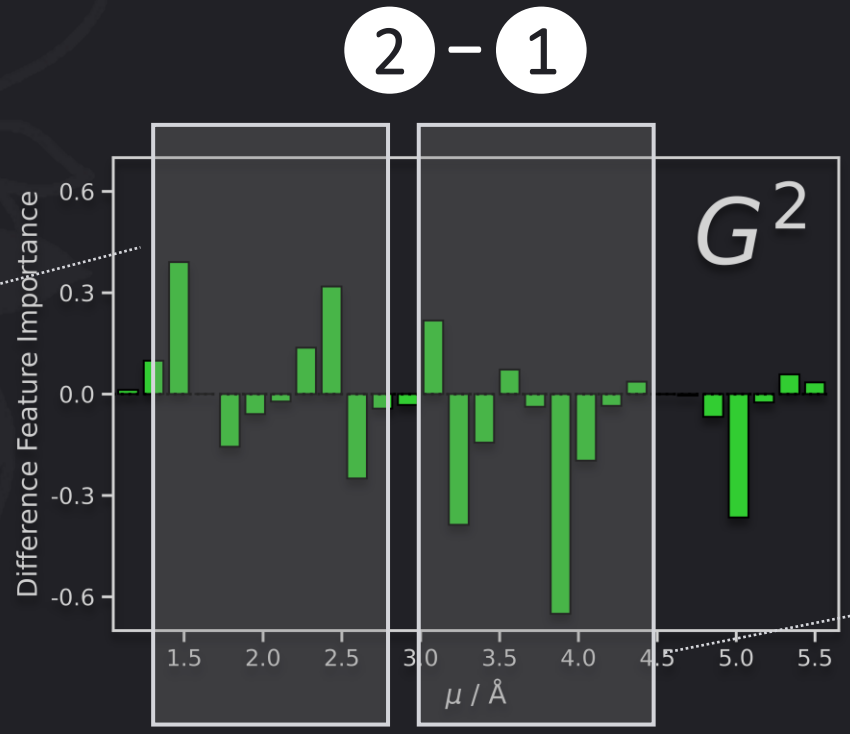
high-energy photoelectrons
single scattering

Teaching Core-Hole Spectroscopy to a Deep Neural Network



features near the X-ray absorption site

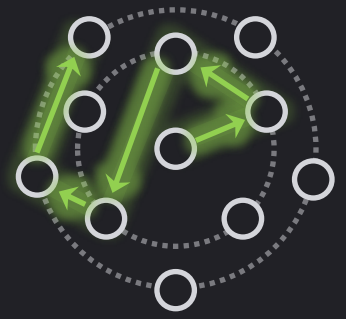
- higher feature importance in the **high-energy window** (+50.0 → +56.0 eV)
- single scattering



bar plot of difference (high-energy minus low-energy) G^2 feature importance for G^2 features placed at a distance μ; data are for the Fe *K*-edge

features far from the X-ray absorption site

- higher feature importance in the **low-energy window** (−3.0 → +3.0 eV)
- multiple scattering



TODAY

- accurate, affordable, fast, and easy predictions of theoretical XAS (and also XES)
- quantification of uncertainty in predictions
- proven applications to ‘real-world’ problems in chemistry and materials science
- explicit inclusion of electronic information, *e.g.* oxidation/charge/spin state, orbital information

TOMORROW

- approaches for data augmentation and intelligent/guided dataset growth
- transfer learning using experimental datasets

Thanks Everyone!

Conor Rankine	>	University of York (ex- Newcastle University)
Tom Penfold	>	Newcastle University
Marwah Madkhali	>	Newcastle University
Luke Watson	>	Newcastle University
Clelia Middleton	>	Newcastle University
Tom Pope	>	Newcastle University

Funding + Support

Newcastle University
University of York
EPSRC
Leverhulme Trust
The Alan Turing Institute
CONEXS