

Nathan Blake^{1,2}

David A. Kelly², Sarah Kapllani-Mucaj³, Rong Wang⁴, Geraint Thomas³, Hana Chockler²

¹ Department of Medical Physics and Biomedical Engineering, UCL, Gower Street, London WC1E 6BT, UK

² Department of Informatics, KCL, The Strand, London WC2R 2LS, UK

³ Department of Cell and Developmental Biology, UCL, Gower Street, London WC1E 6BT, UK

⁴ School of Dentistry, University of Missouri-Kansas City, 650 E 25th Street, Kansas City, MO 64108, USA

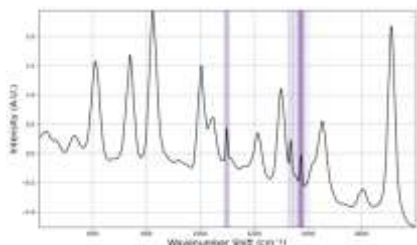
nathan.blake@kcl.ac.uk

Causal Responsibility Based Explainable AI for Raman Spectroscopy

Deep-learning classifiers are increasingly applied to Raman spectra, but their decisions are difficult to interpret at the spectral-feature level. There are both regulatory demands for more transparent AI (e.g. EU AI Act [1]), as well as domain user domain (e.g. clinicians). We introduce Spec-ReX, a causal-responsibility explainable AI method for Raman spectra, based on a computational framework called actual causality [2]. It is a post-hoc model-agnostic method which means it does not require access to model internals and can be applied to any AI model. Spec-ReX recursively partitions a spectrum into contiguous intervals, masks selected regions, and assigns causal responsibility according to the smallest intervention needed to change the model output [3].

However, it is notoriously difficult to robustly assess the performance of any explainable AI technique. A key impediment is separating the AI model itself from the explanations of that model, both of which have their own potential modes of failure. To assess Spec-ReX's suitability to biomedical tasks we developed a three-stepped approach in which an *in silico* dataset with a known ground truth signal is first utilised, followed by an *in vitro* dataset in which laboratory mixtures are created for a controlled but realistic dataset and finally assessing it against an *ex vivo* dataset of oesophageal tissue for cancer diagnostics which has no known ground truth signal. We compare Spec-ReX with gradient SHAP and Grad-CAM.

When a definitive ground truth signal is present Spec-ReX perform better across a range of performance metrics. When a ground truth signal is ambiguous or lacking Spec-ReX displays better specificity, meaning its attributions to spectral regions are more targeted while other methods have more diffuse attributions. These attributions explain trained model behaviour rather than biological causality, and their clinical value requires user-centred validation



References

- [1] European Parliament and Council of the European Union: Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union (2024), Article 13(1).
- [2] Halpern, J. Y.: Actual Causality. The MIT Press (2019).
- [3] Blake, Nathan, et al. "Causal responsibility based explainable AI for vibrational spectroscopy applied to oral FTIR and oesophageal Raman diagnostics." *Scientific Reports* (2026).