MARA-NIR: MASKED AUTOENCODER REPRESENTATION ANALYSIS OF NEAR-INFRARED SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Near-infrared (NIR) spectroscopy predicts the chemical composition of a sample from a single one-dimensional spectrum. Before calibration, chemometricians preprocess this spectrum with fixed formulas, such as scatter corrections and derivative filters. These formulas tell us in advance what a useful representation of NIR spectra computes, which is rare in representation learning. We use them to ask what a masked autoencoder learns without labels. We pretrain one on 7.2 million NIR spectra of food and feed and compare its encoder with the formulas, with untrained encoders and with supervised models. Its first layer rediscovers Savitzky-Golay derivative filters, which neither untrained encoders nor supervised models trained on soil carbon learn. Deeper in the encoder, soil organic carbon becomes more predictable, while untrained encoders stay flat. We then test the frozen encoder on the public LUCAS soil library, a new domain measured on a new instrument. With a linear model, it matches tuned chemometric models for soil organic carbon when all labels are used, and with 50 to 400 labels it stays within 0.05 in R^2 of the best of them.

1 INTRODUCTION AND RELATED WORK

Near-infrared (NIR) spectroscopy measures chemical composition from light absorption between roughly 780 and 2500 nm. The spectrum reports chemistry through overtone and combination bands of C-H, O-H and N-H bonds. The measurement is fast, inexpensive and non-destructive, and industry uses it to quantify moisture, protein, fat and carbon (Pasquini, 2018). Each measurement returns one spectrum for a whole sample, such as a cup of flour: a one-dimensional signal, not an image. The spectrum also carries physical effects unrelated to chemistry. Light scattering from particle size and packing adds sample-dependent offsets and slopes (Rinnan et al., 2009). Chemometrics removes these effects before calibration. Scatter corrections, such as the standard normal variate (SNV) (Barnes et al., 1989) and multiplicative scatter correction (MSC) (Geladi et al., 1985), and derivative filters (Savitzky & Golay, 1964; Norris & Williams, 1984) preprocess the spectrum. A linear model such as partial least squares (PLS) then fits the reference values (Wold et al., 2001).

Calibration has two bottlenecks. Reference values are expensive, because each requires a wet-chemistry analysis. Calibrations are also narrow. Practitioners build them per instrument, product and property, and transfer between instruments is possible but non-trivial. Labelled NIR datasets are therefore small and specific. Calibration methods are designed for this regime. They include PLS, rule-based Cubist (Quinlan, 1992), memory-based learning (MBL) (Ramírez-López et al., 2013) and convolutional networks (Cui & Fearn, 2018; Tsakiridis et al., 2020), each fitted to one task. Self-supervised learning can ease both bottlenecks. It learns a representation from unlabelled spectra, and small labelled sets then suffice for each calibration. Masked autoencoders (MAE) (He et al., 2022) learn such representations by reconstructing hidden parts of the input. Several studies have pretrained masked autoencoders on NIR spectra and applied them to soil, apples, resins and pharmaceutical tablets (Wan et al., 2023; Yang et al., 2023; Gilman et al., 2026). Self-supervised and semi-supervised objectives have also been applied to NIR spectra of fruit (Zhang et al., 2025; Wohlers et al., 2026), and NIR spectra of soil have been aligned to a representation pretrained on mid-infrared spectra (Sun et al., 2026). Each pretrains on the materials it evaluates and reports ac-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

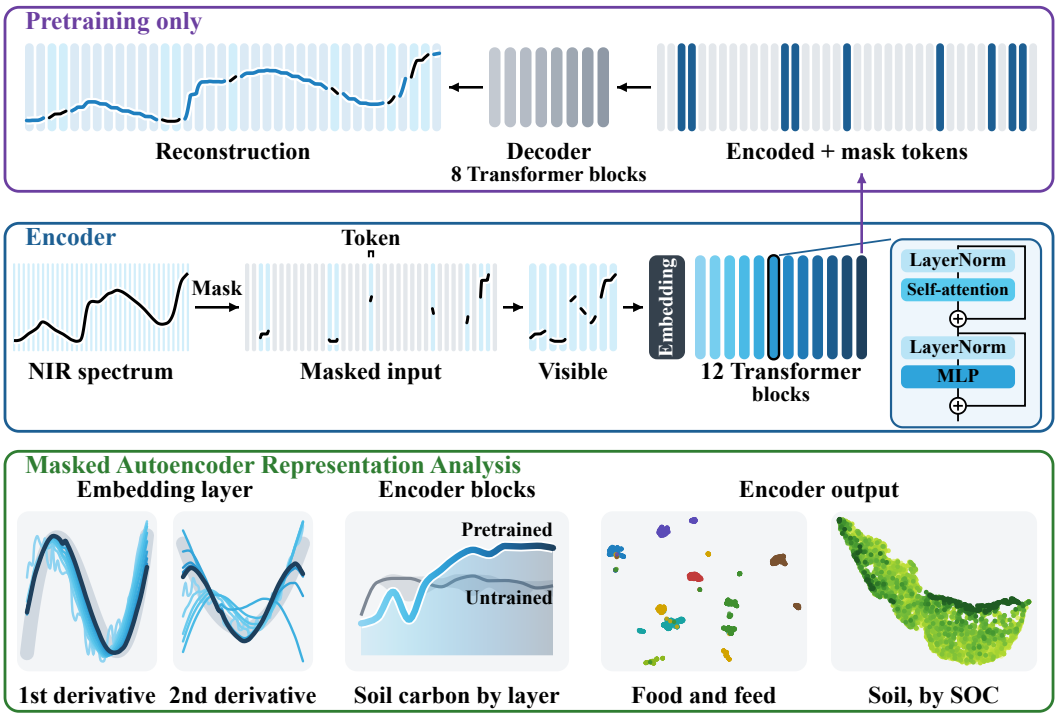

Figure 1: A masked autoencoder learns from NIR spectra by reconstructing hidden patches, and we then open its frozen encoder at three depths to see what it has learned: whether its first layer applies the derivative operators of chemometrics, whether soil organic carbon (SOC), the carbon in soil organic matter, becomes easier to read out with depth than in untrained encoders, and how its output arranges food, feed and soils.

curacy, not what the encoder learns. We pretrain on food and feed and test on soil, a new domain measured on a new instrument.

We study one-dimensional NIR spectra, one per sample, used to predict chemical properties. Self-supervised encoders also exist for spectral images (Baumann et al., 2026; Kong et al., 2025; Sumbul et al., 2025; Ji et al., 2026; Braham et al., 2025). There, every pixel of a scene carries a spectrum, and most models use spatial context. HyperSL (Kong et al., 2025) instead pretrains on the spectra of individual pixels of a scene. Here, one measurement is a single spectrum of a whole sample, without spatial context. Mid-infrared and Raman spectra are also one-dimensional, but they resolve fundamental vibrations as sharper bands. NIR bands are broad overtones that overlap, and scatter shifts and tilts the whole spectrum. This is why NIR calibration relies on scatter correction and derivatives, and these operations give us a target that the other modalities do not share. For infrared spectra, ACT (Huang et al., 2025) builds chemometric preprocessing into a supervised transformer as learnable modules. We test whether such operations emerge without labels.

Looking inside trained networks has a long history in vision. The first layer of an image network learns features that resemble Gabor filters (Yosinski et al., 2014). For vibrational spectra, the first convolutional layer of supervised networks can behave like a Savitzky-Golay derivative filter (Cui & Fearn, 2018; Zhang et al., 2020). Later work compared the layers of vision transformers and masked image models (Raghu et al., 2021; Park et al., 2023; Xie et al., 2023). Untrained networks serve as a control, because random features can already perform well (Saxe et al., 2011).

What pretrained NIR encoders learn remains open. We ask this question for a masked autoencoder pretrained on NIR spectra of food and feed. The preprocessing formulas above tell us in advance what a useful representation could contain, because chemometricians apply them to separate the chemical signal from physical effects. This gives us a known target to check the encoder against. We compare the encoder's filters with the derivative formulas, against filters of the same smoothness and against an untrained encoder. We measure at which depth soil chemistry becomes predictable, and

Table 1: Pretraining yields two to seven times more derivative filters than chance: share of first-layer filters matching a Savitzky-Golay template ($|r| > 0.9$).

Masking	Pretrained	Same smoothness	Untrained
75%	21%	10%	0%
90%	65%	9%	0%

when during pretraining this happens. Finally, we test the frozen encoder on soil, outside the food and feed domain of pretraining. The pretraining spectra come from a handheld MEMS instrument with a narrower band and a lower spectral resolution than the benchtop instrument of the soil spectra, which the encoder never saw (Beć et al., 2021; Digman et al., 2022). To the best of our knowledge, this is the first layer-by-layer analysis of what a self-supervised model learns from NIR spectra, on the largest corpus of NIR point spectra pretrained without labels. Figure 1 summarises the approach.

We find that the first layer learns derivative filters of the kind that chemometricians apply by hand, while supervised training of the same architecture on soil carbon learns none. Soil chemistry becomes more predictable with depth. The filters take their shape and the soil predictions settle within the first 300 of 800 pretraining epochs. On public soil data, measured on an instrument the encoder never saw, a frozen encoder with a linear model performs on par with tuned chemometric models for organic carbon, nitrogen and clay when all labels are used. With 50 to 400 labels, it stays within 0.05 in R^2 of the best tuned chemometric model.

2 Method and Data

We pretrain on about 7.2 million diffuse-reflectance spectra of food and feed materials, acquired with the ProxiScout, a handheld MEMS-based NIR instrument covering approximately 1350 to 2550 nm. The spectra are replicate scans of about 1.3 million measurements, and none carries a reference value. We remove spectra that fail a reference-free integrity check and apply no preprocessing to the remaining spectra. We resample every spectrum by linear interpolation onto a fixed wavenumber grid from 3800 to 25,000 cm^{-1} with a step of 2 cm^{-1}. An instrument covers only part of this grid. The first operation of the model normalises each spectrum to zero mean and unit variance over the covered points. We divide the grid into contiguous, non-overlapping patches of 48 points, each spanning 96 cm^{-1}. The ProxiScout covers about 37 of them. A linear projection maps the 48 values of a patch, together with the fraction of the patch that the instrument covers, to a token of width 768. We add to each token a fixed Fourier encoding (Tancik et al., 2020) of the wavenumber ν at the centre of the patch,

$$\gamma(\nu) = \left[\sin(2\pi f_k x),\ \cos(2\pi f_k x)\right]_{k=1}^{384}, \qquad x = \frac{\nu - \nu_{\min}}{\nu_{\max} - \nu_{\min}}, \tag{1}$$

where $\nu_{\min}$ and $\nu_{\max}$ are the ends of the grid and the frequencies f_k are spaced geometrically from 1 to 10^4. Unlike the index-based code of a vision transformer, this code depends only on the physical position of the patch. A patch centred at the same wavenumber therefore receives the same code on every instrument, and the encoder can read spectra from instruments with a different range. Patches outside an instrument's range are excluded from attention and from the mean over tokens. A standard pre-norm transformer encoder (Vaswani et al., 2017) of ViT-Base size (Dosovitskiy et al., 2021) processes the tokens, following the patch-based design that PatchTST (Nie et al., 2022) popularised for time series; here the sequence axis is wavenumber. We take the mean of the output patch tokens as the representation of the spectrum.

We pretrain the encoder with the masked autoencoder objective of He et al. (2022). We hide 75% of the patches at random. The encoder processes only the visible patches. A small decoder receives the encoded patches together with learnable mask tokens and reconstructs the hidden patches. We keep the training recipe of He et al. (2022) and train for 800 epochs. In each epoch, we draw one replicate scan per measurement at random. We save a checkpoint every 50 epochs. A sweep over masking ratios from 15% to 90% with a 100-epoch schedule places 75% and 90% ahead of lower ratios, at $R^2 = 0.43$ and 0.50 with 100 labels against at most 0.19 (Appendix B): a spectrum must be mostly hidden before reconstruction requires more than interpolation. We therefore also train a model with 90% masking for the full 800 epochs. Both models reach similar soil accuracy, and we keep 75%, the ratio of He et al. (2022) and ahead by 0.01 to 0.04 in R^2 at every label count, as

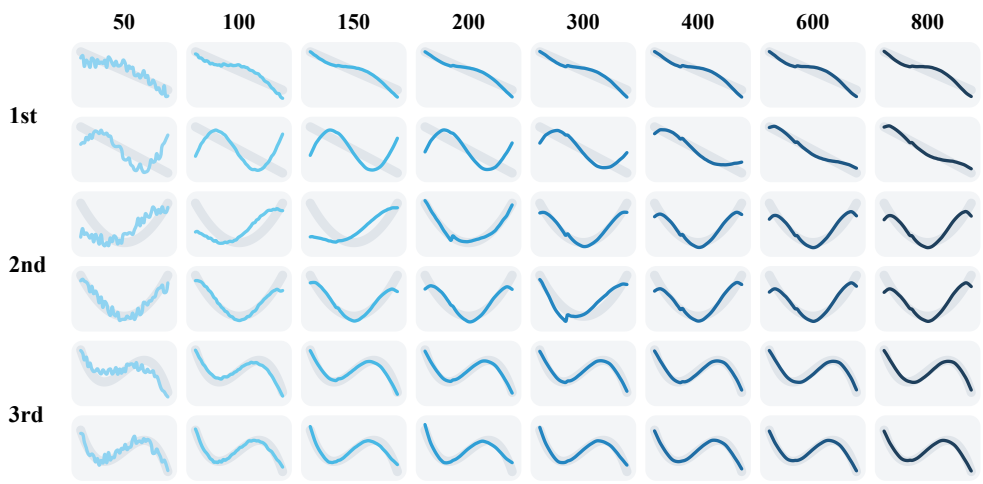

Figure 2: As pretraining proceeds, filters of the first layer gradually take the shape of the derivative operators that chemometricians apply by hand. Each row follows one filter of the model with 75% masking from epoch 50 to epoch 800, until it matches a first-, second- or third-derivative Savitzky-Golay template (grey), without any label.

the main model (Appendix B). As a control, we initialise three encoders of the same architecture at random and never train them. We refer to them as untrained encoders. Pretraining one model takes more than 100 GPU hours, in line with the cost of masked pretraining in vision (He et al., 2022). We therefore measure variance over ten draws of labelled samples, three untrained encoders and 16 checkpoints per model, from epoch 50 to epoch 800.

We evaluate the encoder on the 2009 topsoil library of the Land Use/Cover Area frame statistical Survey (LUCAS) (Orgiazzi et al., 2018), a public collection of about 19,000 soil samples from across the European Union. It is one of the largest public libraries of NIR spectra with reference values, and its organic matter absorbs in the same C-H, O-H and N-H bands as food and feed. Each sample has a laboratory spectrum from 400 to 2500 nm, measured with a FOSS XDS benchtop instrument, whose range starts about 950 nm below that of the ProxiScout, and wet-chemistry reference values, including soil organic carbon (SOC). We use the spectra as distributed, in absorbance, while the pretraining spectra are reflectance. The encoder reads the part of each spectrum between 1000 and 2500 nm. Unless stated otherwise, we predict SOC.

3 Verifying the Representation

We start with the first layer, where chemometric preprocessing would act. It maps each patch of 48 points to a token, and each of its 768 weight vectors is therefore a filter over 48 points, like a convolution kernel in a CNN. We compare every filter with Savitzky-Golay templates of the same

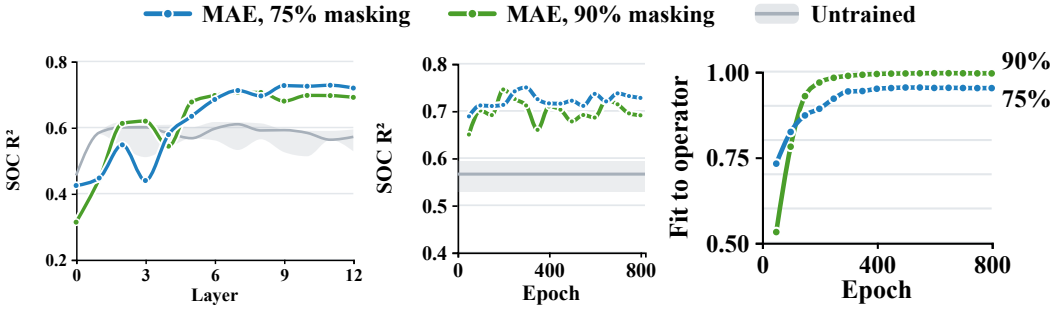

Figure 3: Pretraining makes soil carbon easier to read out in the deeper layers, where untrained encoders (grey band) stay flat (left), and both the soil predictions at the output (middle) and the shape of the derivative filters (right) settle early in pretraining. SOC R^2 from ridge models with 100 labelled LUCAS samples; layer 0 is the patch embedding.

length (Savitzky & Golay, 1964), for polynomial orders 2 to 5 and derivative orders 0 to 3, and count a match when the absolute correlation exceeds 0.9. Because a smooth filter can match a smooth template by chance, two controls calibrate this count: the untrained encoders, and phase-randomised copies of the learned filters, which have the same smoothness but a random shape. Pretraining produces two to seven times more matches than filters of the same smoothness (Table 1). In the model with 75% masking, 21% of the filters match, against 10% for filters of the same smoothness and none in the untrained encoders; with 90% masking, 65% match, against 9% and none. At a threshold of 0.95, 11% and 58% match, against 3% (Appendix C). The matching filters are derivatives of order 1 to 3, the operators that chemometricians apply against baseline offsets and slopes, and Figure 2 shows six of them as they form. The layer as a whole is low-rank: four directions carry 99.9% of the energy of the weights, against 12% in an untrained encoder, and polynomials of degree four or lower hold 99.5% and 99.8% of it in the two models. A Savitzky-Golay filter works in the same way: it fits a low-degree polynomial to its window and reads off a smoothed value or a derivative. The first layer therefore fits each patch with a low-degree polynomial, and each matching filter reads one value or derivative from that fit. The smoothness of NIR spectra explains part of this, as three of the four directions coincide with the leading principal components of the input patches, but the learned directions are more polynomial, with 95% and 99.6% of their energy in cubic polynomials against 78% for the components.

We then follow soil chemistry through the deeper layers. At each layer, we average the tokens of a spectrum and fit a ridge model to soil organic carbon on LUCAS with 100 labelled samples, averaged over 10 random draws (Figure 3, left). In the untrained encoders, R^2 stays between 0.51 and 0.62 at every layer after the patch embedding. The pretrained encoders start lower, at 0.42 and 0.31, and rise through the blocks; they exceed the untrained encoders from block 5 and level off between 0.68 and 0.73 from block 7. Soil chemistry therefore becomes more predictable in the deeper blocks, not in the first layer (Appendix D).

We then test how the embedding responds to changes that leave the chemistry of a sample unchanged, as the cosine distance between the embeddings of a held-out spectrum before and after, relative to the mean distance between different samples (Appendix E). Scale and offset have no effect on any encoder, because the input normalisation removes them. A repeated scan of the same sample moves the embedding of the model with 75% masking by 0.7% of the distance between samples. A shift of the bands by up to 8 cm^{-1}, a simple model of the band changes that temperature causes (Wülfert et al., 1998), moves it by 2%, against 0.1% for the untrained encoders, in line with derivative filters that respond to band position. Extended multiplicative scatter correction (Martens & Stark, 1991) moves every encoder, trained or not, by about 40%: the encoder keeps the information that this correction removes. The encoder is also sensitive to white noise, which at 1% of the standard deviation of the spectrum moves the embedding by 6%.

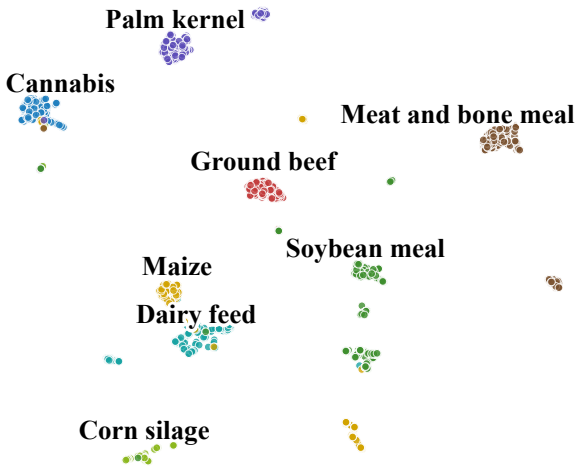

Figure 4: Without any label, the embedding keeps eight food and feed materials apart: UMAP of 2,676 held-out measurements, model with 75% masking.

Table 2: In the full embedding, measurements of one material lie closer together, a second scan finds the first more often, and the neighbours of a soil predict its properties better than with the spectra or untrained encoders. Silhouette and 10-nearest-neighbour accuracy on held-out food and feed; retrieval: second scan is the nearest neighbour; soil: 10-nearest-neighbour R^2 on LUCAS.

Metric	75%	90%	Untrained	Spectra	Spectra, PCA
Silhouette, material	0.09	0.13	−0.18	−0.11	−0.04
Silhouette, family	0.23	0.27	0.04	0.10	0.09
Accuracy, material	0.72	0.74	0.62	0.65	0.65
Retrieval	0.79	0.79	0.58	0.67	0.66
Soil organic carbon	0.85	0.86	0.79	0.78	0.79
Clay	0.58	0.59	0.53	0.54	0.58
pH	0.67	0.67	0.57	0.63	0.65
Calcium carbonate	0.69	0.71	0.53	0.64	0.70

Finally, we measure how the embedding organises samples, in the full 768-dimensional space with cosine distance (Table 2). On 6,761 held-out measurements of 40 food and feed materials, a 10-nearest-neighbour classifier identifies the material of 72% of the measurements from the embedding of the model with 75% masking, against 62% for the untrained encoders and 65% for the normalised spectra, at a chance level of 2.5%. A second scan of a measurement is its nearest neighbour in 79% of cases, against 58% and 67%, and on LUCAS a 10-nearest-neighbour regression of soil organic carbon reaches $R^2 = 0.85$, against 0.79 and 0.78. Chemically similar products lie close together in every representation, and the encoder preserves this structure: the nearest other material lies in the same product family in 88% of cases, against 12% by chance. Soils lie away from the food and feed materials in every representation, and the encoder places them closer: the mean distance from a soil to the nearest food or feed material is 13 times the distance between neighbouring materials, against 278 times for the spectra (Appendix F; Figure 4 shows a UMAP projection (McInnes et al., 2018)).

4 Learning Dynamics

We now ask when, during pretraining, these properties appear, and repeat two measurements on the checkpoints saved every 50 epochs. The soil predictions at the output of the encoder settle early (Figure 3, middle): after only 50 epochs, the model with 75% masking already reaches $R^2 = 0.69$, and over the remaining 750 epochs it moves only between 0.69 and 0.75, as the model with 90% masking does between 0.65 and 0.75. The derivative filters take longer (Figure 3, right). Their median correlation with their final template rises from 0.73 to 0.95 with 75% masking and from

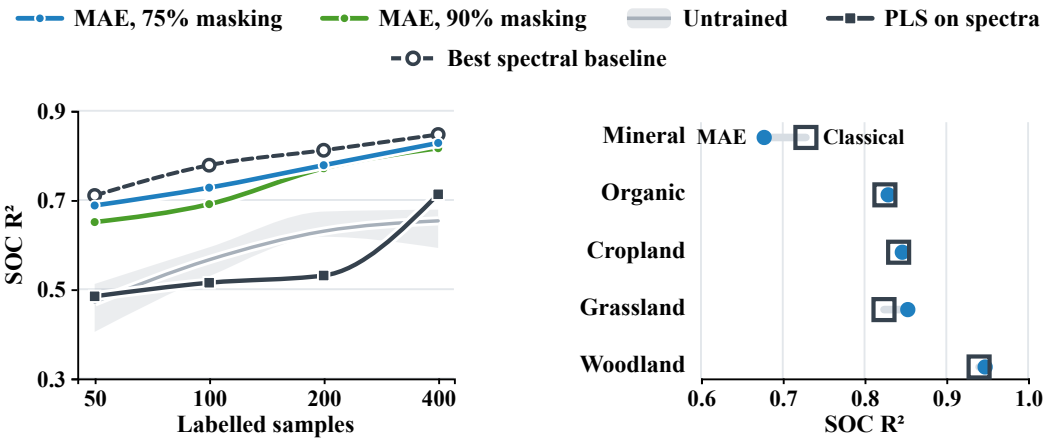

Figure 5: Pretrained on food and feed, the frozen encoder carries over to soil measured on another instrument. With 50 to 400 labels (left), one ridge model on the embedding stays within 0.02 to 0.05 of the best chemometric model tuned on the spectra (Cubist), and exceeds PLS with cross-validated preprocessing by 0.12 to 0.25 and the untrained encoders (grey band). With all labels (right), it matches or exceeds the best of five chemometric models on four of five subsets.

Table 3: The embedding matches chemometric models for organic carbon, nitrogen and clay, the properties tied to organic matter and texture, and falls behind for pH and calcium carbonate. Mean R^2 over the five subsets with all labels, and subsets where the embedding matches or exceeds the best chemometric model per subset, at three decimals.

Property	Embedding	Best chemometric	Subsets
Organic carbon	0.83	0.83	4 of 5
Nitrogen	0.78	0.78	4 of 5
Clay	0.80	0.77	4 of 5
pH	0.77	0.86	0 of 5
Calcium carbonate	0.78	0.86	0 of 5

0.53 to 0.99 with 90% masking, and both curves level off by epoch 300. Within the first 300 of 800 epochs, the matching filters therefore reach their final shape and the soil predictions settle. On ImageNet, in contrast, linear-probe accuracy keeps improving up to 1600 epochs (He et al., 2022).

5 Transfer to Soil

We then evaluate the frozen encoder on LUCAS, soil measured on another instrument, and first compare it with chemometric models when all labels are used. We follow the design of Stevens et al. (2013), which splits LUCAS into mineral and organic soils and, separately, into cropland, grassland and woodland. Within each subset, the Kennard-Stone algorithm (Kennard & Stone, 1969) selects two thirds of the samples for calibration and the rest serve as test; as the original split is not published, all models share our own. The chemometric models receive the preprocessing of Stevens et al. (2013), and for each subset we keep the best of PLS, ridge regression, Cubist, memory-based learning and a 1D CNN, while the embedding gets a single ridge model (Hoerl & Kennard, 1970). Against this per-subset choice, the embedding matches or exceeds the best chemometric model on organic soils, cropland, grassland and woodland (Figure 5, right), and falls behind only on mineral soils, with $R^2 = 0.68$ against 0.73; Appendix G reports every model on every subset.

The result extends to other soil properties (Table 3). For nitrogen and clay as well, the embedding matches or exceeds the best chemometric model on four of five subsets, while for pH and calcium carbonate the chemometric models stay ahead on every subset, by 0.09 and 0.08 in R^2 on average. The embedding also carries information that a linear chemometric model does not: after removing what 15 PLS components of the same input explain, it still predicts every property (Appendix G).

A new calibration often starts with a few dozen labels, the regime in which a pretrained representation can help most. We therefore draw between 25 and 400 labelled samples at random and test on a held-out quarter of LUCAS, averaged over 10 draws. On the embedding we fit a ridge model, which

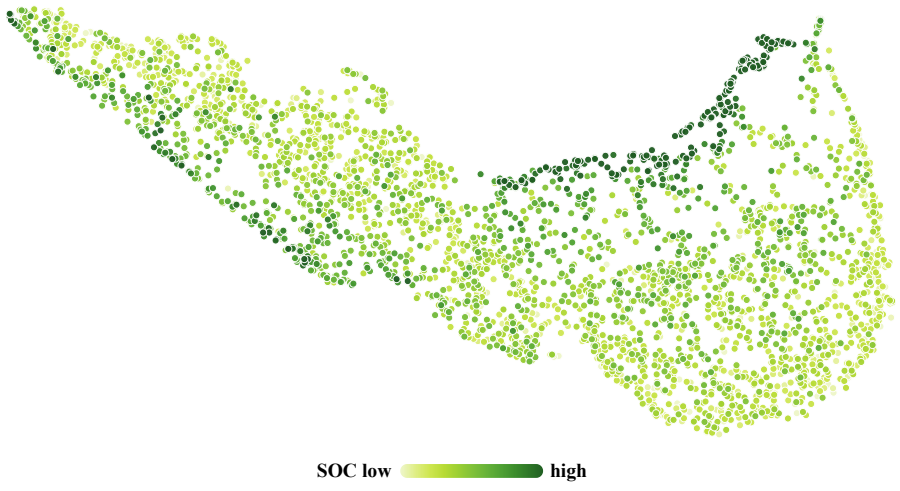

Figure 6: The soils richest in organic carbon gather along one edge of the embedding, although the encoder never saw a carbon value: UMAP of 3,500 LUCAS soils, model with 75% masking.

Table 4: Supervised models trained on soil carbon predict it as well as the masked autoencoder, but they learn no derivative filters and build a different representation. Filters: share matching a Savitzky-Golay template ($|r| > 0.9$). CKA: linear centred kernel alignment with the masked autoencoder at layer 12 (CNN: last hidden layer against its most similar layer). Band shift: embedding change relative to the distance between samples, at the same layers. R^2: soil organic carbon on a held-out fifth of LUCAS, all other labels.

	Masked autoencoder	Supervised transformer	Supervised CNN	Untrained
Filters	21%	0%	0%	0%
CKA with the MAE	1	0.09	0.31	0.93–0.94
Band shift	1.1%	5.3%	3–6%	0.06%
R^2, all labels	0.90	0.92	0.89–0.90	0.89

stays ahead of PLS, Cubist and a small neural network on the same embedding from 50 labels on. The spectral baselines are tuned extensively: we tune ridge and PLS over two spectral ranges and SNV with or without a first or second Savitzky-Golay derivative at several windows and scalings, fit Cubist, memory-based learning and a 1D CNN to first-derivative spectra, and report the best at each size, chosen on the test set. The best is Cubist at every size (Figure 5, left). From 50 labels on, the single ridge model on the embedding stays within 0.02 to 0.05 in R^2 of it, at 0.69 against 0.71 with 50 labels and 0.83 against 0.85 with 400, and exceeds PLS with cross-validated preprocessing, the standard chemometric workflow, by 0.12 to 0.25. The untrained encoders reach only 0.40 to 0.51 with 50 labels and stay below the pretrained encoders at every size. With 25 labels, the tuned spectral models lead (Appendix H).

Figure 6 shows the soil embedding in a UMAP projection (McInnes et al., 2018), coloured by soil organic carbon, a value the encoder never saw.

6 Discussion

Our central finding is that masked pretraining rediscovers the preprocessing that chemometricians apply by hand, and that supervised training on the same data does not (Figure 7). We trained the same transformer, and a 1D CNN with 32 first-layer filters of 48 points, to predict soil organic carbon on LUCAS (Appendix I). Both predict it as well as the encoder, yet neither learns a single filter that matches a Savitzky-Golay template, even at a threshold of 0.75 (Table 4).

The filters of the supervised transformer barely move from their initialisation, with a median cosine similarity of 0.97 to their initial weights, although it reaches the highest R^2. Supervised training on one property therefore does not produce these operators, whereas masked reconstruction of the spectrum does. Earlier work observed derivative-like behaviour in supervised networks, in the output of a convolutional layer for NIR calibration on one of two datasets (Cui & Fearn, 2018) and in the first layer for Raman and mid-infrared spectra (Zhang et al., 2020). Our test compares the weights with the templates, against two controls, for a network trained without labels.

The two kinds of training also produce different representations (Figure 8). Compared layer by layer with linear centred kernel alignment (CKA) (Kornblith et al., 2019), the masked autoencoder stays close to an untrained encoder of the same architecture and keeps the geometry of the input spectra,

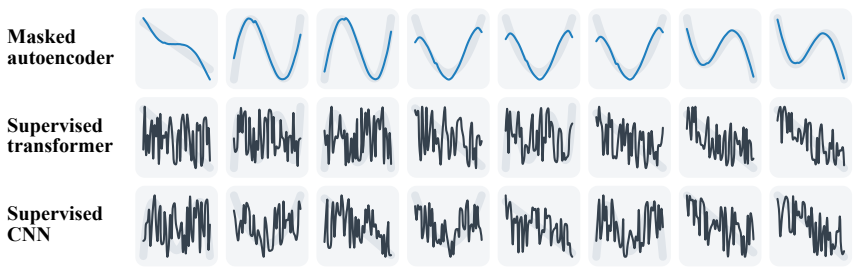

Figure 7: Only masked pretraining turns first-layer filters into derivative operators; filters of supervised models trained on soil carbon stay noisy. Each filter is drawn over its best-matching Savitzky-Golay template (grey): for the masked autoencoder, filters with $|r| > 0.9$; for the supervised models, filters at evenly spaced quantiles of their best correlation.

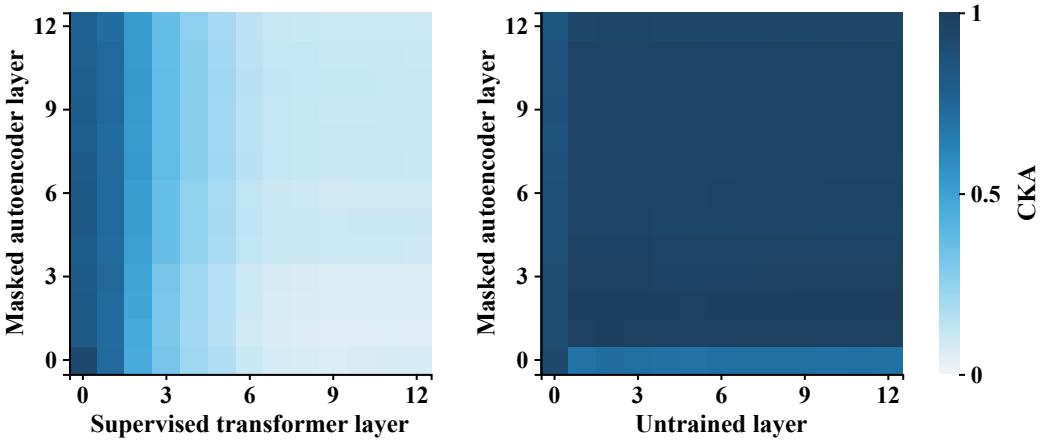

Figure 8: Supervision reshapes the deep layers, while masked pretraining keeps them close to an untrained encoder of the same architecture: linear CKA between the layers of the masked autoencoder and those of the supervised transformer (left) or an untrained encoder (right), on 960 LUCAS spectra.

whereas the supervised transformer reorganises its deep layers around the target. Both supervised models end far from the masked autoencoder, with a CKA of 0.09 and 0.31 (Table 4). This mirrors vision, where masked image modelling keeps local, high-frequency information while contrastive learning builds global features (Park et al., 2023), and supervised models lose the diversity of their attention heads in the last layers (Xie et al., 2023). The supervised transformer is also more sensitive to band shifts than the masked autoencoder, relative to the spread between samples.

Much of the soil information is, however, available without any pretraining. With all labels, a ridge model on an untrained encoder reaches $R^2 = 0.89$ on LUCAS, within 0.01 to 0.03 of the pretrained encoder and the supervised models (Table 4). The architecture and the input normalisation already give a representation from which a linear model with enough labels recovers soil organic carbon, in line with the strength of random features (Saxe et al., 2011). Pretraining helps most when labels are few (Section 5), where the pretrained encoders lead the untrained ones by 0.18 to 0.29 in R^2 with 50 labels, the regime in which new calibrations start.

7 Limitations and Future Work

Our pretraining data come from a single instrument family, and how the learned operators change when pretraining mixes instruments remains open. The next step is to pretrain across many instruments and on materials beyond food and feed, towards a foundation model for NIR spectra. Public libraries of NIR point spectra with reference values are scarce and mostly small, which limits our evaluation to LUCAS, one of the largest of them. The analysis itself is not tied to masked autoencoders: any encoder of NIR spectra can be checked against the same operators.

8 Conclusion

We asked what a masked autoencoder learns from NIR spectra. Pretrained on food and feed without labels, the encoder rediscovers the preprocessing that chemometricians apply by hand: its first layer fits each patch with the low-degree polynomials of Savitzky-Golay filters and reads off their derivatives, which supervised training on soil carbon does not produce. Its deeper layers make soil organic carbon easier to read out than those of untrained encoders of the same architecture, and both the filters and the soil predictions settle within the first 300 of 800 epochs. On public soil data from a benchtop instrument it never saw, the frozen encoder with one linear model performs on par with tuned chemometric models for organic carbon, nitrogen and clay when all labels are used, and with 50 to 400 labels it stays within 0.02 to 0.05 in R^2 of the best chemometric model tuned on the soil spectra. NIR spectra therefore offer a rare test bed in which what a self-supervised model learns can be checked against operators known in closed form.

AI USE STATEMENT

In this work, we used large language models (LLMs), Claude Opus 4.6 and Claude Opus 5.5 through Claude Code, for the following tasks. Implementing methods: the LLMs wrote and ran code for the analyses and evaluations to the authors' specification. The authors reviewed the code and checked the reported numbers against the result files. Design of experiments and interpretation of results: the LLMs suggested controls and baselines and proposed readings of results; the authors decided which to use and checked every interpretation against the data. Literature: the LLMs searched for and summarized related work; the authors read the cited works and added all references by hand. Writing: the LLMs drafted and edited parts of the text; the authors reviewed, edited and approved all of it. No figure, image or data set in this paper was generated by an LLM; all figures plot our own results. The research question, the choice of model and data, and the conclusions are the authors'. The authors take full responsibility for the content of the paper, in line with the ICLR 2027 AI policy.

ETHICS STATEMENT

The pretraining spectra are measurements of food and feed materials and contain no personal data. The evaluation uses the public LUCAS 2009 topsoil library under its terms of use. The work involves no human subjects and no animal experiments.

REPRODUCIBILITY STATEMENT

Section 2 describes the model and pretraining, which follow He et al. (2022), and Appendix A lists every setting. All evaluations use the public LUCAS 2009 library. Sections 3 to 5 and Appendices G and H specify the splits, preprocessing and models. The supplementary material contains the code for pretraining, analysis and evaluation, the first-layer weights and the frozen LUCAS embeddings of both pretrained encoders, and a guide that maps each script to the table or figure it produces. With these files and the public LUCAS data, the first-layer analysis and all soil results can be repeated. The untrained encoders are rebuilt from the released configuration. A public API, free of charge for anyone, at `https://mara-nir.delightfulcoast-ab42997a.westeurope.azurecontainerapps.io`, returns the frozen embeddings of both encoders for any NIR spectrum, at the output and at every layer, which also allows the layer-wise analyses to be repeated. The supplementary material documents its use. During review, the API is hosted anonymously. It accepts spectra on a wavelength or wavenumber axis.

REFERENCES

R. J. Barnes, M. S. Dhanoa, and Susan J. Lister. Standard normal variate transformation and detrending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43(5):772–777, 1989.

Alexander Baumann, Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Alexander Studier-Fischer, Berkin Özdemir, Lena Maier-Hein, and Slobodan Ilic. Carl: Camera-agnostic representation learning for spectral image analysis. In *International Conference on Learning Representations*, volume 2026, pp. 23335–23357, 2026.

Krzysztof B Beć, Justyna Grabska, and Christian W Huck. Principles and applications of miniaturized near-infrared (nir) spectrometers. *Chemistry–A European Journal*, 27(5):1514–1532, 2021.

Nassim Ait Ali Braham, Conrad M Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. Spectralearth: Training hyperspectral foundation models at scale. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:16780–16797, 2025.

Chenhao Cui and Tom Fearn. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems*, 182:9–20, 2018.

Matthew F Digman, Jerry H Cherney, and Debbie JR Cherney. The relative performance of a benchtop scanning monochromator and handheld fourier transform near-infrared reflectance spectrometer in predicting forage nutritive value. *Sensors*, 22(2):658, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Paul Geladi, Duncan MacDougall, and Harald Martens. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39(3):491–500, 1985.

Lauren Gilman, Hui Wang, Nick Birse, and Louise Manning. Sstlnetwork: a self-supervised spectral reconstruction network with hybrid attention for near-infrared spectral transfer learning. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, pp. 128082, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Shiluo Huang, Yining Jin, Wei Jin, and Ying Mu. Analytical-chemistry-informed transformer for infrared spectra modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17440–17448, 2025.

Renjie Ji, Xue Wang, Chao Niu, Wen Zhang, Yong Mei, and Kun Tan. Specaware: a spectral-content aware foundation model for unifying multi-sensor learning in hyperspectral remote sensing mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 234:242–260, 2026.

Ronald W Kennard and Larry A Stone. Computer aided design of experiments. *Technometrics*, 11 (1):137–148, 1969.

Weili Kong, Baisen Liu, Xiaojun Bi, Changdong Yu, Xinyao Li, and Yushi Chen. Hypersl: A spectral foundation model for hyperspectral image interpretation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Harald Martens and Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy. *Journal of pharmaceutical and biomedical analysis*, 9(8):625–635, 1991.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Karl H. Norris and Phil C. Williams. Optimization of mathematical treatments of raw near-infrared signal in the measurement of protein in hard red spring wheat. I. influence of particle size. *Cereal Chemistry*, 61(2):158–165, 1984.

Alberto Orgiazzi, Cristiano Ballabio, Panagiotis Panagos, Arwyn Jones, and Oihane Fernández-Ugalde. Lucas soil, the largest expandable soil dataset for europe: a review. *European Journal of Soil Science*, 69(1):140–153, 2018.

Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023.

Celio Pasquini. Near infrared spectroscopy: A mature analytical technique with new perspectives – a review. *Analytica Chimica Acta*, 1026:8–36, 2018.

J. Ross Quinlan. Learning with continuous classes. *Proc. Australian Joint Conf. on Artificial Intelligence*, pp. 343–348, 1992.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

L. Ramírez-López, T. Behrens, K. Schmidt, A. Stevens, J. A. M. Demáttê, and T. Scholten. The spectrum-based learner: A new local approach for modeling soil vis–nir spectra of complex datasets. *Geoderma*, 195–196:268–279, 2013.

Åsmund Rinnan, Frans Van Den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10): 1201–1222, 2009.

Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

Andrew M Saxe, Pang Wei Koh, Zhenghao Chen, Maneesh Bhand, Bipin Suresh, and Andrew Y Ng. On random weights and unsupervised feature learning. In *Icml*, volume 2, pp. 6, 2011.

Antoine Stevens, Marco Nocita, Gergely Tóth, Luca Montanarella, and Bas van Wesemael. Prediction of soil organic carbon at the european scale by visible and near infrared reflectance spectroscopy. *PLoS ONE*, 8(6):e66409, 2013.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. Smarties: Spectrum-aware multisensor auto-encoder for remote sensing images. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5569–5578. IEEE, 2025.

Luning Sun, José L Safanelli, Jonathan Sanderman, Katerina Georgiou, Colby Brungard, Kanchan Grover, Bryan G Hopkins, Shusen Liu, and Peer-Timo Bremer. Self-supervised and multi-fidelity learning for extended predictive soil spectroscopy. *Geoderma*, 468:117764, 2026.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Nikolaos L. Tsakiridis, Konstantinos D. Keramaris, John B. Theocharis, and George C. Zalidis. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multichannel 1-D convolutional neural network. *Geoderma*, 367:114208, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Midi Wan, Taiyu Yan, Guoxia Xu, Aibing Liu, Yangbin Zhou, Hao Wang, and Xiu Jin. MAE-NIR: A masked autoencoder that enhances near-infrared spectral data to predict soil properties. *Computers and Electronics in Agriculture*, 215:108427, 2023. doi: 10.1016/j.compag.2023.108427.

Mark Wohlers, Andrew McGlone, Eibe Frank, and Geoffrey Holmes. Barlow twins for semi-supervised learning in nir spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, pp. 105664, 2026.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, 2001.

Florian Wülfert, Wim Th Kok, and Age K Smilde. Influence of temperature on vibrational spectra and consequences for the predictive ability of multivariate models. *Analytical chemistry*, 70(9): 1761–1767, 1998.

Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14475–14485. IEEE, 2023.

Yu Yang, Shangpeng Sun, Min Huang, and Qibing Zhu. Peamatl: A strategy for developing near-infrared spectral prediction models under domain shift using self-supervised transfer learning. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2023.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

Liu Zhang, Jincun Liu, Yaoguang Wei, Dong An, and Xin Ning. Self-supervised learning-based multi-source spectral fusion for fruit quality evaluation: A case study in mango fruit ripeness prediction. *Information Fusion*, 117:102814, 2025.

Xiaolei Zhang, Jinfan Xu, Jie Yang, Li Chen, Haibo Zhou, Xiangjiang Liu, Haifeng Li, Tao Lin, and Yibin Ying. Understanding the learning mechanism of convolutional neural networks in spectral analysis. *Analytica Chimica Acta*, 1119:41–51, 2020.

A PRETRAINING SETTINGS

We follow the image recipe closely. We keep the recipe of He et al. (2022) for a ViT-Base encoder and change only what the input requires: the patch length and the positional code of Section 2. Table 5 lists every setting of the model with 75% masking. The model with 90% masking differs only in the masking ratio. As in He et al. (2022), the decoder reconstructs each hidden patch after normalising it to zero mean and unit variance, and the loss covers the hidden patches only. One epoch draws one replicate scan of every measurement at random, and the replicates therefore act as data augmentation. Figure 9 shows the reconstruction loss of the model with 75% masking over all 800 epochs. The validation loss keeps decreasing slowly and reaches its minimum at epoch 767. The loss therefore keeps improving long after the soil predictions and the derivative filters have settled (Section 4), and it is therefore not a guide to when the representation is ready.

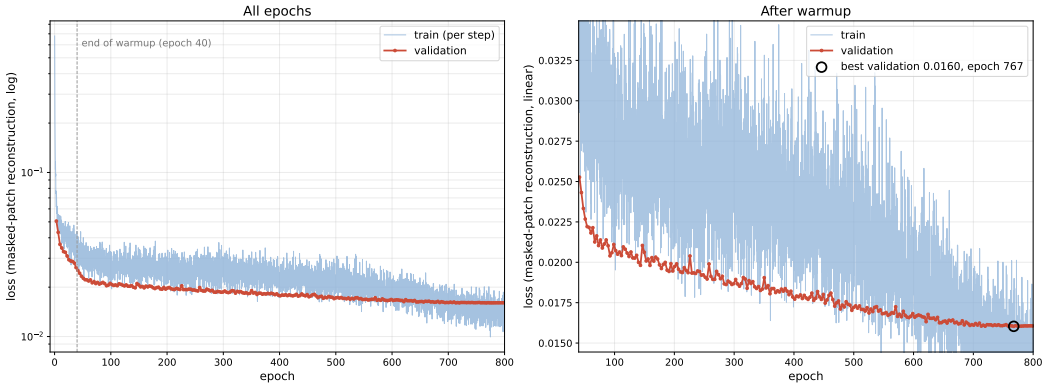

Figure 9: Pretraining loss of the model with 75% masking, on training batches and on held-out pretraining spectra. Left: all epochs, logarithmic scale. Right: after the 40 warmup epochs, linear scale, with the epoch of the lowest validation loss marked.

Table 5: Pretraining settings.

Setting	Value
Patch length	48 points
Encoder	12 blocks, width 768, 12 heads, MLP ratio 4
Decoder	8 blocks, width 512, 16 heads
Positional code	Fourier features of absolute wavenumber
Masking ratio	75%
Loss	mean squared error on normalised hidden patches
Optimiser	AdamW, $\beta = (0.9, 0.95)$, weight decay 0.05
Learning rate	$1.5 \times 10^{-4} \times 4096/256$
Schedule	40 warmup epochs, then cosine decay
Batch size	4096
Epochs	800
Precision	bfloat16

B MASKING RATIO

He et al. (2022) chose a masking ratio of 75% for images. NIR spectra are smooth, and a hidden patch can partly be interpolated from its neighbours. A different ratio could therefore suit them better. To choose the ratios for full pretraining, we pretrain five encoders with masking ratios of 15%, 40%, 60%, 75% and 90% for 100 epochs each, with the recipe of Appendix A compressed to this length. We then fit a ridge model on the frozen embeddings to predict soil organic carbon on LUCAS with N labelled samples, with the protocol of Section 5, averaged over 10 random draws.

Table 6 reports the result. With 100 labels, the ratios of 60% and below stay at or below $R^2 = 0.19$. The ratios of 75% and 90% reach 0.43 and 0.50 and lead at every size. The short schedule lowers every value: with 100 labels, the model with 75% masking reaches 0.43 after 100 epochs and 0.73 after the full 800 epochs. We therefore pretrained both for 800 epochs. They end within 0.04 of each other at every size.

Three considerations lead us to keep 75% as the main model. First, it is the ratio of He et al. (2022), and keeping it leaves their recipe unchanged. Second, after 800 epochs it is ahead of 90% at every label count, by 0.01 to 0.04 in R^2, although 90% led after 100 epochs. The short schedule therefore does not predict the final ranking. Third, 90% masking produces more matching first-layer filters, 65% against 21%, almost all of them second derivatives (Appendix C). A main model with fewer matches keeps the first-layer finding conservative. We report both models throughout the paper.

Table 6: Soil organic carbon R^2 on LUCAS, ridge model on frozen embeddings, N labelled samples.

Masking	Epochs	$N = 50$	$N = 100$	$N = 200$	$N = 400$
15%	100	0.05	0.19	0.35	0.42
40%	100	−0.25	0.13	0.13	0.29
60%	100	−0.07	0.13	0.33	0.50
75%	100	0.41	0.43	0.51	0.68
90%	100	0.43	0.50	0.65	0.72
75%	800	0.69	0.73	0.78	0.83
90%	800	0.65	0.69	0.77	0.82

C FIRST-LAYER FILTERS IN DETAIL

This appendix explains how we decide that a learned filter is a chemometric operator, and why chance does not explain the matches. The first layer projects each patch of 48 points with a weight matrix of 768 rows. We read each row as a filter over the 48 points and drop the weight on the coverage fraction. A Savitzky-Golay filter fits a polynomial to a window of points and returns its value or one of its derivatives at the centre (Savitzky & Golay, 1964). Our template bank contains these filters for a window of 48 points, polynomial orders 2 to 5 and derivative orders 0 to 3. We centre the derivative templates to zero mean. For each learned filter, we take the largest absolute correlation with any template in the bank. The absolute value makes the match independent of the sign of the filter.

A smooth filter correlates with some smooth template by chance. The untrained filters are random noise and therefore give a weak control. The stricter control keeps the magnitude of the Fourier transform of each learned filter and draws its phases at random, for three seeds. These control filters have exactly the smoothness of the learned ones but a random shape. Table 7 varies the threshold. At 0.8, the control already matches 38% of the time, and the threshold is too lax to separate shape from smoothness. At 0.9 and 0.95, the learned filters match two to twenty times more often.

Table 7: Share of first-layer filters that match a Savitzky-Golay template, by threshold on the absolute correlation. Same-smoothness values are the mean of three seeds.

Threshold	75%	Same smoothness	90%	Same smoothness	Untrained
0.8	81%	38%	80%	39%	0%
0.9	21%	10%	65%	9%	0%
0.95	11%	3%	58%	3%	0%

Table 8 counts the derivative order of the matching filters at a threshold of 0.9. The two masking ratios divide their matching filters differently over derivative orders. With 75% masking, first and third derivatives are equally common and second derivatives less so. With 90% masking, 463 of the 501 matching filters are second derivatives. A second derivative removes both an additive offset and a linear slope, the two effects of light scattering described in Section 1 (Rinnan et al., 2009). The model with 90% masking therefore concentrates on the one operator that removes both scatter effects at once, while the model with 75% masking keeps a broader set of derivatives.

The first layer as a whole is low-dimensional. In a singular value decomposition of the first-layer weights, four directions carry 99.9% of the energy, against 12% in an untrained encoder. Polynomials of degree four or lower hold 99.5% and 99.8% of the weight energy in the models with 75% and 90% masking. Three of the four directions coincide with leading principal components of the input patches, which are smooth. The learned directions are, however, more polynomial than these components, with 95% and 99.6% of their energy in cubic polynomials, against 78% for the principal components. The first layer is therefore low-rank and polynomial, like a bank of Savitzky-Golay filters.

Table 8: Derivative order of the first-layer filters that match a template with absolute correlation above 0.9, out of 768 filters.

Masking	First	Second	Third	Total
75%	64	33	63	160
90%	29	463	9	501

D CHEMISTRY BY LAYER AND BY EPOCH

To see where in the network soil chemistry becomes available, we read out the representation after every block. For each layer, we average the output tokens over the patches that the instrument covers. Layer 0 is the patch embedding with the positional code added, which is the input to the first block. We then fit a ridge model, with its regularisation chosen by cross-validation, to predict soil organic carbon on LUCAS from 100 labelled samples, averaged over 10 random draws, with the protocol of Section 5 and the same draws for every encoder and layer.

Table 9 gives the values behind Figure 3, left, for each encoder. The untrained encoders stay between $R^2 = 0.51$ and 0.62 after the patch embedding and do not improve with depth. The pretrained encoders start lower and rise through the blocks. The rise is not monotonic in the first four blocks, and the pretrained encoders overtake the untrained ones from block 5.

Table 9: Soil organic carbon R^2 from each encoder layer, ridge model with 100 labelled LUCAS samples, mean of 10 draws. Layer 0 is the patch embedding; layer 12 is taken before the final normalisation of the encoder.

Encoder	0	1	2	3	4	5	6	7	8	9	10	11	12
75%	0.42	0.45	0.55	0.44	0.58	0.63	0.68	0.71	0.70	0.73	0.72	0.73	0.72
90%	0.31	0.44	0.61	0.62	0.54	0.68	0.70	0.71	0.71	0.68	0.70	0.70	0.69
Untrained 1	0.46	0.58	0.62	0.61	0.61	0.57	0.62	0.62	0.61	0.59	0.59	0.59	0.59
Untrained 2	0.45	0.59	0.54	0.51	0.54	0.57	0.56	0.53	0.56	0.53	0.51	0.56	0.53
Untrained 3	0.46	0.58	0.60	0.60	0.58	0.60	0.60	0.61	0.59	0.59	0.58	0.56	0.57

Table 10 repeats the measurement on the checkpoints saved every 50 epochs, at the output of the encoder and with 50, 100 and 400 labelled samples, and gives the values behind Figure 3, middle and right. From epoch 50 on, both models stay close to their final accuracy. With 50 labels, the values vary by up to 0.09 between neighbouring checkpoints, which reflects the variance of small calibration sets. With 400 labels, they vary by at most 0.04. The last columns give the median correlation between each matching first-layer filter and the template it matches at epoch 800. It rises until about epoch 300 and then stays constant. What the encoder knows about soil is thus in place long before pretraining ends, while the first layer takes about 300 epochs to reach its final shape.

Table 10: Soil organic carbon R^2 at the output of the encoder for N labelled samples, and median correlation of the matching first-layer filters with their final template, at every saved checkpoint.

	75% masking				90% masking			
Epoch	$N = 50$	$N = 100$	$N = 400$	Filters	$N = 50$	$N = 100$	$N = 400$	Filters
50	0.66	0.69	0.79	0.731	0.64	0.65	0.80	0.531
100	0.66	0.71	0.83	0.824	0.68	0.70	0.81	0.781
150	0.67	0.71	0.82	0.872	0.65	0.69	0.80	0.928
200	0.64	0.71	0.80	0.891	0.71	0.75	0.83	0.968
250	0.73	0.74	0.84	0.921	0.66	0.73	0.82	0.982
300	0.71	0.75	0.83	0.942	0.65	0.71	0.81	0.988
350	0.70	0.72	0.82	0.943	0.64	0.66	0.81	0.991
400	0.68	0.72	0.83	0.950	0.69	0.71	0.82	0.993
450	0.64	0.72	0.82	0.952	0.67	0.70	0.82	0.994
500	0.68	0.72	0.82	0.954	0.65	0.68	0.81	0.995
550	0.65	0.71	0.82	0.954	0.66	0.69	0.81	0.995
600	0.71	0.74	0.84	0.952	0.63	0.69	0.81	0.995
650	0.69	0.72	0.83	0.953	0.65	0.72	0.83	0.995
700	0.69	0.74	0.83	0.952	0.66	0.71	0.82	0.995
750	0.69	0.73	0.83	0.952	0.67	0.69	0.82	0.995
800	0.69	0.73	0.83	0.952	0.65	0.69	0.82	0.995

E RESPONSE TO PHYSICAL CHANGES

A spectrum can change without any change in the chemistry of the sample. The sample can be scanned again, its temperature can shift the absorption bands (Wülfert et al., 1998), and light scattering can change its offset and slope. We apply such changes to 960 held-out food and feed spectra, 120 from each of eight materials, and to 960 LUCAS spectra. The band shift is a first-order shift of the spectrum along the wavenumber axis by up to 8 cm^{-1}. The baseline adds an offset and a slope of up to 5% of each spectrum's standard deviation. Extended multiplicative scatter correction (Martens & Stark, 1991) removes a multiplicative factor and a quadratic baseline relative to the mean spectrum. The repeated scan replaces a spectrum with a second replicate of the same measurement, which exists for the food and feed set only. White noise adds Gaussian noise with a standard deviation of 1% of each spectrum's standard deviation.

For each spectrum, we compute the cosine distance between the embeddings before and after the change. Embeddings of different encoders differ in spread: those of an untrained encoder barely change for any input. We therefore divide the mean distance by the mean cosine distance between different samples of the same set. Table 11 reports the result at the output of the encoder. Scale and offset change no embedding by more than 10^{-6}, for every encoder and layer, because the input normalisation removes them before the first layer.

Table 11: Embedding change relative to the mean distance between samples, in percent, at the output of the encoder. Untrained values are the mean of three encoders.

	Food and feed			LUCAS		
Change	75%	90%	Untrained	75%	90%	Untrained
Repeated scan	0.71	0.99	0.41	–	–	–
Band shift	2.0	3.3	0.12	0.67	0.52	0.06
White noise, 1%	5.8	5.6	0.11	0.15	0.13	0.04
Extended MSC	40	41	43	49	56	56
Baseline	0.04	0.05	0.02	0.02	0.02	0.02

Three results hold on both sets. A repeated scan moves the pretrained embedding by less than 1% of the distance between samples. A band shift moves it by 9 to 30 times more than it moves an untrained embedding, relative to spread, in line with derivative filters that respond to band position. Extended multiplicative scatter correction moves every encoder, trained or not, by about half the distance between samples. The encoder therefore keeps the information that this correction removes. White noise moves the pretrained embedding more on the food and feed set than on LUCAS, whose

laboratory spectra are smoother. The encoder is therefore stable where the sample is unchanged, responds to the band shifts that its derivative filters track, and does not remove scatter by itself.

F STRUCTURE OF THE EMBEDDING

A useful representation should place chemically similar samples close together, and this appendix measures how the embedding does so. We measure its structure in the full 768-dimensional space, not in a projection. All metrics use L2-normalised embeddings and cosine distance. We compare five representations: the two pretrained encoders, the three untrained encoders, the normalised spectrum, and the normalised spectrum reduced by PCA to the components that explain 99.9% of its variance.

Figure 4 uses 2,676 held-out measurements of eight materials. The broader food and feed set has 6,761 held-out measurements of 40 materials, with at most 200 measurements per material. We group the materials into 12 product families by generic product type, such as maize grain or animal protein meal. The silhouette compares, for each measurement, the mean distance to its own material with the mean distance to the nearest other material, and ranges from -1 to 1. The material accuracy is the share of measurements whose material a 10-nearest-neighbour classifier predicts correctly, with 5-fold cross-validation. Retrieval takes one scan of each measurement and asks whether its nearest neighbour is the second scan of the same measurement. On the eight materials of Figure 4, every representation reaches a material accuracy above 0.97. The broader set of 40 materials separates the representations, as Table 2 shows.

The soil set is the full LUCAS library of 19,036 samples. For each property, a 10-nearest-neighbour regression with 5-fold cross-validation measures how well the neighbours of a soil predict its value. In a check with the LUCAS spectra converted to reflectance and restricted to the ProxiScout range, the nearest-neighbour R^2 for soil organic carbon rises for the encoders and for the spectra, from 0.85 to 0.89 with 75% masking and to 0.87 for the spectra.

Table 12 compares distances between the centres of materials, of product families and of soils. Raw distances differ in scale between representations, and we therefore report ratios and shares. For 34 materials in families with more than one member, the nearest other material lies in the same family in 88% of cases in every representation except the untrained encoders, against 12% by chance ($p < 0.001$, permutation test). When the encoder confuses two materials, the two belong to the same family in 93% of cases. Its family ratio is lower than that of the spectra: the encoder spreads materials within a family further apart. Soils lie farther from the food and feed materials than those lie from each other in every representation, but the encoder places them much closer to food and feed than the spectra do. The encoder therefore separates materials better, preserves the product families of the input, and places soils closer to food and feed.

Table 12: Distances between cluster centres. Family ratio: mean distance between family centres over mean distance between material centres within a family. Soil ratio: mean distance from a soil to the nearest food or feed centre over the mean distance between neighbouring food or feed centres. Soils beyond all food distances: share of soils farther from every food or feed centre than the largest distance between two food or feed centres.

Metric	75%	90%	Untrained	Spectra	Spectra, PCA
Nearest material in same family	0.88	0.88	0.85	0.88	0.88
Misclassified within true family	0.93	0.94	0.88	0.91	0.91
Family ratio	3.6	3.3	7.5	5.9	3.1
Soil ratio	13	15	2541	278	576
Soils beyond all food distances	0%	0%	100%	100%	100%

G SOIL EVALUATION WITH ALL MODELS

Figure 10 describes LUCAS. Land cover and soil type split the same 19,036 samples in two ways. Organic soils carry far more carbon and nitrogen and have a lower pH than mineral soils, which is why we model the subsets separately.

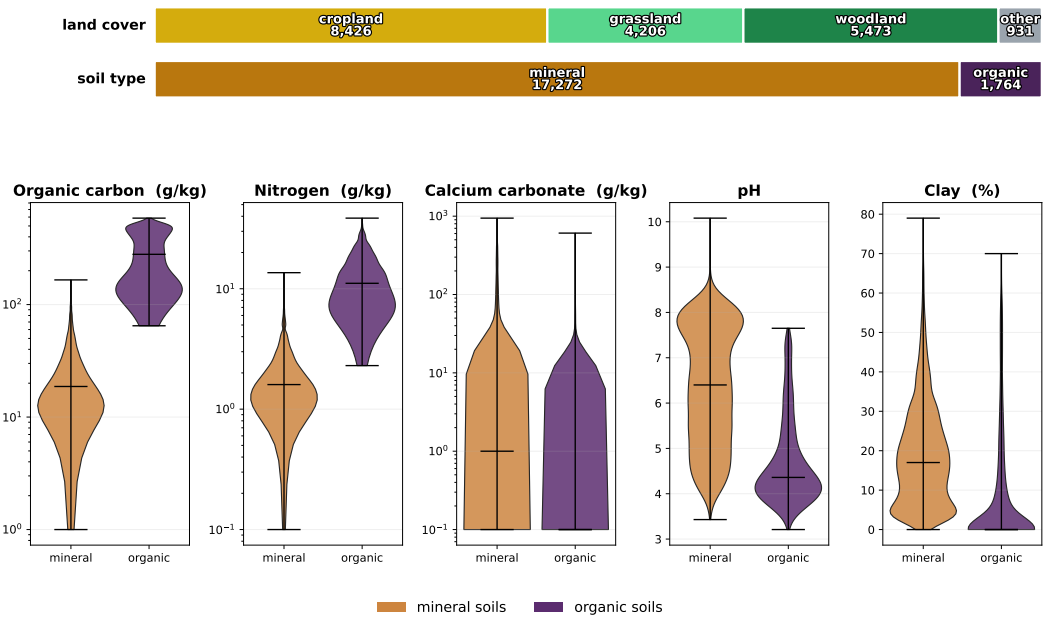

Figure 10: The LUCAS 2009 topsoil library. Top: number of samples by land cover and by soil type; other land covers are shrubland, bare land, wetland, water and artificial land. Bottom: distributions of the five properties in mineral and organic soils.

The full-data comparison of Section 5 follows the design of Stevens et al. (2013). We model each subset separately: mineral and organic soils, and cropland, grassland and woodland. Within each subset, the Kennard-Stone algorithm (Kennard & Stone, 1969) selects two thirds of the samples for calibration so that they cover the spectral space evenly, and the remaining third is the test set. The chemometric models receive the preprocessing of Stevens et al. (2013): Savitzky-Golay smoothing with a first derivative over 101 points and a third-order polynomial, then SNV, keeping every twentieth band. On these spectra we fit PLS with its number of components chosen by cross-validation, ridge regression with its regularisation chosen by cross-validation, Cubist, memory-based learning and the 1D CNN of Cui & Fearn (2018). On the frozen embedding we fit ridge, PLS, Cubist, memory-based learning and a 5-nearest-neighbour regression.

The split of Stevens et al. (2013) is not published, and ours differs in the following respects. We take mineral and organic soils from the classification distributed with LUCAS, which gives 17,272 and 1,764 samples. The land-cover subsets include organic soils. We run Kennard-Stone on the first ten principal components of the preprocessed spectra. For mineral soils and cropland, we test on 2,000 random samples of the held-out third, the same for every model. The chemometric models use the spectrum from 500 to 2500 nm, while the encoder reads 1000 to 2500 nm. All models share the same calibration and test samples.

Table 13 reports soil organic carbon R^2 for every model. On the embedding, ridge regression is the best or tied best model on four of five subsets, and Cubist on cropland. On the spectra, Cubist and memory-based learning are the strongest models. The main text therefore compares a single ridge model on the embedding with the best of five chemometric models per subset.

Table 13: Soil organic carbon R^2 with all calibration samples, on the frozen embedding of the model with 75% masking and on the preprocessed spectra.

Input	Model	Mineral	Organic	Cropland	Grassland	Woodland
Embedding	Ridge	0.68	0.83	0.85	0.85	0.95
	PLS	0.57	0.82	0.80	0.80	0.93
	Cubist	0.67	0.82	0.87	0.82	0.93
	MBL	0.64	0.79	0.84	0.85	0.93
	kNN	0.49	0.59	0.76	0.74	0.88
Spectra	PLS	0.53	0.79	0.68	0.71	0.89
	Ridge	0.64	0.80	0.79	0.77	0.89
	Cubist	0.73	0.80	0.83	0.82	0.94
	MBL	0.59	0.82	0.84	0.80	0.94
	1D CNN	0.42	0.79	0.76	0.72	0.92

The same design applies to four further LUCAS properties: nitrogen, clay, pH and calcium carbonate. Table 14 gives the values behind Table 3 for every property and subset. For organic carbon, nitrogen and clay, the embedding matches or exceeds the best chemometric model on four of five subsets, with the largest gain for clay in organic soils, at 0.87 against 0.79. For pH and calcium carbonate, the chemometric models are ahead on every subset.

Table 14: LUCAS soil properties with all calibration samples: R^2 of a ridge model on the frozen embedding of the model with 75% masking, and of the best of five chemometric models on the spectra (in parentheses), per subset of the design of Stevens et al. (2013).

Property	Mineral	Organic	Cropland	Grassland	Woodland
Organic carbon	0.68 (0.73)	0.83 (0.82)	0.85 (0.84)	0.85 (0.82)	0.95 (0.94)
Nitrogen	0.65 (0.70)	0.81 (0.81)	0.74 (0.72)	0.80 (0.74)	0.92 (0.92)
Clay	0.82 (0.82)	0.87 (0.79)	0.82 (0.82)	0.73 (0.70)	0.77 (0.72)
pH	0.85 (0.91)	0.82 (0.88)	0.81 (0.88)	0.62 (0.80)	0.73 (0.84)
Calcium carbonate	0.94 (0.96)	0.55 (0.67)	0.95 (0.96)	0.82 (0.91)	0.62 (0.78)

To test whether the embedding carries information that a linear chemometric model does not, we hold out a random 30% of LUCAS and fit PLS with 15 components to the standardised spectra the encoder reads. We then regress the standardised embedding on the 15 PLS scores and keep the residual, which is the part of the embedding that PLS does not explain. A ridge model on this residual still predicts every property on the held-out soils (Table 15). The embedding therefore contains the PLS subspace and additional information.

Table 15: Soil properties on a held-out 30% of LUCAS predicted from the part of the embedding of the model with 75% masking that 15 PLS components of the spectra do not explain.

	Organic carbon	Nitrogen	Calcium carbonate	pH	Clay
Residual R^2	0.05	0.11	0.08	0.14	0.11

H SOIL WITH FEW LABELS

A calibration for a new material often starts with a few dozen reference values. To measure how the representation performs in this regime, we hold out a random quarter of LUCAS, grouped by sample and capped at 5,000 samples, as the test set. From the remaining samples, we draw N labelled samples at random, for N from 25 to 400, and repeat each draw 10 times. All models share the same draws and the same test set. Ridge regression and PLS choose their regularisation and number of components by cross-validation within each draw.

Table 16 gives every model. On the embedding, we fit ridge, PLS, Cubist and a small neural network with two hidden layers of 64 and 16 units. Ridge is the best of them from 50 labels on, and Cubist and the network overfit with few labels. On the spectra, we fit ridge and PLS on two ranges, the full spectrum and the range the encoder reads, with three preprocessings. On the range the encoder

reads with the first derivative, we also fit Cubist, memory-based learning and the 1D CNN of Cui & Fearn (2018). Cubist uses 10 committees on both inputs. Memory-based learning needs at least 60 samples to form local neighbourhoods and starts at 100 labels. The standard chemometric workflow selects the preprocessing by cross-validation; it selected the first derivative on the full spectrum at every size. SNV alone on the full spectrum is unstable with few labels. Cubist on the spectra is the best model at every size. From 50 labels on, it leads the ridge model on the embedding by 0.02 to 0.05, and with 25 labels by 0.29. The untrained encoders were evaluated from 50 labels on.

Table 16: Soil organic carbon R^2 against the number of labelled samples, mean of 10 draws. Spectral models use the full spectrum (400 to 2500 nm) or the range the encoder reads (1000 to 2500 nm), after SNV alone (SNV), or followed by a first (D1) or second (D2) Savitzky-Golay derivative. CV: preprocessing selected by cross-validation.

Input	Model	$N = 25$	$N = 50$	$N = 100$	$N = 200$	$N = 400$
Embedding, 75%	Ridge	0.25	0.69	0.73	0.78	0.83
	PLS	0.33	0.66	0.71	0.76	0.81
	Cubist	−0.04	0.49	0.55	0.68	0.77
	MLP	0.24	0.57	0.53	0.73	0.77
Untrained (3)	Ridge	–	0.40–0.51	0.53–0.59	0.62–0.67	0.59–0.68
Full, CV	PLS	0.32	0.48	0.51	0.53	0.71
	Ridge	0.18	0.54	0.59	0.62	0.72
Full, SNV	PLS	0.37	0.30	0.58	0.26	0.41
	Ridge	0.33	−0.35	0.58	0.15	0.63
Full, D2	PLS	0.45	0.59	0.63	0.65	0.71
	Ridge	0.42	0.57	0.62	0.64	0.73
Encoder range, SNV	PLS	0.38	0.55	0.66	0.62	0.72
	Ridge	0.44	0.57	0.68	0.68	0.72
Encoder range, D1	PLS	0.48	0.53	0.65	0.68	0.75
	Ridge	0.45	0.53	0.67	0.73	0.77
	Cubist	0.54	0.71	0.78	0.81	0.85
	MBL	–	–	0.73	0.75	0.81
	1D CNN	0.32	0.55	0.56	0.68	0.78
Encoder range, D2	PLS	0.44	0.57	0.63	0.68	0.72
	Ridge	0.47	0.60	0.66	0.71	0.74

The derivatives in Table 16 use a Savitzky-Golay window of 11 points, and every wavelength is scaled to unit variance before fitting. To make sure the baselines are not handicapped by these choices, we repeat the models on the range the encoder reads with windows of 21, 51 and 101 points and with mean centring instead of scaling. Table 17 gives the result. Wider windows help the first derivative with 50 labels, where ridge reaches 0.61, and do not change the other sizes. Without scaling, ridge on second derivatives fails. None of these variants reaches Cubist at any size. The main text therefore reports Cubist as the best spectral model.

Table 17: Soil organic carbon R^2 of spectral models on the range the encoder reads, with mean centring and Savitzky-Golay windows of w points, mean of the same 10 draws as Table 16.

Derivative	Model	w	$N = 25$	$N = 50$	$N = 100$	$N = 200$	$N = 400$
First	PLS	21	0.39	0.50	0.66	0.67	0.73
	Ridge	21	0.42	0.61	0.64	0.70	0.73
	PLS	51	0.39	0.49	0.65	0.69	0.73
	Ridge	51	0.44	0.61	0.64	0.69	0.72
	PLS	101	0.36	0.57	0.66	0.64	0.72
	Ridge	101	0.44	0.59	0.62	0.68	0.71
Second	PLS	21	0.18	0.41	0.58	0.67	0.71
	Ridge	21	−0.05	0.04	0.06	0.14	0.19
	PLS	51	0.26	0.42	0.61	0.68	0.72
	Ridge	51	−0.06	0.02	0.04	0.10	0.15
	PLS	101	0.33	0.52	0.62	0.67	0.72
	Ridge	101	−0.07	0.00	0.01	0.05	0.08

I SUPERVISED MODELS

The Discussion compares the masked autoencoder with a transformer and a CNN trained with labels. We split LUCAS by sample into 80% for training and 20% for testing, with seed 0, and hold out 10% of the training samples for early stopping. The target is soil organic carbon after $\log(1 + x)$ and standardisation. The supervised transformer has the architecture and input path of the masked autoencoder. We initialise it from an untrained encoder, add a linear head to the mean of the output tokens, and train it end to end with AdamW, a learning rate of 2×10^{-4}, weight decay 0.05, warmup and cosine decay, and a batch size of 256. Its best checkpoint is at epoch 88, and training stops at epoch 114. The supervised CNN has a first convolutional layer with 32 filters of 48 points, two further convolutional layers, global average pooling and a small fully connected head. We train it with three seeds. For the comparison in Table 4, the masked autoencoder and the untrained encoders are evaluated on the same split with a ridge model on all training samples. CKA and the band shift use 960 test spectra.