Task-Aware Self-Supervised Pretraining for Few-Shot Soil NIR Regression

Iván Garzón[1,3][0000−0001−6043−1425]⋆,
Guillermo Gomez-Trenado[2,3][0000−0003−3366−6047], and
Sergio Damas[1,3][0000−0002−8377−8349]

[1] Department of Software Engineering, University of Granada, Granada, 18071, Spain
[2] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, 18071, Spain
[3] Andalusian Research Institute in Data Science and Computational Intelligence, University of Granada, Granada, 18016, Spain

Abstract. Accurate soil-property estimation from Near-Infrared (NIR) spectra is limited by the cost and destructiveness of laboratory analyses, which restricts the number of labelled samples available for each target property or deployment scenario. This work studies task-aware self-supervised pretraining for few-shot soil-property regression. One-dimensional NIR spectra are encoded as two-dimensional Gramian Angular Difference Fields (GADF), enabling the use of an image-based I-JEPA objective with a Vision Transformer backbone. Since GADF images are antisymmetric, we adapt the masking strategy so that target regions and their transposed counterparts are jointly removed from the context, reducing shortcut solutions during masked latent prediction. To align representation learning with the downstream regression objective, we train I-JEPA jointly with a Prototypical Network loss computed on labelled few-shot episodes from related soil-property tasks. On held-out tasks, the proposed joint model reaches a cross-validated test R^2 of 0.467, improving over both pure I-JEPA pretraining ($R^2 = 0.369$) and the same ViT ProtoNet architecture trained from scratch ($R^2 = 0.411$). Although it remains below a convolutional ResNet18 ProtoNet reference ($R^2 = 0.492$), the results show that task-aware pretraining improves the transformer representation and suggest that the remaining gap is mainly related to backbone–data mismatch.

Keywords: Few-shot regression · Self-supervised learning · I-JEPA · Prototypical Networks · Near-infrared spectroscopy · Soil analysis · Gramian Angular Difference Fields

1 Introduction

Accurate soil characterization is essential for precision agriculture, environmental monitoring, and sustainable land management. However, conventional laboratory analyses of soil properties are often costly, time-consuming, and destructive,

⋆ Corresponding author: ivangarzon@ugr.es

which limits the amount of labelled data that can be collected for each target property and deployment scenario. Near-Infrared Spectroscopy (NIR) offers a rapid and non-destructive alternative, as it provides high-dimensional spectral signatures that are informative about the physicochemical composition of soil samples [12]. The challenge is that translating these spectra into accurate property estimates remains difficult when only a small number of labelled examples are available for a new soil property, region, or acquisition condition.

This setting naturally motivates few-shot learning. Instead of training a model independently for each new regression problem, few-shot methods aim to exploit knowledge acquired from related tasks and adapt to a new one using only a limited support set. This is particularly relevant in spectroscopy, where labels are expensive but collections of spectra from related problems may be available. Recent work by Garzón et al. showed that meta-learning methods are effective for few-shot spectroscopic regression, with Prototypical Networks [14] emerging as a particularly strong metric-based approach [7]. This motivates our use of Prototypical Networks as the task-aware adaptation mechanism in the present work.

Beyond the choice of the adaptation mechanism, recent few-shot learning studies have increasingly emphasized the role of the representation itself. Under controlled comparisons, several works have shown that the quality of the embedding and the backbone can explain a large part of the performance previously attributed to more complex meta-learning algorithms [3, 11, 15]. This representation-first view is particularly relevant for spectroscopy, where raw or weakly annotated spectra are often more abundant than fully labelled examples for each downstream property.

Self-supervised learning provides a natural way to exploit such data before few-shot adaptation, and has recently been explored for small-sample NIR spectral analysis [19]. Joint-embedding predictive architectures, such as I-JEPA, learn by predicting latent representations of masked input regions rather than reconstructing pixels directly [2]. However, this objective is not necessarily aligned with downstream regression: a representation that is predictive under masking may not organize spectra according to the soil properties of interest. This motivates adding a task-aware signal during representation learning, rather than relying on a purely self-supervised objective.

In this work, we combine I-JEPA with Prototypical Networks for few-shot regression on soil NIR spectra. The I-JEPA objective exploits the available spectra through masked latent prediction, whereas the prototypical loss uses labelled episodes from related training tasks to organise the embedding space according to property-relevant variation. The resulting training scheme directly tests whether adding a task-aware episodic signal during self-supervised pretraining improves downstream few-shot regression.

To apply image-based I-JEPA to spectra, each one-dimensional spectrum is encoded as a two-dimensional Gramian Angular Difference Field (GADF) [17], which represents pairwise relations between wavelength positions as an image. This makes it possible to use a patch-based Vision Transformer backbone, the

natural architecture for I-JEPA, and to compare the proposed pretrained model against a scratch Vision Transformer (ViT) ProtoNet control with the same input representation, backbone, episodic protocol, and evaluation split. Since GADF images have a deterministic antisymmetric structure, we also adapt the masking strategy to prevent masked regions from being inferred from their visible transposed counterparts.

Empirically, the joint model reaches a cross-validated test R^2 of 0.467, improving over both pure I-JEPA pretraining ($R^2 = 0.369$) and the scratch ViT ProtoNet control ($R^2 = 0.411$). The latter comparison isolates the effect of joint pretraining under the same architecture and evaluation protocol. The model remains slightly below the ResNet18 + ProtoNet reference [7], suggesting that convolutional priors and model scale remain relevant under the current pretraining-data regime.

The main contributions of this work are as follows:

- We propose a joint I-JEPA and Prototypical Network pipeline for few-shot regression on Gramian Angular Difference Fields (GADF) representations of soil NIR spectra.
- We introduce a task-aware pretraining objective in which a prototypical regression loss is optimised jointly with masked latent prediction, injecting episodic supervision during representation learning.
- We adapt the I-JEPA masking strategy to the antisymmetric structure of GADF inputs, preventing masked regions from being inferred from their visible transposed counterparts.
- We show, under a controlled same-backbone comparison, that joint pretraining improves a ViT ProtoNet model from $R^2 = 0.411$ to $R^2 = 0.467$, while analysing the remaining gap with a convolutional ResNet18 + ProtoNet reference in terms of convolutional priors, model scale, and data regime.

2 Related Work

Few-shot learning was originally developed around mechanisms for rapid adaptation from a small support set. Metric-based methods, such as Prototypical Networks, perform prediction through distances in a learned embedding space [14], whereas optimisation-based methods, such as MAML, learn an initialisation that can be adapted with a few gradient steps [6]. More recent controlled studies have shown, however, that much of the performance previously attributed to complex meta-learning algorithms can be explained by the quality of the representation, backbone, and evaluation protocol [3]. This has led to a representation-first view, where strong embeddings combined with simple adaptation rules are often competitive with more sophisticated episodic methods [15, 11]. For spectroscopic regression, this view supports the use of simple adaptation mechanisms on top of transferable embeddings, rather than relying exclusively on complex meta-learning procedures. Recent work on few-shot spectroscopic regression showed that Prototypical Networks provide a strong and simple metric-based baseline [7], motivating their use here as the task-aware adaptation mechanism.

Self-supervised learning provides a natural way to learn representations from unlabelled or weakly labelled data before few-shot adaptation. I-JEPA learns by predicting target representations in latent space rather than reconstructing pixels, which makes it suitable for representation learning from image-like inputs [2]. More broadly, domain- and task-adaptive pretraining have been shown to improve downstream performance when labelled data are scarce [8]. Similar motivations have recently appeared in spectral modelling, including masked autoencoding for NIR soil-property prediction and self-supervised learning for small-sample NIR classification [16, 19]. These works mainly consider self-supervision as a representation-learning stage optimised before the downstream predictor. Less attention has been paid to settings where masked latent prediction and episodic few-shot regression supervision are optimised jointly during pretraining.

To apply image-based self-supervised learning to spectra, one-dimensional signals must first be mapped to two-dimensional inputs. Gramian Angular Fields were introduced to encode sequential signals as images while preserving pairwise relations between positions [17]. This family of encodings has also been used in spectral analysis; for example, recent work combines GADF representations of Vis-NIR spectra with a lightweight vision transformer for regression of apple soluble solids content [10]. In our setting, GADF is not merely an input conversion step. Its antisymmetric structure affects the pretraining task itself, since a masked region can otherwise be partly inferred from its visible transpose. This motivates the symmetric masking strategy described in Section 3.

Vision Transformers have become a prominent backbone in modern vision and self-supervised pretraining pipelines. The original ViT work showed that a pure transformer over image patches can compete with or surpass convolutional networks when trained at sufficient scale [4], and I-JEPA itself is formulated as a patch-based visual self-supervised architecture [2]. This makes a ViT backbone a natural choice when spectra are converted into image-like inputs. At the same time, the literature also shows that this choice should be evaluated case by case. Compared with CNNs, ViTs encode weaker image-specific priors such as locality, translation equivariance, and hierarchical feature extraction, which can reduce data efficiency in small or specialised regimes [4, 5]. Recent work has therefore revisited the role of convolutional structure, showing that local inductive bias can improve ViTs on small datasets [1] and that ConvNets can match Vision Transformers under controlled large-scale pretraining conditions [13]. Existing work therefore leaves a relevant architectural tension: ViTs are well suited to patch-based self-supervised learning, whereas convolutional priors may remain advantageous when data are limited or highly structured.

3 Method

3.1 Spectral Encoding via GADF

We cast few-shot NIR regression as an image problem by mapping each one-dimensional reflectance spectrum to a single-channel GADF image [17] (see Fig.

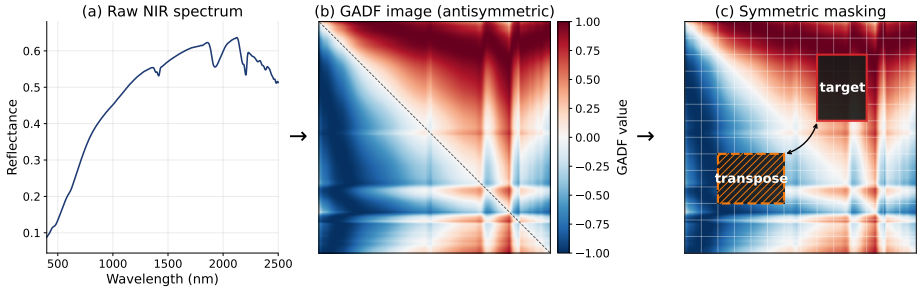


Fig. 1. From a 1D NIR spectrum to a masked 2D GADF representation. (a): a raw soil spectrum. (b): its GADF image (Eq. 2), antisymmetric about the diagonal. (c): a sampled target configuration under symmetric masking, where a target block and its transpose are jointly hidden from the context encoder.

1). Each spectrum contains $L = 1051$ wavelengths in the 400–2499 nm range. Depending on the experiment, a Savitzky–Golay filter ($w = 15$, $p = 2$) and its first derivative are applied before reducing the spectrum to 224 points with Piecewise Aggregate Approximation. The resulting vector is scaled per sample to $[-1, 1]$, mapped to angular coordinates,

$$\phi_i = \arccos(\tilde{x}_i), \qquad \tilde{x}_i \in [-1, 1], \tag{1}$$

and the GADF is formed as

$$G_{ij} = \sin(\phi_i - \phi_j), \tag{2}$$

yielding a 224×224 matrix, followed by image-level tensor normalisation before the backbone.

The encoding in Eq. (2) is antisymmetric: since the sine is odd,

$$G_{ji} = \sin(\phi_j - \phi_i) = -\sin(\phi_i - \phi_j) = -G_{ij}, \qquad G_{ii} = 0. \tag{3}$$

The matrix is therefore fully determined by either triangle: the entry (j, i) is the negation of (i, j). As discussed in Section 3.2, this redundancy directly motivates the symmetric masking scheme, since a masked region can otherwise be partly recovered from its visible mirror.

3.2 I-JEPA Backbone with Symmetric Masks

We pretrain the backbone with Image-based Joint-Embedding Predictive Architecture (I-JEPA) [2], which predicts in representation space rather than reconstructing pixels. The model comprises three components: a context encoder f_θ (a ViT operating on 16×16 patches, i.e., 196 tokens for a 224×224 input), a narrow predictor g_ϕ, and a target encoder $f_{\bar{\theta}}$ whose weights are an exponential moving average (EMA) of the context encoder.

Given a GADF image, the target encoder processes the full input to produce patch-level representations; M target blocks are sampled, and the targets are the (stop-gradient) teacher representations at those positions. The context encoder receives only a context block from which the target regions have been removed, and the predictor maps the context representation, together with positional mask tokens for each target block, to predicted target representations. The objective is the average ℓ_2 distance between predicted and target representations over the M blocks,

$$\mathcal{L}_{\mathrm{JEPA}} = \frac{1}{M} \sum_{m=1}^{M} \left\| g_\phi\big(f_\theta(\mathbf{x}_{\mathrm{ctx}}), \mathbf{p}_m\big) - \mathrm{sg}\big[f_{\bar{\theta}}(\mathbf{x})_m\big] \right\|_2^2, \tag{4}$$

where $\mathbf{p}_m$ are the positional tokens of block m and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The target encoder is not trained by backpropagation but updated as $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$. This stop-gradient teacher–student design provides stable non-trivial targets and helps prevent representational collapse, where different inputs are mapped to nearly identical embeddings.

Symmetric masking. Standard I-JEPA samples target and context blocks independently. On GADF inputs this is problematic because of Eq. (3): an entry hidden inside a target block can be inferred from its visible transpose. We therefore couple each target block with its transposed mirror: for a block at position (t, l, h, w) (top, left, height, width), both this block and (l, t, w, h) are excluded from the context. The encoder observes neither the target region nor its anti-symmetric counterpart, removing this shortcut without modifying the backbone or positional encoding.

3.3 Joint Training with Prototypical Networks

The I-JEPA objective in Eq. (4) learns representations from unlabelled spectra, but it is agnostic to the downstream regression target: it shapes the embedding space to be internally predictive across masked regions, yet nothing in $\mathcal{L}_{\mathrm{JEPA}}$ encourages that space to encode how a soil property is read off from a spectrum. At pretraining time we do, however, have access to labelled episodes from soil-property tasks that are related to—but disjoint from—the tasks used for downstream adaptation. We exploit this signal by attaching a Prototypical Network [14] head and optimising both objectives jointly:

$$\mathcal{L} = \mathcal{L}_{\mathrm{JEPA}} + \lambda\, \mathcal{L}_{\mathrm{Proto}} \tag{5}$$

Joint optimisation is used so that the episodic regression signal shapes the backbone during representation learning, rather than being introduced only after a task-agnostic self-supervised stage has converged. The task term therefore encourages the embedding space to preserve spectral structure that is useful for few-shot property prediction, while the I-JEPA term exploits the unlabelled spectra through masked latent prediction.

Prototypical head. Each few-shot episode provides a support set $\{(\mathbf{x}_j^s, y_j^s)\}_{j=1}^{N_s}$ and a query set $\{\mathbf{x}_i^q\}_{i=1}^{N_q}$ drawn from the same regression task. Support and query images are encoded by the context encoder f_θ, whose patch tokens are mean-pooled and projected by a lightweight head h_ψ:

$$\mathbf{z} = h_\psi\Big(\tfrac{1}{196}\textstyle\sum_{t=1}^{196} f_\theta(\mathbf{x})_t\Big). \tag{6}$$

Following the metric-learning principle of Prototypical Networks but adapting it to a continuous target, we replace per-class prototypes and a classification softmax with a support-weighted prediction [14]. Each query is predicted as a weighted average of the support labels, where the weights are a temperature-scaled softmax over its distances to the support embeddings:

$$y\hat{}i = \sum_{j=1}^{N_s} w_{ij}\, y_j^s, \qquad w_{ij} = \frac{\exp\big(-\|\mathbf{z}_i^q - \mathbf{z}_j^s\|_2/\gamma\big)}{\sum_{k=1}^{N_s}\exp\big(-\|\mathbf{z}_i^q - \mathbf{z}_k^s\|_2/\gamma\big)}, \tag{7}$$

with temperature γ. The task loss is the mean squared error of these predictions against the query labels,

$$\mathcal{L}_{\text{Proto}} = \frac{1}{N_q}\sum_{i=1}^{N_q}\big(\hat{y}_i - y_i^q\big)^2. \tag{8}$$

Crucially, in the joint objective L_{Proto} is computed on the context encoder, so its gradients flow into the backbone and shape the representation during pretraining; the EMA target encoder is used only to produce I-JEPA targets with stop-gradient. The projection h_ψ admits several designs, which we compare in Section 5.1.

4 Experimental Setup

4.1 Dataset and Task Protocol

Experiments are conducted on the soil NIR subset of the task-oriented few-shot spectroscopy benchmark introduced by Garzón et al. [7], which is derived from the Open Soil Spectral Library (OSSL) [18]. The subset contains soil reflectance spectra in the near-infrared range, with 1051 wavelengths covering 400–2499 nm. Following the preprocessing pipeline described in Section 3.1, each spectrum is converted into a 224×224 single-channel GADF image before being passed to the visual backbone.

The soil NIR subset comprises 185 soil-property regression tasks, following the benchmark protocol of Garzón et al. [7]. These tasks are divided into three disjoint groups: 91 tasks are used to sample labelled episodes for the prototypical loss during joint pretraining, 54 tasks are used for online validation during pretraining, and 40 held-out tasks are reserved for the final offline evaluation. The self-supervised I-JEPA objective is optimized over the available spectra without

using their task labels, whereas the ProtoNet term is computed only on labelled episodes sampled from the training tasks.

During episodic training and online validation, episodes are sampled with $k_{spt} = 25$ support examples and $k_{qry} = 25$ query examples, following the base protocol of [7]. This 25-shot setting provides a reasonable compromise for few-shot regression: it remains challenging due to the limited support set, while still providing enough samples to estimate a stable local input–target relationship within each task. For the final offline evaluation, we use a cross-validated leave-one-block-out protocol, where each support block contains 25 samples and all remaining samples of the task are used as query instances.

4.2 Architectures and Optimisation

The main model uses a ViT encoder as the shared backbone for both the I-JEPA objective and the ProtoNet head. The same backbone, input representation, episodic protocol, and evaluation split are also used in the scratch ProtoNet control, so that the comparison isolates the effect of joint pretraining rather than changes in model capacity.

The masking strategy follows I-JEPA [2], but uses the symmetric masking scheme described in Section 3.2 to account for the antisymmetric structure of GADF inputs. In all experiments, target blocks and their transposed counterparts are removed from the context, preventing the model from recovering masked entries through their visible mirror locations.

Optimisation details are summarised in Table 1. The main reported configuration uses AdamW with a cosine learning-rate schedule and an EMA target encoder whose momentum is annealed from 0.996 to 1.0. For the remaining model variants, the learning rate was selected independently through a small grid search. The ProtoNet loss weight was set to $\lambda = 0.025$ for the main transformer-projection and Savitzky–Golay plus first-derivative variants, and to $\lambda = 0.05$ for the no-projection variant. The temperature γ in Eq. (7) was fixed to 0.5, a standard value for softmax-weighted distance-based prediction in metric-based few-shot learning.

Table 1. Main architectural and training hyperparameters.

Hyperparameter	**Value**	**Hyperparameter**	**Value**
Backbone	ViT-Base, $\sim$ 86M	Input / patch size	224×224 / 16×16
Depth / heads	12 / 12	Tokens	196
Batch size	128	Support / query	25 / 25
Target blocks	4	Context scale	[0.85, 1.0]
Target scale	[0.15, 0.2]	EMA momentum	$0.996 \rightarrow 1.0$
Optimiser	AdamW	Epochs	750
Learning-rate peak	10^{-3}	Proto loss weight λ	0.025 / 0.05
Temperature γ	0.5		

4.3 Baselines and Model Variants

The central comparison is between the proposed joint I-JEPA + ProtoNet model and a ProtoNet trained from scratch with the same ViT backbone. Since both models share the architecture, input representation, episodic protocol, and evaluation split, this control provides the fairest estimate of the contribution of joint pretraining.

We also evaluate an I-JEPA-only variant, which removes the task-aware ProtoNet term, and a sequential I-JEPA followed by ProtoNet training variant. Backbone capacity is assessed with ViT-Small, ViT-Base, and ViT-Large joint models. The ResNet18 + ProtoNet model of Garzón et al. [7] is included as a protocol-comparable convolutional reference, although it changes the backbone family and is therefore not an isolated test of pretraining. Finally, MAML [6] is included as an optimisation-based meta-learning baseline, and TabPFN [9] as an external small-data tabular baseline applied directly to the NIR feature vector.

5 Results and Discussion

Table 2 reports the cross-validated performance on the 40 held-out soil-property regression tasks. Metrics are computed independently for each fold, averaged within each task, and then averaged across tasks. Since tasks correspond to different soil properties, MSE and MAE are reported in standardised target units.

The proposed joint I-JEPA + ProtoNet model obtains the best result among the ViT-based methods, improving the same ViT ProtoNet architecture from $R^2 = 0.411$ to $R^2 = 0.467$. This corresponds to a gain of 5.6 percentage points, while also reducing MSE from 0.575 to 0.534 and MAE from 0.449 to 0.440. This same-backbone comparison is the central result of the paper: *joint task-aware self-supervised pretraining learns a more useful representation than episodic training of the ViT backbone alone.*

The comparison with I-JEPA-only and sequential training clarifies the role of the task-aware signal. Pure I-JEPA reaches only $R^2 = 0.369$, below the scratch ProtoNet control, indicating that masked latent prediction alone does not organise the embedding space sufficiently for few-shot soil-property regression. Sequential I-JEPA followed by ProtoNet training improves this result to $R^2 = 0.414$, but remains close to the scratch control and is worse in MSE and MAE. The benefit therefore appears when the prototypical loss shapes the backbone during representation learning.

The strongest overall result is obtained by the convolutional ProtoNet ResNet18 reference, with $R^2 = 0.492$. The proposed model reduces the gap to 2.5 percentage points, but does not surpass it. Since this comparison changes the backbone family, it should not be interpreted as a failure of task-aware pretraining. Rather, together with the controlled ViT comparison, it suggests that joint pretraining is beneficial while the remaining limitation is likely related to the match between backbone, data regime, and spectral structure.

Table 2. Cross-validated performance on the held-out few-shot soil-property regression tasks. The best overall result is shown in bold, and the best proposed model is underlined.

Model	$R^2 \uparrow$	MSE $\downarrow$	MAE $\downarrow$
TabPFN	0.160	0.987	0.627
MAML ResNet18	0.210	0.759	0.585
I-JEPA only	0.369	0.606	0.510
Sequential I-JEPA + ProtoNet	0.414	0.582	0.465
ProtoNet ViT scratch	0.411	0.575	0.449
<u>Joint I-JEPA + ProtoNet</u>	<u>0.467</u>	<u>0.534</u>	<u>0.440</u>
ProtoNet ResNet18 [7]	**0.492**	**0.510**	**0.424**

Table 3. Additional ablations of the proposed pipeline. Each block isolates a different design factor.

Factor Variant	$R^2 \uparrow$
Spectral preprocessing	
Savitzky–Golay	0.455
Savitzky–Golay + first derivative	**0.467**
Projection head	
Direct pooled embedding	0.448
Transformer head	**0.467**
Backbone size	
ViT-Small	0.392
ViT-Base (main ViT)	**0.467**
ViT-Large	0.323

5.1 Additional Ablations and Architectural Factors

Table 3 summarises the main ablations of the proposed pipeline. Savitzky–Golay filtering reaches $R^2 = 0.455$, while adding the first derivative increases performance to $R^2 = 0.467$. The projection head also has a moderate effect: directly using the pooled embedding obtains $R^2 = 0.448$, whereas the Transformer head reaches $R^2 = 0.467$. These gains are complementary to the joint training strategy, whose effect was isolated in Table 2.

The backbone-size study is more diagnostic. ViT-Small reaches $R^2 = 0.392$, while ViT-Large drops to $R^2 = 0.323$; therefore, performance does not improve monotonically with transformer capacity. ViT-Base remains the strongest transformer backbone evaluated. This supports the interpretation that the remaining gap with ResNet18 is linked not only to preprocessing or projection design, but to a broader backbone–data mismatch: ViT-Base has approximately 86M parameters, whereas the pretraining corpus contains roughly 45k spectra, and convolutional biases such as locality, weight sharing, and hierarchical feature extraction may remain better aligned with smooth local spectral structure. Future work should therefore explore spectroscopy-oriented transformers with stronger structural bias, such as hybrid convolutional stems, finer patch sizes, or local-attention variants.

6 Conclusions

This work studied task-aware self-supervised pretraining for few-shot soil-property regression from NIR spectra. We encoded one-dimensional spectra as GADF images, adapted the I-JEPA masking strategy to their antisymmetric structure, and jointly optimised masked latent prediction with a prototypical regression loss. The proposed joint model improves the same ViT ProtoNet architecture from $R^2 = 0.411$ to $R^2 = 0.467$, whereas pure I-JEPA and sequential pretraining provide weaker results. This shows that the task-aware signal is most useful when it shapes the backbone during representation learning. Nevertheless, the method remains below the convolutional ResNet18 ProtoNet reference ($R^2 = 0.492$), suggesting that the main limitation is the match between a large ViT backbone, the available pretraining corpus, and the local structure of NIR spectra. Future work should therefore explore spectroscopy-oriented transformer variants with stronger structural bias, including hybrid convolutional stems, finer patch sizes, local attention, and self-supervised objectives designed directly for one-dimensional spectra.

Acknowledgments This publication is part of the R&D&I project PID2024-156434NB-I00 (CONFIA2), funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU. This publication is also part of the Project "Ethical, Responsible and General Purpose Artificial Intelligence: Applications In Risk Scenarios" (IAFER) Exp.:TSI-100927-2023-1 funded through the Creation of university-industry research programs (Enia Programs), aimed at the research and development of artificial intelligence, for its dissemination and education within the framework of the Recovery, Transformation and Resilience Plan from the European Union Next Generation EU through the Ministry for Digital Transformation and the Civil Service.

References

1. Akkaya, I.B., Kathiresan, S.S., Arani, E., Zonooz, B.: Enhancing performance of vision transformers on small datasets through local inductive bias incorporation. Pattern Recognition **153**, 110510 (2024)
2. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
3. Chen, W.Y., Liu, Y.C., Kira, Z., Wang, Y.C.F., Huang, J.B.: A closer look at few-shot classification. In: International Conference on Learning Representations (2019)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations

5. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: Convit: Improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning. pp. 2286–2296. PMLR (2021)
6. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135. PMLR (2017)
7. Garzón, I., Gomez-Trenado, G., Damas, S.: A task-oriented few-shot spectroscopy benchmark: Empirical evaluation of meta-learning approaches. Engineering Applications of Artificial Intelligence **167**, 113593 (2026)
8. Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A.: Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8342–8360 (2020)
9. Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., Hutter, F.: Accurate predictions on small data with a tabular foundation model. Nature **637**(8045), 319–326 (2025)
10. Li, Y., Sun, H., Zheng, Y., Wei, Q., Chen, Z., Zhang, J., Qi, H., Zhang, C., Chen, F.: Combined gramian angular difference field image coding and improved mobile vision transformer for determination of apple soluble solids content by vis-nir spectroscopy. Journal of Food Composition and Analysis **131**, 106200 (2024)
11. Luo, X., Wu, H., Zhang, J., Gao, L., Xu, J., Song, J.: A closer look at few-shot classification again. In: International Conference on Machine Learning. pp. 23103–23123. PMLR (2023)
12. Oberholzer, S., Summerauer, L., Steffens, M., Ifejika Speranza, C.: Best performances of visible–near-infrared models in soils with little carbonate–a field study in switzerland. Soil **10**(1), 231–249 (2024)
13. Smith, S.L., Brock, A., Berrada, L., De, S.: Convnets match vision transformers at scale. arXiv preprint arXiv:2310.16764 (2023)
14. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30** (2017)
15. Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J.B., Isola, P.: Rethinking few-shot image classification: a good embedding is all you need? In: European Conference on Computer Vision. pp. 266–282. Springer (2020)
16. Wan, M., Yan, T., Xu, G., Liu, A., Zhou, Y., Wang, H., Jin, X.: Mae-nir: A masked autoencoder that enhances near-infrared spectral data to predict soil properties. Computers and Electronics in Agriculture **215**, 108427 (2023)
17. Wang, Z., Oates, T.: Imaging time-series to improve classification and imputation. In: Proceedings of the 24th International Conference on Artificial Intelligence. pp. 3939–3945 (2015)
18. Woodwell Climate Research Center, University of Florida, OpenGeoHub and contributors: Open Soil Spectral Library: technical manual (2021), https://soilspectroscopy.github.io/ossl-manual/
19. Zhao, R., Li, W., Xu, J., Chen, L., Wei, X., Kong, X.: A cnn-based self-supervised learning framework for small-sample near-infrared spectroscopy classification. Analytical Methods **17**(5), 1090–1100 (2025)