Structured spectral reconstruction for scalable soil organic carbon inference

Evan A. Coleman
MIT Climate and Sustainability Consortium
ecol@mit.edu

Sujay Nair
Georgia Institute of Technology
sujaynr@mit.edu

Xinyi Zeng
Coho Climate Advisors
zeng.xinyi@hotmail.com

Elsa Olivetti
MIT Department of Materials Science & Engineering
elsao@mit.edu

Abstract

Measuring soil organic carbon (SOC) inexpensively and accurately is crucial for soil health monitoring and agricultural decarbonization. Hyperspectral imaging is commonly evaluated as an inexpensive alternative to dry combustion for SOC measurement, but existing end-to-end approaches trained to predict SOC content from spectral data frequently fail to generalize when applied outside of their ground-truth geographic sampling distributions. Using stratified data from the USDA Rapid Carbon Assessment (RaCA), we demonstrate a method to improve model generalization out-of-distribution by training SOC regression alongside models that reconstruct input spectra. Because hyperspectra can be collected from remote platforms such as drones and satellites, this approach raises the possibility of using large hyperspectral Earth observation datasets to transfer SOC inference models to remote geographies where geographically-dense ground-truth data collection may be expensive or impossible. By replacing the decoder with a simple physics-informed model, we also learn an interpretable spectral signature of SOC, confirming its dark hue and expected reflectance troughs. Finally, we show that catastrophic generalization failures can be better addressed with these architectures by fine-tuning on large quantities of hyperspectral data.

1 Introduction

Modern intensive agriculture yields direct greenhouse gas emissions from energy consumption, fertilizer production, livestock husbandry, and land use change, accounting for roughly 20% of global anthropogenic contributions to climate change (1). An avenue of significant interest to decarbonize agriculture is to augment organic carbon stored in agricultural soils through local shifts in farming practices. Approaches that have been tested include zero- or reduced-tillage, cover cropping, and crop residue utilization (2). Such practices can bolster the interactions of plants' root systems with bugs and microorganisms, leading to increased deposition of carbon from the atmosphere into soil (2; 3; 4). However, measurements to verify that these sustainable practices have their intended effects on soil organic carbon (SOC) stocks are prohibitively expensive to apply to large areas of land (∼$15-25 USD/sample) (5; 6). These steep costs have motivated investigation into more scalable approaches for SOC inference.

In this work, we investigate hyperspectral imaging (HSI) as a candidate procedure to scale SOC measurement. The standard technique to measure SOC content requires combusting soil to determine its total carbon content, and mixing it with acid to determine its inorganic carbon content (stored in the form of carbonate minerals). SOC is the difference between these quantities (7). Relative to this laboratory procedure, HSI is non-destructive, reagent-free, emissions-free, and may be performed *in situ* or remotely, via drone or satellite (8). However, existing analyses involve end-to-end regression between measured spectra and laboratory SOC measurements which, as we show in this work, can drastically fail when

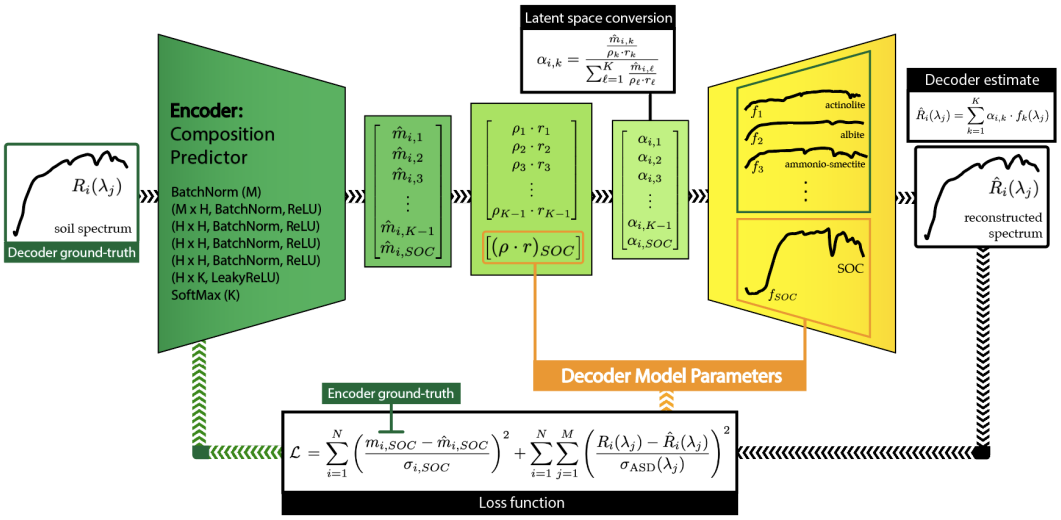

Figure 1: Autoencoder in the physics-informed case (see appendix B). A soil spectrum $R(\lambda)$ is used to predict endmember mass fractions $\vec{m}$ via a differentiable process (here, a fully-connected multi-layer perceptron). To preserve physical interpretability, a correction is applied, converting $\hat{m}_k$ to mixing ratios α_k for pure endmember spectra f_k. The α_k and f_k are then combined to reconstruct an estimate $\hat{R}(\lambda)$ of the original $R(\lambda)$. In the end-to-end approach, the decoder model is replaced by an ANN using $\hat{m}_{i,k}$ as input features.

applied outside of their sampling distributions (OOD) (see appendix A for related work). Such inabilities to generalize undermine the feasibility of applying HSI to SOC measurement. Methods to reliably transfer SOC inference models based on HSI data, i.e. requiring minimal combustion data, are thus of major interest. We demonstrate how to better achieve this goal by training models for SOC measurement alongside models which reconstruct the original input spectrum. A major benefit of our approach is that it permits training models using hyperspectral data alone, which may enable the global scaling of SOC measurement via HSI thanks to emerging capabilities in continuous hyperspectral Earth observation (9; 10).

We present 3 novel contributions: (1) A method for soil property analysis which uses spectral reconstruction as an auxiliary loss to enable the use of unlabeled data during training; (2) An application of this architecture which backs out an interpretable reference spectrum for SOC using physical priors; and (3) a replicable empirical study using a cleaned dataset, which demonstrates the failure of SOC inference models to predict OOD, and shows how our method enables improvements in generalization performance.

2 Method

In this work, we analyze a large library of soil carbon contents and soil reflectances. The SOC content of sample i is represented as a fraction $m_{i,SOC}$, and its reflectance as a vector over M wavelengths: $R_i(\lambda_j)$, with $j = 1, \ldots, M$. Our objective is to estimate the % by mass of SOC $\hat{m}_{i,SOC}$ in a given sample of soil i using only R_i. We refer to the model trained to solve this task as the encoder model $\mathcal{E}[R_i(\vec{\lambda})] = \hat{m}_{i,SOC}$. We then couple $\mathcal{E}$ to a decoder model $\mathcal{D}$ which solves the inverse inference problem: $\mathcal{D}\big[\mathcal{E}[R_i(\vec{\lambda})]\big] = \hat{R}_i(\vec{\lambda})$. We include $K-1$ auxiliary variables in the output space of the encoder model, $\hat{m}_1, \ldots, \hat{m}_{K-1}$, so that the latent space between $\mathcal{E}$ and $\mathcal{D}$ is a K-vector $\vec{m}$. Figure 1 depicts this architecture when $\mathcal{D}$ is the physics-informed model discussed below. We evaluate $\mathcal{E}$'s performance by root mean square error (RMSE), R^2, and mean bias $\langle \hat{m}_{i,SOC} - m_{i,SOC} \rangle$. To evaluate the performance of $\mathcal{D}$, we use the mean RMSE per reconstructed spectrum, which we denote MRMSE.

We analyzed soil spectral data from the USDA Rapid Carbon Assessment Project (RaCA) (11), a campaign which stratified the conterminous United States into 17 disjoint regions based on major land resource area (MLRA) and land use land cover (LULC) classifications (see Figure 2). Hyperspectral scans were performed on over 120,000 sam-

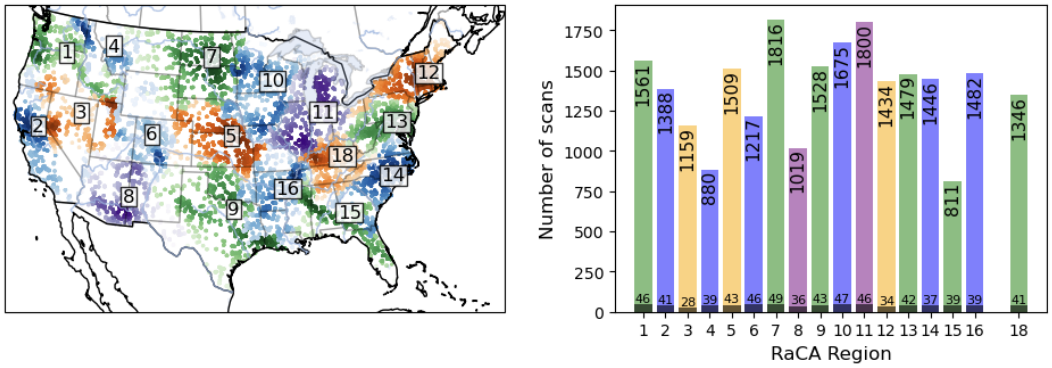

Figure 2: (Left) Geographic distribution of the RaCA data, labeled by RaCA region. (Right) Data available by region. Black bars show example 10-pedon subsamples for Experiment 3.

ples collected from soil cores (or *pedons*) taken from those regions using an ASD LabSpec spectrophotometer at 1 nm resolution within the Visible-Near Infrared (Vis-NIR) spectral window of 350-2500 nm. Prior to scanning, each soil sample was air-dried, ground, and sieved to a particle size below 2 mm. A representative fraction of the samples was sent to a laboratory for total carbon measurement (via dry combustion) and carbonate content measurement (via the modified pressure calcimeter method). Organic carbon was computed as (total carbon) − (inorganic carbon). We identified 23,291 unique pairs of spectral data and SOC content. We removed one outlier with high water content (48% by mass) as well as 40 spectra where the ASD spectrophotometer failed to log reflectances for one or more wavelengths. Due to frequent pixel failures at lower wavelengths, we removed reflectance data below 365 nm. Following selection cuts, we obtain 23,550 soil spectra within a spectral window of 365-2500 nm, alongside laboratory-measured SOC content.

In part of our analysis, we used a physics-informed model in place of $\mathcal{D}$ (see appendix B for details) with 2,152 parameters, including the spectrum of pure SOC $f_{SOC}(\lambda_j)$ and a nuisance parameter $(\rho \cdot r)_{SOC}$. To construct this model, we used 93 characteristic spectra from the United States Geological Survey Spectral Library (12) corresponding to minerals commonly found in soils. Relevant mineral properties such as density, unit cell size, and commonality in soils were scraped from textbook sources (13) and the Mindat.org mineralogical database (14).

With this dataset, we have $M = 2{,}136$ and $K = 94$. The loss function is taken as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{m_{i,SOC} - \hat{m}_{i,SOC}}{\sigma_{i,SOC}}\right)^2 + \frac{1}{N' \cdot M}\sum_{i=1}^{N'}\sum_{j=1}^{M}\left(\frac{R_i(\lambda_j) - \hat{R}_i(\lambda_j)}{\sigma_{\text{ASD}}(\lambda_j)}\right)^2, \qquad (1)$$

where $N \neq N'$ in general. Here $\hat{m}_{i,SOC}$, $\hat{R}_i(\lambda_j) \in [0, 1]$. We take $\sigma_{i,SOC} = 0.0041$ (15; 16) and $\sigma_{ASD}(\lambda_j) = 0.01$. The artificial neural network (ANN) encoder in Figure 1 has $H = 512$ with 1,944,360 parameters. The ANN decoder model has a similar size (take $M \leftrightarrow K$ in Figure 1 and replace SoftMax with Sigmoid). For the physics-informed decoder model, $f_{SOC}(\lambda_j)$ is initialized to 0.5, and $(\rho \cdot r)_{SOC}$ is set to the mean $(\rho \cdot r)$ for mineral endmembers.

3 Experiment

We investigated 3 points for 3 architectures: no decoder ($N' = 0$), ANN decoder, and physics-informed decoder (both with $N' \geq N$). In Experiment 1, we quantified the OOD performance of SOC content regression under worst case conditions, when the validation data comes from a distinct land class. In Experiment 2, we analyzed the regressed values of $f_{SOC}(\lambda_j)$ in the physics-informed decoder model trainings of Experiment 1 for interpretable properties matching observations in the literature. Lastly, in Experiment 3 we tested whether the burden of ground-truth SOC data collection could be reduced via transfer learning, by fine-tuning pretrained models on OOD spectral data. (See appendix C for tabulated results.)

Experiment 1: *Does failure to perform spectral reconstruction signal generalization failure?* In each of 4 runs fixing random seeds, we compared 17 trainings of 1,000 epochs on a dataset

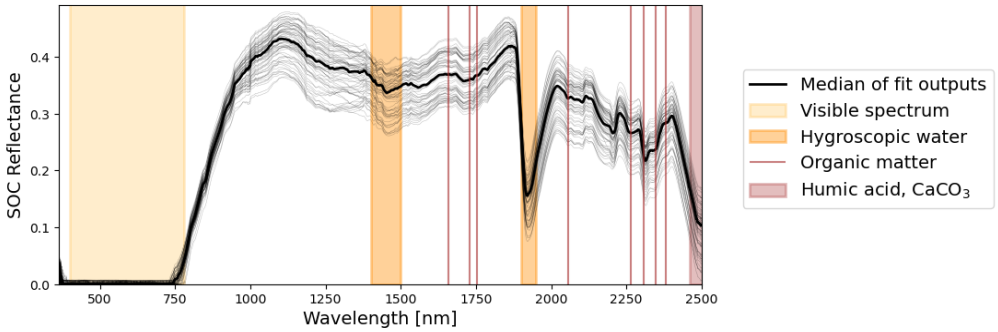

Figure 3: The spectral signatures of SOC discussed in Experiment 2.

comprised of 16 out of 17 RaCA regions, with one region left as validation. For architectures with decoders, we had $N' = N$. Regions 3, 5, 6, 8, 16, and 18 demonstrated catastrophically poor generalization performance, with validation $R^2 < 0$ for all models. The remaining 11 regions had $R^2 > 0$ on the validation set in the vast majority of cases. No significant difference was found between the average encoder performance of the 3 models. However, for the ANN decoder the validation MRMSE was $(3.358 \pm 0.179)\%$ on average for the experiments which generalized, whereas that of the 6 problematic regions was $(4.984 \pm 0.338)\%$. Per a two-tailed Welch's t-test, this is a statistically significant difference in performance ($t = 4.681$, $p < 0.0001$). The same observations held in the physics-informed case ($t = 4.789$, $p < 0.0001$). *Takeaway:* These results provide evidence that poor spectral reconstruction in a given geography can signal OOD generalization failure without the use of combustion data.

Experiment 2: *Is the regressed spectrum of SOC physically interpretable?*
Since SOC is not a single compound, and is not fully separable from soil, the optical spectrum of "pure" SOC cannot be determined from laboratory methods (7). However, our approach enables regression of this unknown via the physics-informed decoder model. For the trainings in Experiment 1, we found $(\rho \cdot r)_{SOC} = 2.471 \pm 0.132$ mg m^{-2}. Figure 3 depicts the regressed $f_{SOC}(\lambda_j)$, demonstrating their variability due to changes in the geographic sampling distribution. These spectra collectively confirm long-standing observations regarding SOC in the literature: the Munsell color is black (8), and it has reflectance troughs noted in e.g. figure 2 of (17) (red lines in figure 3). The troughs near 1400 and 1900 nm are likely due to the presence of trapped moisture (18; 19). That near 2500 nm is attributed to both humic acid (an extractible component of SOC) (18) and calcium carbonate (20), suggesting that the model may not perfectly distinguish SOC from one or more sources of inorganic carbon. *Takeaway:* The physics-informed decoder model reverse-engineers known properties of SOC.

Experiment 3: *Can a surplus of HSI data improve scale-up of SOC inference?*
To demonstrate the benefit of HSI's ease of application relative to laboratory methods, we trained on a small subset of ground-truth combustion data ($\sim 5\%$) alongside the complete set of hyperspectral scans, to compare architectures in the more realistic regime where $N' \gg N$. First, we selected 10 pedons at random from each RaCA region. In each comparison, we set aside one RaCA region as a validation dataset. For training, we combined combustion data for all 160 pedons of the remaining 16 regions with the reflectances of all samples in those regions. We trained for 500 epochs. From the validation region, we selected 5 pedons at random and combined their combustion data with the full set of reflectances for the validation region. We fine-tuned on this dataset for 25 epochs. We repeated the full experiment over 4 fixed random seeds. After initial training, we found an average validation $R^2 < 0$ for all architectures, with no significant performance differences across architectures. After fine-tuning, only the models with ANN decoders demonstrated R^2 values significantly larger than 0 on the validation set (0.210 ± 0.068). While we observed generally poor model performance in this regime (likely due to the small size of the training and fine-tuning datasets), these results demonstrate a clear use case for our approach, and motivate further analysis. *Takeaway:* Our approach can better support the transfer of SOC models to local conditions when much more HSI data are available for analysis than combustion data.

4 Discussion

We presented an architecture which can exploit the versatility of HSI to scale soil property inference, demonstrating the utility of spectral reconstruction in: (1) signalling the occurrence of OOD behavior without expensive laboratory analyses; (2) extracting physically-interpretable signatures of SOC; and (3) bolstering model generalization through fine-tuning.

References

[1] F. N. Tubiello, K. Karl, A. Flammini, J. Gütschow, G. Obli-Laryea, G. Conchedda, X. Pan, S. Y. Qi, H. Halldórudóttir Heiðarsdóttir, N. Wanner, R. Quadrelli, L. Rocha Souza, P. Benoit, M. Hayek, D. Sandalow, E. Mencos Contreras, C. Rosenzweig, J. Rosero Moncayo, P. Conforti, and M. Torero. Pre- and post-production processes increasingly dominate greenhouse gas emissions from agri-food systems. *Earth System Science Data*, 14(4):1795–1809, 2022.

[2] Rattan Lal. Soils and sustainable agriculture. A review. *Agronomy for Sustainable Development*, 28:57–64, 2008.

[3] Rattan Lal. Sequestration of atmospheric CO_2 in global carbon pools. *Energy & Environmental Science*, 1(1):86–100, 2008.

[4] Robert B Jackson, Kate Lajtha, Susan E Crow, Gustaf Hugelius, Marc G Kramer, and Gervasio Piñeiro. The ecology of soil carbon: pools, vulnerabilities, and biotic and abiotic controls. *Annual review of ecology, evolution, and systematics*, 48:419–445, 2017.

[5] Dianna K Bagnall, Elizabeth L Rieke, Cristine LS Morgan, Daniel L Liptzin, Shannon B Cappellazzi, and C Wayne Honeycutt. A minimum suite of soil health indicators for North American agriculture. *Soil Security*, 10:100084, 2023.

[6] Charlotte E Norris, G Mac Bean, Shannon B Cappellazzi, Michael Cope, Kelsey LH Greub, Daniel Liptzin, Elizabeth L Rieke, Paul W Tracy, Cristine LS Morgan, and C Wayne Honeycutt. Introducing the North American project to evaluate soil health measurements. *Agronomy Journal*, 112(4):3195–3215, 2020.

[7] Johannes Lehmann and Markus Kleber. The contentious nature of soil organic matter. *Nature*, 528(7580):60–68, 2015.

[8] Marion F Baumgardner, LeRoy F Silva, Larry L Biehl, and Eric R Stoner. Reflectance properties of soils. *Advances in agronomy*, 38:1–44, 1986.

[9] Shen-En Qian. Hyperspectral satellites, evolution, and development history. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:7032–7056, 2021.

[10] Max Joshua, Katie Salvaggio, Mark Keremedjiev, Keely Roth, and Elizabeth Foughty. Planet's upcoming vis-swir hyperspectral satellites. In *Optica Sensing Congress 2023 (AIS, FTS, HISE, Sensors, ES)*, page HM3C.5. Optica Publishing Group, 2023.

[11] USDA NRCS. Rapid Carbon Assessment (RaCA). 11 2023.

[12] RF Kokaly, RN Clark, GA Swayze, KE Livo, TM Hoefen, NC Pearson, RA Wise, WM Benzel, HA Lowers, RL Driscoll, et al. USGS Spectral Library version 7 data: US Geological Survey data release. *United States Geological Survey (USGS): Reston, VA, USA*, 61, 2017.

[13] Joe Boris Dixon, Sterling Barg Weed, and RL Parpitt. Minerals in soil environments. *Soil Science*, 150(2):562, 1990.

[14] Hudson Institute of Mineralogy. Mindat.org. 2023.

[15] Rupert EH Kuveke, Lachlan Barwise, Yara van Ingen, Kanika Vashisth, Nicholas Roberts, Saurabh S Chitnis, Jason L Dutton, Caleb D Martin, and Rebecca L Melen. An international study evaluating elemental analysis, 2022.

[16] Christopher A Stanbery, Jennifer L Pierce, Shawn G Benner, and Kathleen Lohse. On the rocks: Quantifying storage of inorganic soil carbon on gravels and determining pedon-scale variability. *Catena*, 157:436–442, 2017.

[17] Bo Stenberg, Raphael A Viscarra Rossel, Abdul Mounem Mouazen, and Johanna Wetterlind. Visible and near infrared spectroscopy in soil science. *Advances in agronomy*, 107:163–215, 2010.

[18] Cheng-Wen Chang and David A Laird. Near-infrared reflectance spectroscopic analysis of soil C and N. *Soil science*, 167(2):110–116, 2002.

[19] Janice L Bishop, Carle M Pieters, and John O Edwards. Infrared spectroscopic analyses on the nature of water in montmorillonite. *Clays and clay minerals*, 42:702–716, 1994.

[20] Roger N Clark et al. Spectroscopy of rocks and minerals, and principles of spectroscopy, 2020.

[21] F Deng, M Knadel, Y Peng, G Heckrath, M Greve, and B Minasny. Soil profile organic carbon prediction with visible-near infrared reflectance spectroscopy based on a national database. *Digital soil assessments and beyond*, page 409, 2012.

[22] Yufeng Ge, Cristine LS Morgan, and Jason P Ackerson. VisNIR spectra of dried ground soils predict properties of soils scanned moist and intact. *Geoderma*, 221:61–69, 2014.

[23] Mervin St Luce, Noura Ziadi, and Raphael A Viscarra Rossel. GLOBAL-LOCAL: A new approach for local predictions of soil organic carbon content using large soil spectral libraries. *Geoderma*, 425:116048, 2022.

[24] Marco Nocita, Antoine Stevens, Gergely Toth, Panos Panagos, Bas van Wesemael, and Luca Montanarella. Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry*, 68:337–347, 2014.

[25] Said Nawar, Muhammad Abdul Munnaf, and Abdul Mounem Mouazen. Machine learning based on-line prediction of soil organic carbon after removal of soil moisture effect. *Remote Sensing*, 12(8):1308, 2020.

[26] Nuwan K Wijewardane, Yufeng Ge, Skye Wills, and Terry Loecke. Prediction of soil carbon in the conterminous United States: visible and near infrared reflectance spectroscopy analysis of the rapid carbon assessment project. *Soil Science Society of America Journal*, 80(4):973–982, 2016.

[27] J Padarian, B Minasny, and AB McBratney. Transfer learning to localise a continental soil vis-NIR calibration model. *Geoderma*, 340:279–288, 2019.

[28] Zefang Shen, Leonardo Ramirez-Lopez, Thorsten Behrens, Lei Cui, Mingxi Zhang, Lewis Walden, Johanna Wetterlind, Zhou Shi, Kenneth A Sudduth, Philipp Baumann, et al. Deep transfer learning of global spectra for local soil carbon monitoring. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188:190–200, 2022.

[29] José M Soriano-Disla, Les J Janik, Raphael A Viscarra Rossel, Lynne M Macdonald, and Michael J McLaughlin. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Applied spectroscopy reviews*, 49(2):139–186, 2014.

[30] Nichola M Knox, Sabine Grunwald, ML McDowell, Gregory L Bruland, DB Myers, and WG Harris. Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. *Geoderma*, 239:229–239, 2015.

[31] Wartini Ng, Budiman Minasny, Sang Ho Jeon, and Alex McBratney. Mid-infrared spectroscopy for accurate measurement of an extensive set of soil properties for assessing soil functions. *Soil Security*, 6:100043, 2022.

[32] Jason P Ackerson, CLS Morgan, and Y Ge. Penetrometer-mounted VisNIR spectroscopy: Application of EPO-PLS to in situ VisNIR spectra. *Geoderma*, 286:131–138, 2017.

[33] Nuwan K Wijewardane, Sarah Hetrick, Jason Ackerson, Cristine LS Morgan, and Yufeng Ge. VisNIR integrated multi-sensing penetrometer for in situ high-resolution vertical soil sensing. *Soil and Tillage Research*, 199:104604, 2020.

[34] Jerzy Weber, Yona Chen, Elżbieta Jamroz, and Teodoro Miano. Preface: humic substances in the environment. *Journal of Soils and Sediments*, 18(8):2665–2667, 2018.

[35] José M Bioucas-Dias, Antonio Plaza, Nicolas Dobigeon, Mario Parente, Qian Du, Paul Gader, and Jocelyn Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE journal of selected topics in applied earth observations and remote sensing*, 5(2):354–379, 2012.

A Related Work

A significant body of work, spanning decades, has applied HSI to the task of soil property inference (8; 21; 22; 23). Measurements of interest include soil moisture, pH, and nutrient contents such as organic carbon, nitrogen, and potassium content (17). Prior to the advent of machine learning, and still in recent years, common analysis methods have included generalizations of linear regression techniques such as partial least squares regression (PLSR) and Cubist models (24; 25). The most common metrics of performance include root mean square error (RMSE), mean bias, and correlation coefficients R^2 between measurements and predictions.

In one of the first works performing end-to-end regression analysis on the RaCA dataset (26), the same dataset used in our analysis, the authors compared PLSR against artificial neural networks (ANNs), finding superior performance with the ANN. Using a 60:40 training-validation split over the full dataset, they achieved an RMSE of 3.59%, $R^2 = 0.96$ (training), and an RMSE of 3.61%, $R^2 = 0.96$ (validation). As the goal of the study was to develop algorithms useful for carbon stock assessment of the conterminous United States, leave-one-region-out cross-validation was not performed; there was no emphasis on the generalization of trained models outside of the geographic sampling distribution.

More recent efforts have focused on optimizing local model development for SOC prediction utilizing large soil spectral libraries (SSLs). In (27), the authors train ANNs to predict SOC content using the LUCAS database, but with an emphasis on out-of-distribution transfer of results. They demonstrate that the use of an ANN pre-trained on a global SSL and subsequently fine-tuned on local data performed better than local models in 91% of cases, with an average reduction in RMSE on the order of 10%. More recently, the authors of (28) extended this approach on 1D CNNs, weighing the benefit of transferring pre-trained model weights versus instance data from SSLs when developing local models. They found that transferring instance data directly led to the most consistent improvement, an average reduction in RMSE of 25.8% compared with local modeling. However, in these studies, no attempt was made to simultaneously solve the inverse inference problem, i.e. using regressed soil properties such as SOC to reconstruct the original input spectrum. This means that transfer learning techniques will still require large soil spectral libraries to be developed which contain paired combinations of spectra and ground-truth SOC data. In this work, we demonstrate how training a model to solve the forward problem (SOC estimation) alongside the inverse problem (spectral reconstruction) improves the efficacy of transfer learning techniques for SOC estimation without necessitating the collection of expensive label data from laboratory combustion analysis.

There is active interest in applying HSI to soil property inference within other spectral ranges, such as MIR (mid-infrared, 2,500-25,000 nm) (29; 30; 31), and also in combination with proximal measurement devices such as penetrometers (32; 33). In this work, we focused on Vis-NIR spectroscopy because it presents the greatest immediate opportunity to scale global soil data collection and calibrate universal models for soil property inference, using only those input features which can be accessed remotely, such as from a drone- or satellite-borne

platform. Existing and planned hyperspectral satellites utilize sensors calibrated to the Vis-NIR spectral window (9; 10). The experiments we perform in this work are designed to be deployed, with the objective of scaling SOC inference to regions where geographic stratification of soil classes is possible but direct sampling is expensive. The methodology presented here can enable the utilization of emerging capabilities in continuous hyperspectral Earth observation in order to transfer models for soil property inference to new environments and geographies.

B Physics-informed model: motivation and assumptions

The implementation of the decoder network $\mathcal{D}$ presents an opportunity to regress spectral features of SOC which are not otherwise accessible through laboratory methods. Taking the latent space $\vec{m}$ to represent mass fractions of well-characterized soil components, or endmembers, then $\mathcal{D}$ may be interpreted as reconstructing the spectrum of the admixture of those components, including SOC. A physics-informed approach to this spectral mixing generically requires synthesizing the spectra for pure samples of each endmember. However, known methodologies to isolate organic compounds from soil are inherently destructive (34), so it is not possible to perform a spectral scan of "pure" SOC. At the same time, SOC does not have a unique chemical structure, but rather is an amalgamation of organic compounds present in soil whose distribution is sensitive to multiple local sources of variation, such as vegetation types or average soil temperature and moisture (4). A tool to extract the spectral fingerprint of SOC appropriate to a given region can be used as a diagnostic, to analyze *post facto* why a trained model for soil property inference may have failed to generalize.

In part of this work, we considered the case where an end-to-end decoder model is replaced by a physics-informed linear mixing model, combining known soil endmember signatures in a weighted average to reconstruct the input soil spectra while regressing the spectral signature of SOC. While many approaches exist to perform spectral mixing, the linear case is the simplest, and is a common sub-component of more sophisticated techniques (35). We model the reflectance of a given soil sample $R_i(\lambda_j)$ in terms of the spectra $f_k(\lambda_j)$ of pure endmember samples as:

$$\hat{R}_i(\lambda_j) = \sum_{k=1}^{K} \alpha_{i,k} \cdot f_k(\lambda_j), \qquad \alpha_{i,k} = \frac{\frac{\hat{m}_{i,k}}{\rho_k \cdot r_k}}{\sum_{\ell=1}^{K} \frac{\hat{m}_{i,\ell}}{\rho_\ell \cdot r_\ell}} \tag{2}$$

where $\sum_{k=1}^{K} \alpha_{i,k} = 1$. We note that $\alpha_{i,k} \neq \hat{m}_{i,k}$. Preserving physical interpretability of the model parameters requires a unit conversion factor, since combustion measurements reflect percentages by mass, whereas reflectance measurements relate to the area of the reflective surface attributable to a given endmember. Per dimensional analysis, we perform a weighted average by $\rho_k \cdot r_k$, where ρ_k is the density of endmember k and $r_k \equiv V^{1/3}$ is its characteristic length scale computed from the unit cell volume V. The $M + 1 = 2{,}152$ parameters of this decoder should be interpreted as modeled characteristics of pure SOC: both its spectral fingerprint $f_{SOC}(\lambda_j)$ and its corresponding size factor $(\rho \cdot r)_{SOC}$.

To understand the effect of the assumptions introduced to maintain the physical interpretability of model parameters, we ran the same experiments, but disabling the latent space conversion (equivalent to taking all $\rho_k \cdot r_k \to 1$ or $\alpha_{i,k} \to \hat{m}_{i,k}$). This did not lead to a significant change in performance; it effectively rescaled the regressed values of f_{SOC} by an overall multiplicative factor. Qualitative observations regarding f_{SOC} were likewise unaffected.

C Performance statistics

	Training				Validation			
	Encoder			Decoder	Encoder			Decoder
	RMSE (%)	R^2	Bias (%)	MRMSE (%)	RMSE	R^2	Bias	MRMSE
No decoder	0.827 ± 0.028	0.988 ± 0.001	0.012 ± 0.021	N/A	2.587 ± 0.034	0.887 ± 0.003	-0.053 ± 0.027	N/A
ANN decoder	0.814 ± 0.021	0.989 ± 0.001	0.014 ± 0.024	0.616 ± 0.006	2.589 ± 0.028	0.887 ± 0.002	-0.049 ± 0.031	0.669 ± 0.011
Physics-informed	0.840 ± 0.032	0.988 ± 0.001	0.026 ± 0.024	1.155 ± 0.009	2.579 ± 0.044	0.888 ± 0.003	-0.050 ± 0.030	1.177 ± 0.008

Table 1: Baseline performance statistics when all architectures are trained on 60:40 training-validation splits of the entire RaCA dataset, for 10 fixed random seeds, for comparison against (26).

	Training				Validation			
	Encoder			Decoder	Encoder			Decoder
	RMSE (%)	R^2	Bias (%)	MRMSE (%)	RMSE	R^2	Bias	MRMSE
All regions								
No decoder	0.834 ± 0.008	0.988 ± 0.000	-0.016 ± 0.009	N/A	4.076 ± 0.260	-0.200 ± 0.159	-0.222 ± 0.177	N/A
ANN decoder	0.854 ± 0.008	0.988 ± 0.000	0.011 ± 0.008	0.571 ± 0.004	4.236 ± 0.267	-0.404 ± 0.200	-0.208 ± 0.185	4.123 ± 0.209
Physics-informed	0.865 ± 0.009	0.987 ± 0.000	-0.007 ± 0.008	1.126 ± 0.003	4.132 ± 0.260	-0.306 ± 0.180	-0.129 ± 0.180	4.131 ± 0.207
Problematic regions: 3, 5, 6, 8, 16, 18								
No decoder	0.839 ± 0.009	0.989 ± 0.000	-0.010 ± 0.013	N/A	3.095 ± 0.148	-1.216 ± 0.230	0.638 ± 0.078	N/A
ANN decoder	0.881 ± 0.010	0.987 ± 0.000	0.032 ± 0.014	0.574 ± 0.007	3.346 ± 0.189	-1.637 ± 0.301	0.714 ± 0.087	4.984 ± 0.338
Physics-informed	0.897 ± 0.013	0.987 ± 0.000	-0.015 ± 0.013	1.132 ± 0.004	3.226 ± 0.157	-1.444 ± 0.265	0.765 ± 0.083	4.987 ± 0.342
Generalizing regions: 1, 2, 4, 7, 9-15								
No decoder	0.829 ± 0.013	0.988 ± 0.000	-0.021 ± 0.013	N/A	4.949 ± 0.424	0.704 ± 0.015	-0.987 ± 0.269	N/A
ANN decoder	0.830 ± 0.011	0.988 ± 0.000	-0.007 ± 0.009	0.569 ± 0.005	5.028 ± 0.439	0.691 ± 0.018	-1.027 ± 0.276	3.358 ± 0.179
Physics-informed	0.837 ± 0.011	0.987 ± 0.000	-0.000 ± 0.011	1.121 ± 0.005	4.938 ± 0.431	0.705 ± 0.015	-0.923 ± 0.271	3.371 ± 0.166

Table 2: Compiled results from Experiment 2. Averages are performed over the specified RaCA regions and all 4 random seeds considered. In the regions labeled "problematic," none of the fits converged to a validation $R^2 > 0$.

	Training				Validation			
	Encoder			Decoder	Encoder			Decoder
	RMSE (%)	R^2	Bias (%)	MRMSE (%)	RMSE	R^2	Bias	MRMSE
After pre-training								
No decoder	3.601 ± 0.047	0.776 ± 0.007	-0.075 ± 0.027	N/A	5.037 ± 0.315	-1.443 ± 0.447	-0.070 ± 0.197	N/A
ANN decoder	3.599 ± 0.048	0.776 ± 0.006	-0.001 ± 0.024	0.859 ± 0.005	4.843 ± 0.277	-0.999 ± 0.317	-0.009 ± 0.184	4.254 ± 0.207
Physics-informed	3.635 ± 0.050	0.771 ± 0.007	-0.123 ± 0.026	1.518 ± 0.009	4.791 ± 0.299	-1.040 ± 0.375	-0.051 ± 0.189	4.431 ± 0.206
After fine-tuning								
No decoder	1.715 ± 0.181	0.202 ± 0.310	-0.348 ± 0.077	N/A	4.682 ± 0.369	0.005 ± 0.146	-0.591 ± 0.172	N/A
ANN decoder	1.628 ± 0.178	0.570 ± 0.049	-0.282 ± 0.084	4.494 ± 0.191	4.478 ± 0.352	0.210 ± 0.068	-0.575 ± 0.157	2.250 ± 0.092
Physics-informed	1.561 ± 0.156	0.562 ± 0.063	-0.286 ± 0.056	4.262 ± 0.171	4.665 ± 0.385	0.111 ± 0.101	-0.559 ± 0.180	2.198 ± 0.077

Table 3: Compiled results from Experiment 3. Averages are performed over all 17 RaCA regions and all 4 random seeds considered.