Structured spectral reconstruction for scalable soil organic carbon inference

Evan Coleman[1], Sujay Nair[2], Xinyi Zeng[1,3], **Elsa Olivetti**[1]

[1] MIT Climate & Sustainability Consortium, Massachusetts Institute of Technology, Cambridge, MA 02139
[2] School of Computer Science, Georgia Institute of Technology, Atlanta, GA 30332
[3] Coho Climate Advisors, Bethesda, MD 20814

Abstract

- Sequestering atmospheric carbon into soils via natural accrual processes presents a major opportunity to reverse climate change [1].
- Scalable monitoring of soil properties such as organic carbon (SOC) abundance is *critical for optimizing land management practices* [2].
- Soil carbon *measurements from laboratory analyses are prohibitively expensive* due to soil's significant geostatistical variability [3].
- Common methods to infer SOC from end-to-end regression between lab data and imaging systems *fail to generalize geographically.*
- By training to solve the inverse inference problem simultaneously, we demonstrate a method to signal generalization failures, back out physically-interpretable SOC signatures, and use purely optical measurements to *improve performance scale-up in new geographies*.

Background and Method

Soils and Hyperspectral Imaging (HSI)

- Measuring soil organic carbon costs ~$25-50 USD per sample [3], requires acid treatments and combustion, destroys samples, and emits CO_2.
- HSI is an alternative which measures light at nanometer-scale wavelength resolution (~100x RGB), and may have significant scaling potential.
- Large databases of soil hyperspectra such as RaCA, KSSL, and OSSL have been developed, and are mined for signals of critical soil properties such as SOC [2].
- Current SOC estimates involve end-to-end regression between spectra and lab analyses *(encoder model)*.
- We add in spectral reconstruction as an auxiliary loss *(decoder model)* to analyze and improve inference performance in new geographies.

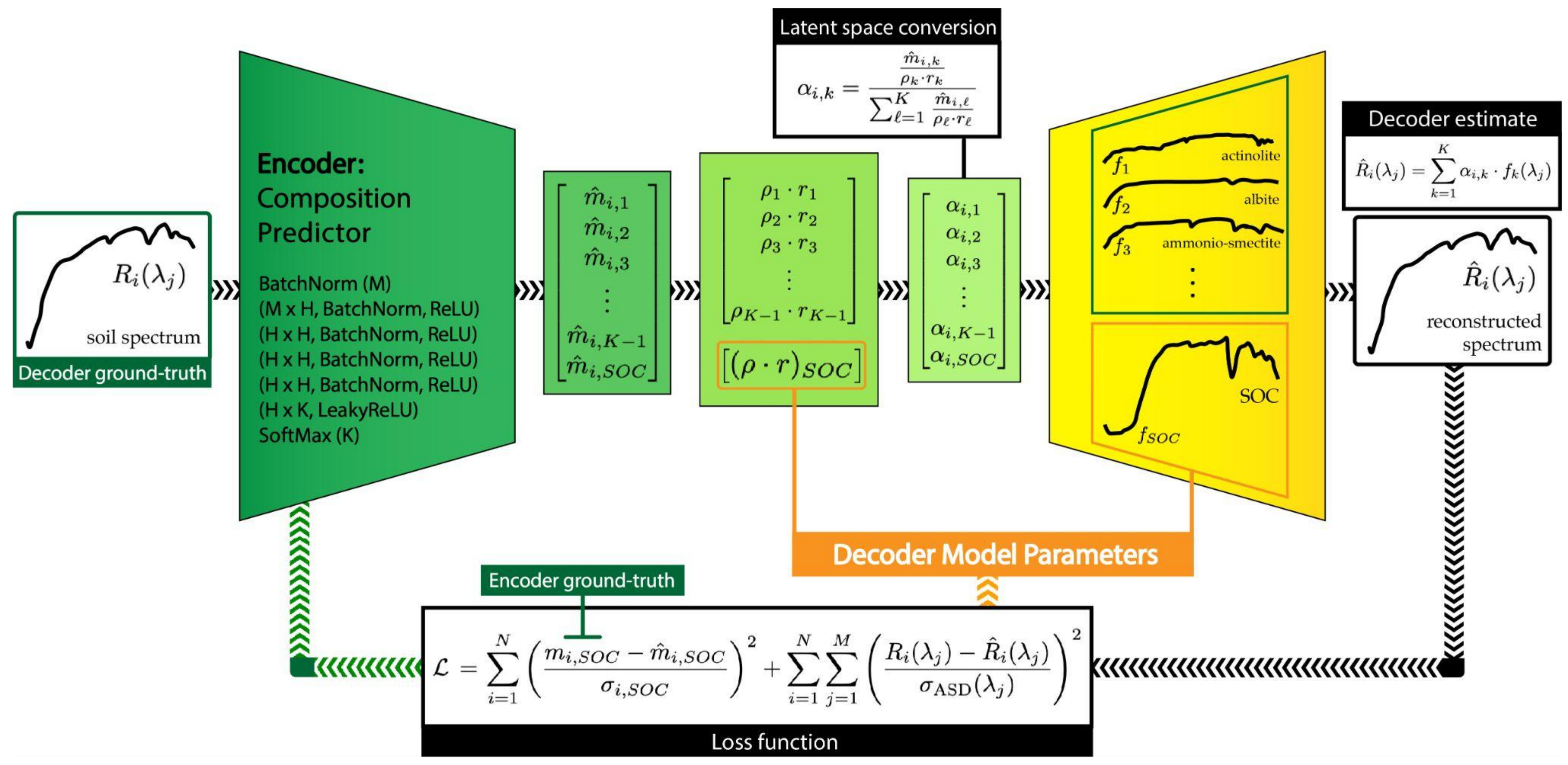


Figure 1: Example autoencoder architecture for Experiment 2. A soil sample's hyperspectrum R(λ) is used to predict the mass fractions m_k of its contents via a multi-layer perceptron. The decoder model reconstructs the input spectrum.

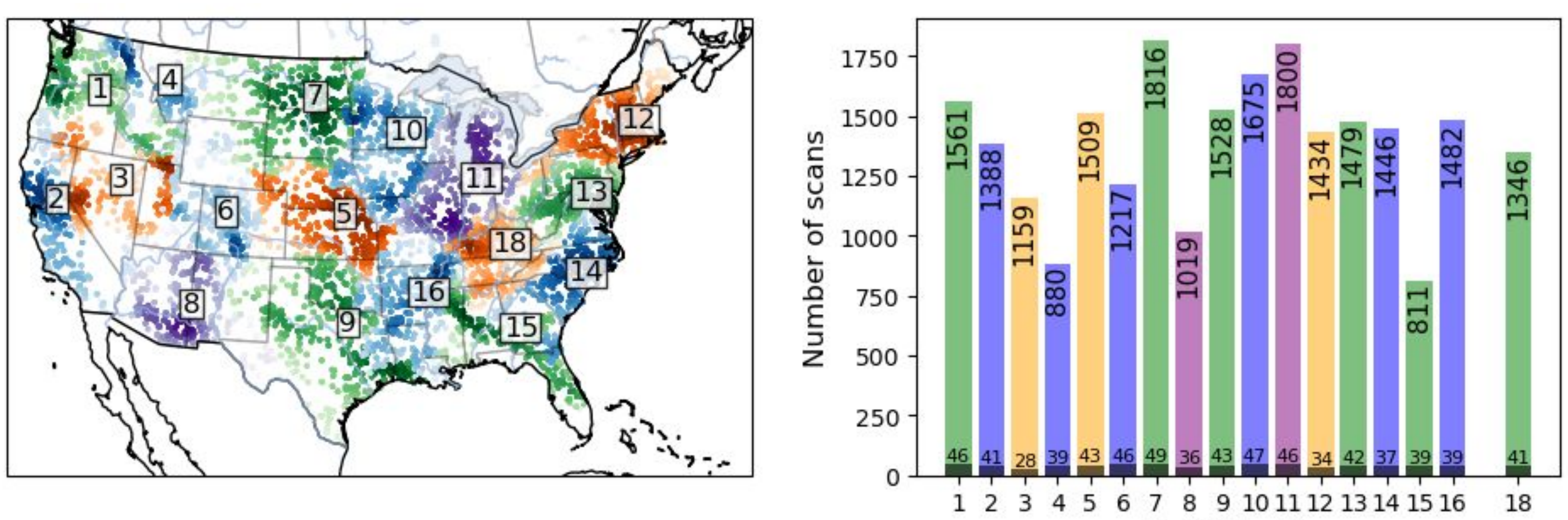


Figure 2: (Left) Geographic distribution of the RaCA data, labeled by RaCA region. (Right) Data available by region. Black bars show example 10-pedon subsamples for Experiment 3.

The USDA RaCA Soil Spectral Library

- Over 20,000 hyperspectral soil scans and laboratory SOC contents collected from across the conterminous United States.
- Each scan contains 2,135 measurements of color data; wavelengths between 365-2,500 nm.
- Soil was air-dried, sieved to <2mm particulate size, and pressed prior to proximal imaging.
- Data collection coordinated to achieve uniform sampling across covariates: split into 17 "RaCA Regions" by MLRA and LULC classifications [2].

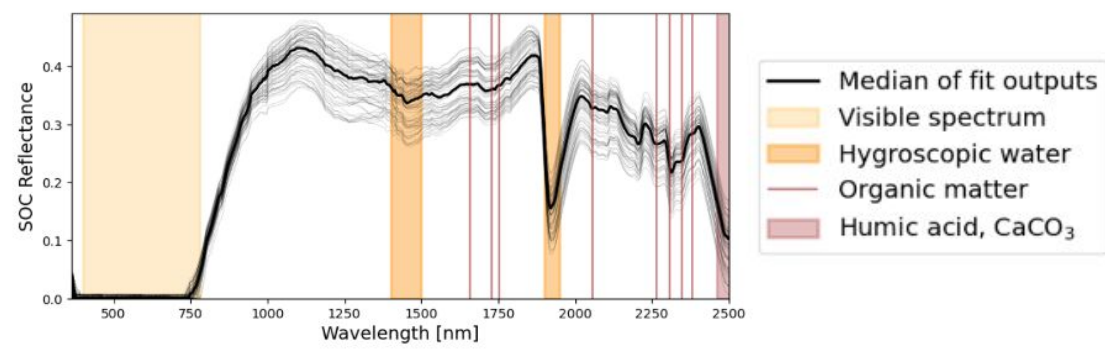
Experiments and Results

1 Reconstruction failure as an OOD signal

- Due to uniform sampling across covariates, RaCA is pre-optimized for leave-one-region-out cross-validation experiments.
- We considered individual RaCA regions as validation regions to study out-of-distribution (OOD) performance of models with and without decoders. The remaining 16 regions formed the training data.
- Evaluated model performance for each RaCA region and 4 random seeds.
- Considered 3 architectures: no decoder, a physics-informed decoder, and an ANN decoder.
- **Result:** For 6/17 RaCA regions, the *R^2 between predicted and ground-truth SOC content was <0*. For remaining 11/17 regions, the R^2 was 0.70 on average.
- No significant differences in performance for models with decoders vs. without decoders. However, decoder model performance significantly decreased for regions with R^2<0 ($p < 0.0001$).
- **Takeaway:** decoder failure *signals SOC measurement generalization failures without using laboratory data*.

2 Extraction of the SOC hyperspectrum

- It is *not currently known how to obtain pure SOC* in the laboratory [4], but its signature may be inferred by training a physics-informed decoder model.
- **Result:** Using a physics-informed linear mixing model, we *back out a reference spectrum* which matches the dark hue and certain reflectance troughs ascribed to SOC content.
- Model is as shown in Figure 1. *Considered 92 separate mineral subcomponents* ("endmembers") of soil and pulled characteristic spectra from the USGS Spectral Library [5].
- Unit conversion factor applied to latent space, to maintain physical interpretability of results.



3 Better performance scale-up via auxiliary loss

- The use of spectral reconstruction permits incorporating training data without laboratory labels.
- Due to the scalability of HSI, *we can collect more hyperspectral scans of soil* than laboratory-based combustion data.
  - RaCA contains >90,000 auxiliary scans of samples which were not analyzed in a lab.
- To understand consequences for OOD generalization failures, we re-performed Experiment 1. We *trained on ~5% of the original combustion data* so that validation R^2 was <0 for all 17 RaCA regions.
- We then *fine-tuned all 3 architectures* on ~2.5% of validation combustion data. Architectures with *decoder models received 100% of HSI data*.
- **Result:** statistically significant relative performance improvement after fine-tuning, for models with decoders.
- **Takeaway:** decoder models can exploit a surplus of HSI data to improve generalization performance OOD. This *presents a candidate approach to scale SOC inference.*

MCSC MIT Climate & Sustainability Consortium

Acknowledgements:
Suraj Nair, Cesar Terrer, Sherrie Wang

References:

[1] 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.
[2] 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.
[3] 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.
[4] Johannes Lehmann and Markus Kleber. The contentious nature of soil organic matter. *Nature*, 528(7580):60–68, 2015.
[5] 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.*