# Author Response RC2 - Multiphysics property prediction from hyperspectral drill core data

Akshay Kamath[1], Samuel Thiele[1], Moritz Kirsch[1], and Richard Gloaguen[1]
[1]Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany.
**Correspondence:** Akshay Kamath (a.kamath@hzdr.de)

We thank the reviewers for their constructive feedback. Please find a full set of responses, including a revised manuscript and record of the corresponding changes with the changed line numbers in the attached pdf file.

Sincerely,
Akshay Kamath and authors

**Reviewer Comment 2**

Dear Mclean Trott,

We thank you for your time and effort reviewing the submitted manuscript, and are pleased that you appreciated our results. We have incorporated your suggestions into the revised manuscript, as detailed in the following pages. Please note that to facilitate the evaluation of our revision, the page and line numbers of the reviewers' comments refer to the originally submitted manuscript while page and line numbers of our responses refer to our revised manuscript.

Kindest regards,
Akshay Kamath, Samuel Thiele, Moritz Kirsch and Richard Gloaguen

**Q) Some food for thought... If I'm using gamma logs, say as input parameters to predict formation, and I also have VNIR-SWIR-MWIR-LWIR data for the same holes and find that it can accurately predict gamma values, why not directly predict lithology?**

**Same logic for sonic logs... If the reason for acquiring sonic logs is to log porosity/permeability, why not directly predict that rather than travel times?**

**This is just a thought exercise, to spur you to think about end-user applications, it in no way invalidates your work.**

We agree with the reviewer that direct prediction of lithology (or mineralogy) makes sense in many situations, however this is not our aim in this case. Specifically, as stated in the introduction:

"In this contribution, we build on this work to test if 'hyperspectral upscaling' workflows that were developed to predict mineralogy (e.g. Thiele et al., 2024) can also be used to map petrophysical properties along drill cores at high (mm scale) spatial resolution. In doing so, we aim to both enhance the spatial resolution of down-hole petrophysical logs and work towards potentially generalisable methods that could, one day, be used to predict important petrophysical and mechanical properties across large drill core libraries and hyperspectral scans of outcrops"

Our aim here is to better understand the links between hyperspectral response and petrophysical properties, as a first step towards predicting more difficult to measure parameters (e.g., porosity, strength, stiffness, etc.). Hence, the derivation of a lithology log is not our end goal. Additionally, we note that our approach effectively super-resolves the borehole petrophysics data (from ~0.1 m to 0.001 m), facilitating more detailed analyses of these widely used data and, interestingly, in some ways mitigating the coregistration uncertainties addressed in the following comments.

This has been clarified by adding the following in the conclusion at L11.12, and L11.23:

"We effectively super-resolve the borehole petrophysics data (i.e. from ~0.1 m resolution of the logger to 0.001 m resolution of the hyperspectral cameras), which helps us explore the intricacies and variations of these properties that cannot be captured by running log measurements. Additionally, our workflow mitigates the coregistration uncertainties that prevent carrying out machine learning workflows over drill core data."

"Our work also serves as a stepping stone toward predicting secondary properties such as porosity and permeability, and provides insights into the links between hyperspectral data and several important rock properties."

We thank the reviewer for their thought provoking question!

**Q) Downhole geophysical tools typically start measuring distance from surface at 0 and measure in a linear fashion downhole based on how much line has been unspooled. Really they have the best depth registration of virtually all the drillhole analysis methods, including core scanning. Core scanning hardware typically registers depth between driller blocks or on a per-box basis. Not as accurate, and depending on the circumstances may be significantly different from the depths provided by wireline geophysical tools. Section 3.3 does not address this issue of co-registration. Or perhaps there is an underlying assumption that the scanned data depths are accurate and correspond to the geophysical depths? Either way this should be addressed or at least acknowledged. It's one of the greatest barriers to performing ML workflows on drillhole data.**

We completely agree that depth estimates will differ between the core boxes (and hence HSI data) and downhole petrophysics logs. This is why we removed the areas (petrophysical edges, defined using the rolling standard deviation) that will be most sensitive to these coregistration uncertainties. The second paragraph of section 3.3 (L4.28) has been clarified as follows to address this:

"Due to factors like core-loss, co-registration errors are expected between the petrophysical logs and the drill-core boxes. To ensure our training dataset does not contain spectra paired with incorrect petrophysical properties, we use the rolling standard deviation to eliminate points from regions of high property variance, as these will be highly sensitive to coregistration uncertainties. Hence the underlying assumption within our preprocessing steps is that by picking points only from petrophysically homogeneous regions of the drill cores, we can partially mitigate challenges caused by co-registration errors."

**Q) You've used HBSCAN to cluster the data and further identify noise (class -1) which you've**

**removed from the dataset. You've mentioned resiliency to hyperparameter selection– it is actually very well established that HDBSCAN outcomes are highly sensitive to hyperparameter selection, particularly the distance metric, min_samples, and min_clusters parameters. sklearn can automate hyperparameter selection using Randomized Search Cross Validation, which seeks to optimize the validity index for iterated hyperparameters, otherwise hyperparameter tuning is highly manual and hugely impacts the number of clusters and cluster distributions returned. I'd strongly suggest addressing this.**

We agree that the HDBSCAN algorithm is sensitive to hyper-parameter selection; so this statement has been removed (L5.29). In our case we did not optimize these hyperparameters as we're aiming to over-segment our dataset (i.e. classify into more classes than exist in the data), for the sole purpose of performing stratified random sampling. Hence, unlike classification tasks where the output classes have more meaning, our approach is less sensitive to hyperparameter selection. Instead, we tweaked the hyperparameters until we were satisfied that we had over-segmented the dataset, based on visual inspection of the clustering results (Fig. 3C). We have clarified this in the Data-Balancing section, at L6.1:

"Hyperparameters of HDBSCAN were selected manually, and iteratively assessed based on the number and separation between the clusters shown in Fig. 3. We aimed to over-segment the dataset, as the clustering was solely used to help in removing the inherent bias from a larger number of points belonging to one lithology. Any cluster distribution with over-segmented clusters (i.e., the large clusters separated from the rest) would provide similar results during training."

As for the second point, we used the HDBSCAN clusters to do the stratified sampling; this has been clarified in the document now at L6.6:

"…using the labels from the HDBSCAN clusters…"