<u>Response to Reviewer #2</u> - Richard Ketcham

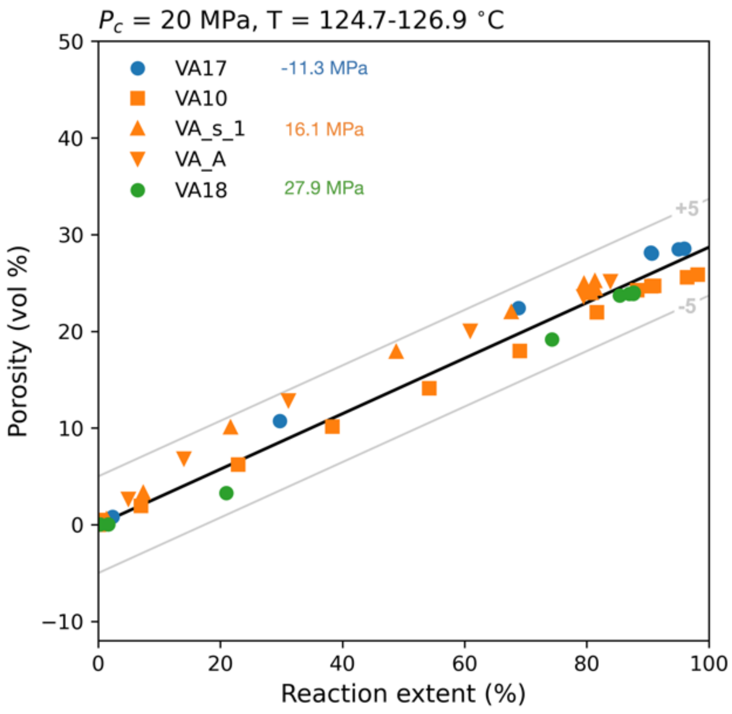
This is a very nice study, using knowledge of the physical system as a means of benchmarking machine learning approaches to segmentation. It will certainly be of interest to the community, and is close to being ready to publish as-is. There are only a few items that might need a little more work.

We thank Prof. Richard Ketcham for the comments and suggestions on our manuscript. We are pleased to read that he found this work valuable contribution. We hope our response to the comments addresses the issues raised.

Overall this method seems to work quite well on this system. However, it relies on manual segmentation to create the training data. This works fine for the more massive phases, but for the ones that have one or more small dimensions relative to the data (celestite, and more importantly water-filled porosity) many of the voxels suffer partial-volume or blurring effects. The segmentation of the porosity in Fig. 3e is very chunky in comparison to how the porosity really is (thin grain boundary layers). Basically, everything with a hint of darkness is being called porosity, but many of those voxels represent mixtures between water and solid material. This makes the method a bit less repeatable, and is likely why the water always comes out a bit high in Fig. 7. This issue is recognized by the authors in lines 341-345, but not given context on how it is affecting the analysis presented.

We acknowledge the limitations of our segmentation method. The manuscript demonstrates the precision of segmentation by comparing it with theoretical molar volumes (please refer to Figure 7). It is shown that the deviation from theoretical values for porosity—representing water in the pore space—falls within a 95% accuracy threshold (as evidenced by the proximity of the data points to the theoretical lines in Figure 7a).

To maintain focus and clarity in the paper, we have presented the accuracy analysis for only one time-series volume. However, in a recently accepted publication in Geology (Gilgannon et al., 2023), we have demonstrated the consistent accuracy across all volume data. The accompanying image below illustrates the extracted porosity from various experiments (represented by dots, triangles, diamonds, and squares) alongside the theoretical curve for the dehydration reaction



Also, the calibration of CT values does not necessarily need to rely on external calibration – you can probably just measure the grayscales off your images (and use the volume balance in your system to check that your results are about right). Since this would essentially be a post-processing step reinterpreting some segmented voxels, it seems a complementary add-on.

We are grateful to Prof. Ketcham for his insightful comment on calibrating μCT values. The recommendation to derive grayscale values directly from our μCT images and utilise the volume balance of our system for validation is well-received. We will explore how to incorporate this method into our existing procedures, mindful of our dataset's unique characteristics and constraints. While implementing this technique may pose challenges, particularly in matching grayscale values to specific phases or compositions, it represents a promising direction for advancing and fine-tuning our segmentation approach.

Detailed comments:

[line 72] Replace “results” with “provides”

Modified as suggested.

[lines 163-165] This sentence is garbled (“histogram thresholding… or… histogram thresholding”?). Rewrite.

This issue was also raised by Reviewer 1. We have, therefore, revised the sentence as follows: "To establish the ground truth set for initial image classification, we initially considered methods with differing informational depths. One such method is histogram thresholding, which relies on basic greyscale values and typically results in low-information-level outcomes. However, this approach alone proved inadequate for our purpose, as it often led to gradients with diffuse phase boundaries, underscoring the need for more sophisticated classification methods."

[Table 1] It seems odd that the average DICE for Model E is higher than that for Model RF even though the former scores worse and much worse on two phases. How was the average calculated?

The average is derived from only those classes that obtain a DICE score. To provide a clearer picture, as average DICE values may not fully represent individual class performance, we present the DICE scores for each segmented class in the table.

[line 314] What does "(imaged as porosity [in] the µCT data)" mean? Was the water missing/drained, or is the water just being called porosity because it's pore-filling?

A portion of the water released during the reaction was retained in the porosity formed from the volumetric reduction when gypsum dehydrated to bassanite, while the surplus was expelled. Given that the fluid pressure (Pf) applied surpassed the vapor pressure of water at the reaction temperature, it is likely that all imaged porosity was water-saturated.

[line 331] Another issue with ML-segmentation is that the training may only work well for very similar conditions of material, geometry, imaging parameters, etc.

We fully agree. This is the reason why we implemented a Deep Leaning (DL) model on top of the ML Random Forest. The DL is able to generalise a segmentation model better than ML, particularly when DL is integrated with Data Augmentation.

[line 344-345] Another approach might be re-interpretation of voxels using grayscale, so as to assign affected voxels partial values.

We have now included this into the Discussion session as follows: "Solutions have been proposed and often require complementary techniques (such as using tactile, optical sensors) to calibrate measurements derived from CT data (Torralba, 2018). Reinterpreting segmented voxels using grayscale values can offer a complementary method for calibration. This approach assigns partial values to affected voxels, potentially enhancing accuracy in cases of overlapping mineral phases or partial volume effects."

[line 377] “In future iterations of our method aims to expand…”: change to “we aim” or “Future iterations of our method will aim”

Thanks for the suggestion. Now: “Future iterations of our method will aim to expand its capabilities and applications”