Hybrid Domain-Transfer Approach for Resolution Increase within Data-Constrained Paleoclimatic Proxy Records

William J. Hoffmann

Department of Computer Science

Northwestern University

Supervisors

Zach Wood-Doughty Ph.D. , Jonathan Raberg Ph.D.

In partial fulfillment of the Honors requirements for the degree of

Bachelor of Arts in Computer Science

April 15, 2024

Abstract

In this paper, we use a Hybrid ViT-CNN model to predict proxy values from sediment core records in two Icelandic lakes, Stóra Viðarvatn, SVID, and Litla Viðarvatn, LVID. We predict the values of paleoclimate proxy variables throughout the sediment record in order to create a tool to impute proxy values and increase the resolution of proxy records at hand. We were limited from a few dozen to a few hundred examples (fewer than 200) for each proxy value. This constraint led us to implement Facebook's DeiT model as our embedding model for a 3-layer CNN trained specifically on the task at hand. The model was effective at imputation, yet struggled at out-of-domain prediction. We present our results and potential next steps for going beyond our limitations and improving the predictive power of our model.

Contents

1 Introduction

In the past few years, machine learning has made inroads in many fields of study. Machine learning is attractive for a multitude of reasons. The increasing skill and complexity of modern modeling techniques allow for increased power in context-dependent situations as well as the extraction of more general and low-level features. It has helped revolutionize medical imagery and modern-day business. In both fields, the biggest strides of machine learning have come in the form of the automation of existing tasks. With infinite time and resources, machine learning becomes redundant; however, in our resource-constrained world, automation has helped create a balance between precision and speed previously unattainable with only human labor. In the medical world, models that can process images quickly and accurately have led to faster diagnoses and an expansion of the amount of useful data. By leveraging machine learning, diagnoses that would have previously snuck through the cracks, even common ailments such as tuberculosis and complications of childbirth, can be predicted and handled (16,17). In the business world, increased automation allows for a substitution for high-priced consulting work (18) and increased work output (19,20). As companies, labs, and industries continue to invest and implement machine learning and artificial intelligence (21), the central questions become: 1.) How can we maximize the resources we do have, and 2.) What are the limits of machine learning's usefulness?

In regards to exploring the limits of machine learning, there are still areas of science and industry more adversarial to automation and machine learning that can act as a proving ground for today's computing techniques. The reasons for hesitation are plentiful. In some cases, such as in the service sector or intense surgeries, people prefer to

interact with humans, and humans can react to fast-changing environments or difficult personalities (24,25). In other cases, domain restrictions mean that the context of the problem becomes just as important as the source data (26). It is easy to imagine a world where some problems are unsolvable; however, recent strides in mathematics and machine learning have provided unique solutions and pushed the limits of our understanding of heavily qualitative and random problems (96,97,98).

To get one step closer to finding the limits of machine learning, we chose paleoclimate as a proving ground in the currently underexplored territory of geoscience. Machine learning has seen limited adoption in geoscience due to the field's reliance on strong physical theories that guide understanding of complex natural processes, which can be difficult to encapsulate within purely data-driven models (28,29). Furthermore, geoscience data often exhibit a high degree of randomness due to the variance of earth processes and require a deep contextual understanding to analyze, making straightforward application of machine learning challenging (27). These challenges require tailored approaches that blend domain expertise with machine learning techniques, complicating their broad implementation. Regardless of the current progress, the possibility and opportunity for this to change is not only possible but realistic (5).

Within our constrained research environment, we wanted to explore the impact we could have with the time, resources, and data at our disposal. A very promising recent approach for a multitude of problems has involved domain transfer. This technique is particularly beneficial in enhancing generalization capabilities and efficiency (22,23). Domain transfer in machine learning refers to the process of applying knowledge gained from solving one problem in the source domain to a different problem in a separate target

domain. This is useful because it allows for the leveraging of pre-existing models and data to improve performance on new tasks where labeled data might be scarce or expensive to obtain. When time and data are limited, domain transfer allows for a previously unattainable step-up in modeling power by combining an expensive pre-trained model with a smaller, more specific, and contextual model.

Geoscience is not only interesting for the possibility of exploration but also for the impact the work could have. As human activity increasingly affects the Earth, understanding our world – and its extensive past – becomes even more important. Specifically within the paleoclimate domain of geoscience, the study of past climate history, we can help to understand and tackle the planetary warming we are experiencing today due to anthropogenic climate change (30).

Past environmental conditions (e.g., temperature, precipitation) can only be reconstructed through direct measurements for the past few hundred years. Beyond the end of these instrumental records, geoscientists use geochemical measurements from sedimentary archives as proxies for these environmental parameters. Sediments from the lake and marine environments provide one such archive of past environments, and many geochemical proxies have been developed to reconstruct parameters such as temperature, precipitation, and terrestrial and aquatic ecosystem productivity and composition.

Although extremely valuable, these proxies are expensive and time-consuming to measure by hand. The core collection process can include manual drilling in potentially remote areas, transport across continents, cumbersome storage, and manual measurement; therefore, precise analysis can be limited by the resolution of the data rather than the sheer amount of data itself (5,31). This means that valuable datasets and analyses can be

left on the sideline due to the effort and resources required for the collection and labeling of these sedimentary archives. Rather than replace the human component, we set out to augment the work that is and can be done by solely human contributors by increasing the proxy resolution of these sedimentary cores. A higher resolution dataset allows for richer analysis across time. The Earth system functions on several different timescales. Orbital dynamics cause changes in Earth's climate on the scale of thousands to 100s of thousands of years whereas climate oscillations, such as the El Niño Southern Oscillation (ENSO), operate on annual to decadal timescales. The sediment cores represent about 10,000 years of history, so with 10 samples, one can look at millennial-scale processes, whereas with 1000, one could look at decadal-scale changes.

Not only can geoscience analysis be improved through machine learning but machine learning can be improved by expanding work within the geoscience domain. The process of geoscience research differs from pure computer science research. If scientific problems are only pursued from a computer science perspective, success is bound to be limited (32,33,34). By better understanding what matters to questions in geoscience, computer scientists and machine learning engineers can better implement effective models for the task at hand. Informing models with theory has already had exponential benefits in creating global maps of surface water dynamics (35,36), modeling lake temperature (37), and creating a deeper understanding of climate patterns and relationships (38). Although important in computer science work, explainability is paramount in a hard science like paleoclimatology. Working in such a domain forces the scientist to consider data preparation from more than just a format a neural network can understand but also in a context the broader geoscience community can find helpful and

sensible. Understanding the limits and context-dependent nature of modeling can improve and expedite the process. By informing models with necessary information pre-facto and considering post-facto explanations for the broader earth science community, we ensure the efficacy and clarity of any solution (39).

Regarding the first question of resource maximization stated in the first paragraph, we are limited by both the data at hand and the learning ability of a pre-trained model on a completely different dataset. In a perfect world, whether trained on retail data or sedimentary cores, we would see model performance transfer seamlessly between the domains. The features extracted would be constant across the two datasets, and any sufficiently complex model would lead to a relatively cheap and effective solution. There would be less value in manually increasing the resolution of proxy data if the current resolutions are sufficient for machine learning. Likewise, rather than putting forward the effort and data necessary to create a domain-specific model with millions or billions of parameters, we could use one of the many current available models in the public domain for our task of feature extraction (99,100). At the other end of the spectrum, the disparity between the two datasets and the limitations of even the most powerful and modern models could lead to learning that doesn't transfer. There is a world in which the addition of a completely different model proves to be more harmful than helpful whether due to covariate shift, dataset bias, or another related reason (80, 81, 82).

In comparison, measuring the usefulness of a model is much more straightforward. If the preprocessing model did help and the data was sufficient, we would expect to build a model that could be useful for both the direct task at hand and for other problems within different datasets. At the upper bound of our proposed solution,

we would expect the model to perfectly distinguish between the two lakes and recognize different unseen lakes or samples from other, similar domains such as marine sediment (79). This solution would allow for seamless imputation and even the potential to predict unseen periods within a lake and in other lakes. Imperfection, however, would involve a model confused rather than aided by the additional data. Predictions would be spotty and unreliable over the entire record. Even in this world, we would be able to say that this domain provides limits to the usefulness of machine learning within our proposed solution.

Nonetheless, we hope to add to the geoscience domain by providing modeling techniques that could be applied beyond the sedimentary lake cores we decided to focus on. At the same time, we consider standard paleoclimate approaches and implement these approaches in our pursuit of the stated task. In this paper, we use a hybrid domain transfer learning approach to increase the proxy resolution of lake sedimentary cores in two Icelandic lakes. By applying a prominent pre-trained model to create an embedding, we test the power of domain transfer in a far different domain. We challenge the learning ability of our model by testing its skill at predicting unseen 1000-year chunks and verifying its ability to distinguish between lakes when training on two different lakes simultaneously.

2 Related Work

Image-related work in machine learning has been of interest for many years. Only in the past few years have significant strides been made. The standard process in predictive machine learning involves learning the link between a set of data and a variable of interest. The set of data comes in the form of features that are used as inputs to the model.

For each set of data, the model learns the relationship through the mathematical connection between the data and the target, whether through perceptrons, least-squares regression, or your favorite technique. The first challenge for image data came in the form of the abundance of features within an image. Even a 2 MB image contains 10,500 pixels, so for every labeled 2 MB image, the model needs to take in 10,500 features. Although noise is present in tabular data as well, each image provides a huge range of variation, including but not limited to lighting, placement of objects, scale, and orientation. Before being able to handle such complex inputs, computational resources had to catch up to handle such large and complex data relationships.

The first big image breakthrough was the Convolutional Neural Network (CNN) model (41). Although first introduced in the 1980s, both GPUs and parallel computing have been instrumental to the current ubiquitousness of image-based models (40,42). The CNN builds off the artificial neural network architecture first coined in 1958 by building off of even earlier mathematical concepts (43,47). The cornerstones of the neural network models are backpropagation and gradient descent (44). By sending the prediction error back through the model, iteration and ample complexity allow for learning. These same concepts can be extended to images by using convolutional and pooling layers. Convolutional layers extract the features from the image by taking into account both the location and contents within an input. The layers use a combination of weights, filters, stride, and padding to accomplish this. The weights are trained using backpropagation and gradient descent, in much the same way as a classic neural network. The filter construction determines the number of feature maps extracted from the input, the stride determines the distance moved by the kernel over the input, and padding is used to line

up the input size with the filter size. Pooling layers serve to reduce the dimensionality of an input. By taking either the average or max of subsections of the image, the input loses granularity but reduces complexity and improves efficiency (48). The first CNNs were limited by the size of the models and the amount of data a single model could handle (45,46).

Once computing power caught up to the potential of CNNs, the ImageNet and MNIST datasets provided large and labeled repositories for models to build off of. The ImageNet dataset was conceived in 2009 and included 3.2 million labeled images at its inception. It continues to grow and serve as a backbone for comparing modeling performance (49,50). The MNIST dataset is a collection of 70,000 handwritten images of digits written by high school students and the United States Census Bureau assembled in 1994 (51,52). These have served as benchmarks across the industry, and model performance has improved dramatically since their inception due to an increase in the complexity of the CNN models. In the case of ImageNet classification, AlexNet achieved a score of more than 10.8 percentage points lower than the runner-up (53), and the current best CNN has far surpassed that (54).

More recently, the newest player in the game has been the Vision Image Transformer (ViT). While using the same datasets as benchworks, ViTs have outperformed the best traditional CNN models by 1% with ImageNet (59) and achieved comparable performance on the easier MNIST dataset (56). The ViT model builds off of the text-based Transformer model (1,2). The transformer model is the backbone for the Large Language Models (LLM) we see today in the form of ChatGPT or Google Gemini (57,58). Transformers are an extension of neural networks that compute each word of a

sequence in parallel and use self-attention mechanisms to maintain the context and importance of each word in the input as well as positional encodings to note the relative position of each word within the input. ViTs take these ideas and apply them to image inputs. The ViT can consider both context and position within an image in much the same way a classic transformer works with textual input. It works by breaking down an image into smaller patches, treating each patch like a word in a sentence. These patches are then converted into a sequence of embeddings (numerical representations), similar to how words are represented in natural language processing. The transformer model processes these embeddings using self-attention mechanisms to understand the relationship between different patches. Self-attention mechanisms in models calculate the relevance of each part of the input data to every other part, enabling it to learn contextual information about various parts of the image (1). This process allows the ViT to capture both local and global features of the image.

To maximize the strengths of both the ViT and the CNN, strides have been made by combining the two models within the ImageNet domain (55), crop yield prediction (4), and medical image registration (3). The ViT excels at keeping track of global context and relation understanding, whereas the CNN's power comes from its ability in local feature extraction and translational invariance, in that the position of an object within an image does not influence a CNN's ability to detect it. Feeding the raw image into a ViT allows for the creation of an embedding capturing more of the global and relational features. Then, feeding the embedding into a CNN allows for a more local understanding of the problem through the embedded features leading to a powerful hybrid prediction.

Even while using ViTs trained on a completely different dataset, the nature and skill of a ViT's feature-extracting ability allow it to still shine in disparate areas (60,61,62).

Image-related work in machine learning has been of interest due to its value in other domains. Whether in the form of satellite images, geological maps, or sediment cores, physical image data within geoscience is paramount to understanding the world around us. One method for expanding paleoclimate understanding is to reconstruct climatic conditions during warm periods in Earth's past. This paleoclimate approach is highly valuable, both for providing an understanding of Earth's geologic history and for generating targets for climate model simulations (63).

As mentioned in the introduction, when environmental parameters such as temperature become immeasurable, correlated proxy data is the next best alternative. Proxy data has had great success in helping us understand the changes of the past and the effects of environmental variation in paleoclimatology (64,65,66). Many of these proxy measurements are time-consuming and/or expensive, and new methods are being developed to maximize data and more easily decrease dataset limitations through proxy-to-proxy calibrations (67), cross-domain feature learning (83), or dataset augmentation (84). The amount of samples to process can be so large that many insights are yet to be made from samples in storage (87). Through tabular data, machine learning has already proved helpful in the detection of tree ring width proxies (8) and brGDGT proxy values (12). In recent years, image-related studies employing machine learning techniques have helped alleviate this burden by allowing researchers to automatically detect varves – annual sediment layers – with Deep Varve Net (7) and detect tree rings (68) with high accuracy.

Complex machine learning, however, is not always the answer. As in any domain, there are places where the simpler solution outperforms the more complex one. In geoscience, complex deep learning models proved to be less successful than a simple logistic model in predicting aftershock prediction (85, 86). In the same vein, the process needs to be airtight to be accepted in the geoscience community. Physical laws need to be respected and domain knowledge needs to be incorporated. This becomes difficult when there are complex problems in geoscience with non-unique explanations, and domain knowledge can prove to be hazy and contradictory (88). At the same time, a parallel pressure comes from the machine learning side. Data leakage needs to be prevented and sufficient tools need to be implemented to prevent overfitting to the source data (89). In order to respect both the norms in computing and geoscience, a balance needs to be struck between speed and progress. If proper checks and balances are followed, however, machine learning has the opportunity to go beyond theory and allow for greater exploration within the entire field of geoscience. (87).

We hope to build off these works, apply the results gleaned to our experimental process, and use computer vision to go one step further beyond object detection and predict proxy values directly from the images.

3 Methods

This section contains a description of the data collection process and the decisions made when creating our model.

3.1 Site & Geochemical Proxy Description

The two lakes of focus consist of Stóra Viðarvatn, SVID, and Litla Viðarvatn, LVID. The lakes are positioned less than a kilometer apart, meaning that they've experienced the same climate history, in the Northeast corner of Iceland. However, they differ substantially in their size and limnology; SVID is a large (2.51 km^2), deep (48 m) lake with a fully oxygenated water column while LVID is small (0.21 km^2), shallow (2.5 m), and contains oxygen-depleted waters in the winter (69). Comparisons of proxy reconstructions from the sedimentary archives of these two lakes can thus aid in proxy development by increasing our understanding of the modern links between environmental complexities and proxy measurements.

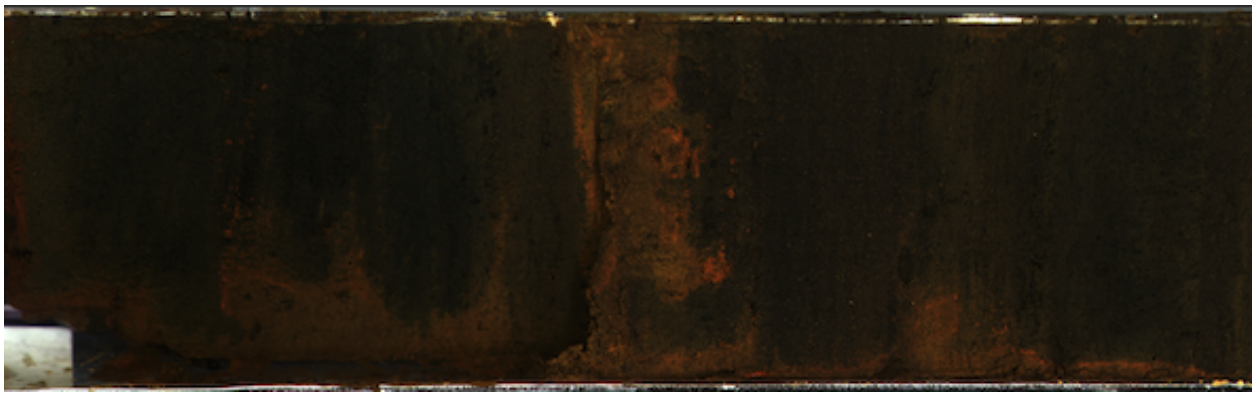


Figure 1: LVID core image over the latest 400 years (20 cm) of sediment

Figure 2: SVID core image over the latest 400 years (20 cm) of sediment

Measurements of the geochemical proxies of %TOC, brGDGT, BSI, and C/N were taken every one to sixteen centimeters.%TOC, total organic content, serves as an indicator of past biological productivity and environmental conditions. High levels of %TOC indicate periods of increased biological activity. This offers insights into climate conditions such as temperature and moisture levels (13). The proxy of brGDGT, branched glycerol dialkyl glycerol tetraethers, is a class of membrane-spanning lipids found in bacteria (70). The distribution of brGDGTs changes with temperature, so brGDGTs are an effective proxy for temperature (71). Work has been done on a global scale to verify brGDGT as a temperature proxy, and it is effective in its use for inference of historical climate conditions (6,12), including in Icelandic lake sediments (72,6). Throughout the paper, we use brGDGT interchangeably with the variable MBT, which stands for methylation of branched tetraethers. MBT captures changes in the number of methyl groups on the brGDGTs, which is correlated with temperature (71). Although powerful, MBT is the most difficult proxy to measure; therefore, the resolution is much lower in our dataset as compared to the other three proxies. We use an updated version of MBT called MBT'_{5Me} (101) that more accurately relates to temperature, but refer to it as MBT for simplicity. Biogenic Silica, BSI, is derived from the remains of siliceous organisms, which incorporate silica from the lake water into their cell walls. High BSI levels in sediment reflect historical levels of primary productivity. By using BSI levels, nutrient availability, water temperature, light conditions, and general environmental health of past climate conditions can be inferred (14,73). Carbon/Nitrogen ratio, C/N, captures the relationship between terrestrial versus aquatic sources of organic matter in sedimentary records. A high ratio indicates a higher amount of terrestrial life and plant material,

whereas a low ratio indicates aquatic algae or phytoplankton. C/N helps to assess historical changes in vegetation, erosion, and nutrient cycling (15). The proxies are all microscopic or molecular and are not visible directly in the core images; however, the sediment records capture an integrated signal of the environmental conditions that influence each of the proxies, so there is potential for prediction from the image data alone.

3.2 Datasets & Preparation

In this study, we utilize two types of data for proxy predictions: i) Tabular proxy data labeled by the ILLUME Project Team (see Acknowledgements) and ii) High-Resolution Core image data generated at the Continental Scientific Drilling facility at the University of Minnesota.

3.2.1 Data Description

There were two twin datasets for both LVID and SVID. As LVID is the shallower of the two lakes, the corresponding image and proxy data record spanned only 724.5 centimeters as opposed to the 902-centimeter-long core of SVID. The data was irregularly sampled throughout each core with anywhere from one to sixteen centimeters between each measurement. There were a total of 182 labeled proxy measurements for %TOC, BSI, and C/N, as opposed to only 84 labeled measurements for the more costly brGDGT proxy within SVID. In LVID, there were fewer examples across the board, as expected. There were a total of 55 labeled brGDGT examples, 151 labeled examples for %TOC and C/N, and 144 labeled depths of BSI. Manual preprocessing of the tabular proxy datasets was minimal beyond the compilation of the data in one place. There was little missingness and no data input errors that we were aware of. The range of the data is as follows: 0.018 to 10.7198 for %TOC, 0.0544 to 0.2964 for MBT, 1.301 to 11.4722 for C/N, and

20.682 to 158.874 for BSi. This complete dataset is still in preparation and has not been published yet, so although our analysis was done on the entire dataset, we provide only SVID %TOC data as a reference.

Figure 3: First 100 centimeters of core data for SVID

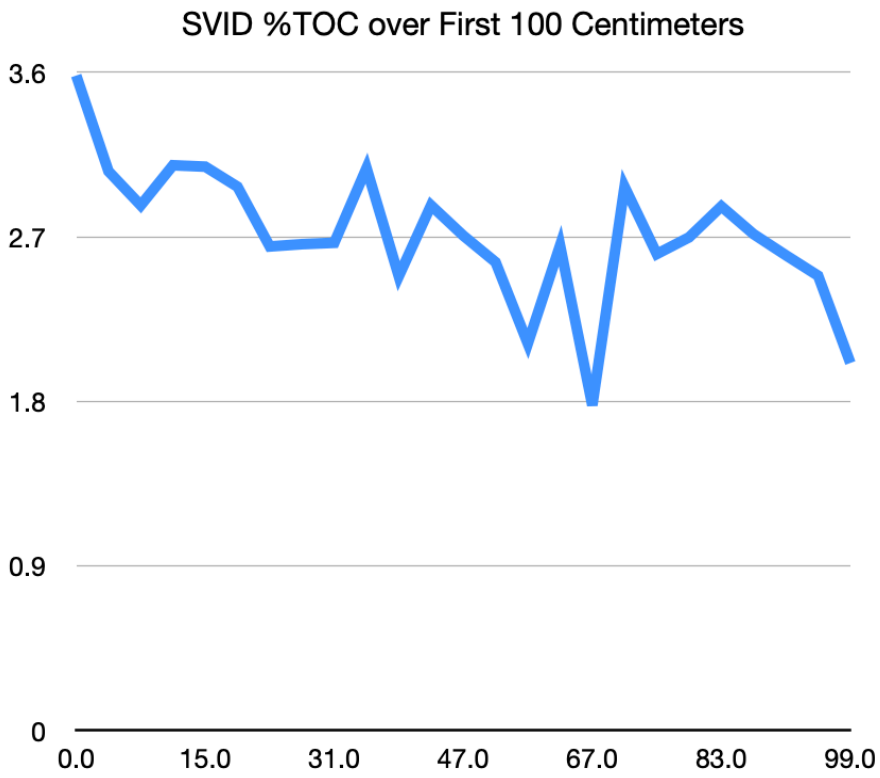


Figure 4: SVID %TOC values over most recent ~1000 years (100 cm) of sediment data, beginning at the 16 cm mark in Figure 3

3.2.2 Image Processing Methods

To create a coherent dataset, we utilized the Python Pillow package. Each core was comprised of multiple segments of images. We first cropped excess reference data from

each image. This included removing the depth markers at the bottom, the packing materials at the beginning, and the color reference data at the back end of each core. To have complete measurements and prevent data loss, we appended each core image together to create a complete record of the core. Then, we went through the process of segmenting the complete core into one or five-centimeter chunks and finally assigned the geochemical variables of interest to each image chunk. More details on the creation of the individual chunks are presented in sections 3.4.1 and 3.4.3.

Figure 5: First 100 Centimeters of SVID sediment after cropping

3.3 Modeling Decisions

Our model consisted of two parts: One being a preprocessor ViT model and the second being a CNN model trained specifically for our task at hand.

3.3.1 Performance and Evaluation

Evaluation of the model was based on root mean squared error (RMSE) values. We chose RMSE to account for the variance in our data. Due to the high variance in the dataset, we wanted to make sure that the best model was not only able to capture the general trend but also the intricacies within the data. RMSE is also a convenient evaluation metric due to values being on the same scale as the target data. Mean squared error was used as the loss function for training the model throughout the process due to the same reasons. There was no need to square root the loss values as they are only used for training, and relative comparisons between train and test loss were the extent of their use for

evaluation. Loss comparisons allowed us to monitor our model's learning ability and verify learning rather than only memorizing the training set. Although test loss being higher than train loss can be expected, a simple check for overfitting was to examine the training loss and test losses and verify that the divergence wasn't so great between the two sets that it seemed the model was merely regurgitating train values for test examples.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Figure 6: RMSE Calculation for all N examples

In addition to quantitative evaluation, we conducted a qualitative evaluation of our model's performance by comparing our model's performance to human domain knowledge. The experiment is laid out in section 3.4.5.

3.3.2 Preprocessor ViT

The ViT architecture we decided to use was a Data-efficient Image Transformer (DeiT) model trained on 1 million ImageNet images and 1000 classes (9). The complete model includes 5 million parameters trained for two to three days on an 8-GPU node. The corresponding performance on the ImageNet dataset is competitive with other larger deep learning frameworks. The DeiT incorporates a standard class token as well as a distillation token. Distillation allows the model to learn from a "teacher model" and increase model efficiency and performance (74,75). The model consists of alternating layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks with LayerNorm applied to the output of every block and a GeLu activation function

following every MLP block. Although the task trained on was far from our intended domain, this specific model has had success in other domain transfer tasks in both psychophysics (10) and active speaker detection (11). We extracted the initial feature extractor and the first two layers of the ViT for our preprocessor model. Roughly 1 million parameters were included in the Preprocessor ViT. The ViT model was not trained any further than it already was, and the focus was instead on training the CNN. Using the ViT model allowed for a decrease in the size of our CNN inputs while still having confidence in the feature representation of the embedding, leading to both increased performance and efficiency

3.3.3 CNN

PyTorch functionality was used to create a CNN model. The model was designed to take in the ViT embeddings and predict the value of a target geochemical variable based on the image chunk. The final CNN architecture we chose had 812,000 trainable parameters. It consisted of three convolutional layers, with output dimensions of 256, 128, and 64, respectively. Each is followed by a batch normalization layer and a Sigmoid activation function. Although specific for any given dataset and task, rather than finding our optimal kernel size, as can be done through modeling (95), we chose a smaller and odd kernel size of 3. A kernel of 3x3 allows for some broader context in each convolution, while not being too large at the expense of granularity. In addition, odd-numbered kernels allow for the preservation of symmetry around the center of the image (76). A padding value of 1 pixel was used to allow the convolution operation to be applied right to the edges, ensuring no information was lost at the borders.

The convolutional layers were then followed by a dropout layer with a 10% dropout rate, an adaptive pooling layer, and a final fully-connected linear layer with 64 nodes which took in the embedding to create our continuous prediction. We considered dropout rates as high as 50%; however, a lower value around 10% was found to be much more effective. A more in-depth discussion of hyperparameter decisions can be found in the following section, 3.3.3.1.

3.3.3.1 Grid Search

To use a model with appropriate complexity and ability to learn we ran a grid search over a collection of hyperparameters: layers and nodes, learning rate, epochs, activation function, dropout, learning rate decay, and batch normalization. We used the experimental design laid out in section 3.4.1. Over the course of a few weeks, we trained a total of 2055 models over different combinations of lakes and proxy variables to discover the best set of parameters.

The proxy variable %TOC within SVID was found to have the least amount of variance among the four geochemical proxy variables. This became our reference point for finding a stable and effective set of parameters to do the majority of our analysis. In addition to %TOC within SVID, we performed similar grid searches with the combined dataset of the two lakes to examine the optimal amount of epochs and whether to use learning rate decay for training.

	Learning Rate									
Values	1.00E-08	5.00E-08	1.00E-07	5.00E-07	5.00E-06	1.00E-06	0.0005	0.0001	0.005	0.001
RMSE Mean	2.08812	1.7669-	1.81753	1.81630	1.41994	2.04223	1.86488	1.91317	1.87965	2.10568
RMSE Variance	0.14357	0.00494	0.05365	0.01993	0.02856	0.00104	0.03849	0.00212	0.02471	0.08337

	Architecture					
Values	[128, 64]	[256, 64]	[128, 128, 128]	[128, 64, 32]	[256, 128, 64]	[512, 128, 64]
RMSE Mean	2.09144	1.98430	12.87466	2.16550	2.79058	4.72578
RMSE Variance	1.22484	0.20018	2165.27367	1.63662	8.81353	154.15435

	Dropout		
Values	0%	10%	50%
RMSE Mean	2.91854	2.25432	5.95588
RMSE Variance	35.26115	0.89345	0.08052

	Activation		
Values	ReLU()	LeakyReLU()	Sigmoid()
RMSE Mean	2.90450	8.40982	1.99730
RMSE Variance	66.76596	1103.96837	0.06840

	Batch Normalization	
Values	True	False
RMSE Mean	1.99481	6.87961
RMSE Variance	0.05785	0.08458

Tables 1-5: Test RMSE Mean and Variance for different hyperparameter values across 5cm SVID disjoint random experiments. [x,y,z] within the architecture parameter signifies the number of layers and the output size of each layer.

Within SVID, the effectiveness of different values of learning rate, architecture, dropout, activation, and batch normalization were examined. In Tables 1-5, we provide the results for each parameter, respectively. From this, we decided to use 10% dropout,

batch normalization, and the sigmoid activation function. We used this to narrow our architecture and learning rate parameter space and then proceeded to test these parameters across both lakes.

	Learning Rate						
Values	1.00E-08	5.00E-07	1.00E-07	5.00E-06	1.00E-06	1.00E-05	0.0001
RMSE Mean	2.08812	1.7669	1.81753	1.8163	1.41994	2.04223	1.86488
RMSE Variance	0.14357	0.00494	0.05365	0.01993	0.02856	0.01004	0.03849

	Epochs				
Values	50	100	200	300	500
RMSE Mean	3.27893	3.11808	2.94337	3.50035	3.46004
RMSE Variance	2.17915	1.94286	2.06004	1.73187	1.82983

	Architecture			
Values	[128, 64]	[256, 64]	[256, 128, 64]	[512, 256, 64]
RMSE Mean	3.33432	3.34764	3.27606	3.30259
RMSE Variance	1.84879	1.75549	1.78667	1.87758

	LR Decay	
Values	True	False
RMSE Mean	2.08812	2.97211
RMSE Variance	0.87236	0.00109

Tables 6-9: Test RMSE Mean and Variance for different hyperparameter values across 5cm combined SVID & LVID disjoint random experiments. [x,y,z] within the architecture parameter signifies the number of layers and the output size of each layer.

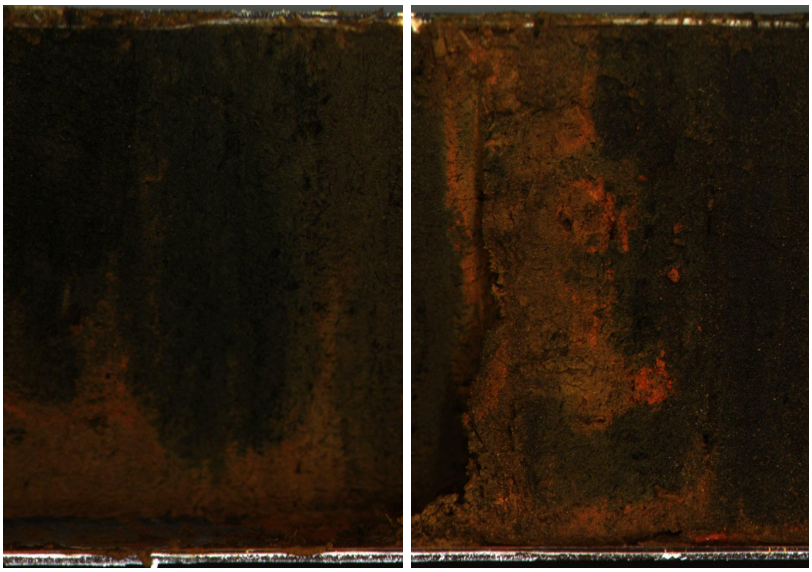
After narrowing our parameter search down, the combined LVID & SVID dataset was used to further hone in on the best learning rate value, the number of epochs, the final architecture, and whether to use learning rate decay. Tables 6-8 show the results from the following grid searches, respectively. Learning rate decay was less variable, albeit worse in terms of final RMSE values; therefore, we decided to leave out decay from our chosen model. After verifying an optimal learning rate of 0.01, the decision was made to prioritize efficiency in our modeling process by choosing 100 epochs for the following experiments. By opting for fewer epochs, we chose a more complex model architecture of three layers and a decreasing number of nodes between each layer from 256 to 128 and then to 64. The best overall model throughout all of our runs in terms of RMSE, was a model with this architecture, a learning rate of 0.01 and only trained for 100 epochs, even though in the aggregate combined dataset 200 epoch models outperformed 100 epoch models.

3.4 Experimental Design

3.4.1 Five-Centimeter Disjoint Experiments

The following experiments were run with varying partitions and forms of non-overlapping five-centimeter image chunks of data. The following experiments were run by creating a five-centimeter chunk for each consecutive chunk of data by sliding along a five-centimeter window along the entirety of the dataset. Each experiment focused on the prediction of a subset of our geochemical variables: %TOC, brGDGT, BSI, and C/N. Labeling each image with the corresponding proxy values involved averaging any measured values within the image depth range. For example, for the first

five-centimeter chunk of SVID data, there were two %TOC values: 3.58 at 0 centimeters and 3.06 at 3 centimeters. The %TOC label for this image chunk was then 3.315. The amount of data held out for validation remained at 20% throughout all experiments.



Figures 7 & 8: Two consecutive chunks from LVID with no overlap

3.4.1.1 Randomized Disjoint Experiment

The first experiment was to verify that our model could learn anything from the data at hand. The task was made simpler by expanding the amount of data available to the model. Instead of predicting at a high resolution of one centimeter, we first verified the effectiveness of the model in predicting five-centimeter chunks of data. In this experiment, 20% of the disjoint chunks were randomly selected as the test set. This experiment was run for all four geochemical variables of interest.

3.4.1.2 Sequential Disjoint Experiment

Next, we further challenged the model by seeing if it could predict completely unseen intervals of data. By feeding it the first 80% of the data, the model was then expected to forecast the changes in proxy values over the deepest 20% of the core. In this experiment, the last 20% of the disjoint chunks were selected as the test set. This experiment was run for all four geochemical variables of interest.

3.4.2 One-Centimeter Disjoint Experiments

Following the five-centimeter tests, we changed the task to learn the proxy values within a one-centimeter segment of sediment core data. The following experiments were run with varying partitions and forms of one-centimeter image chunks of data. The following experiments were run by creating a one-centimeter chunk for each consecutive chunk of data by sliding along a one-centimeter window along the entirety of the dataset. Each experiment focused on the prediction of a subset of our geochemical variables: %TOC, brGDGT, BSI, and C/N. The amount of data held out for validation remained at 20% throughout all experiments.

3.4.2.1 Randomized Disjoint Experiment

This experiment was analogous to the five-centimeter version, only with each training and test example a fifth of the size. In this experiment, 20% of the disjoint chunks were randomly selected as the test set. This experiment was run for all four geochemical variables of interest.

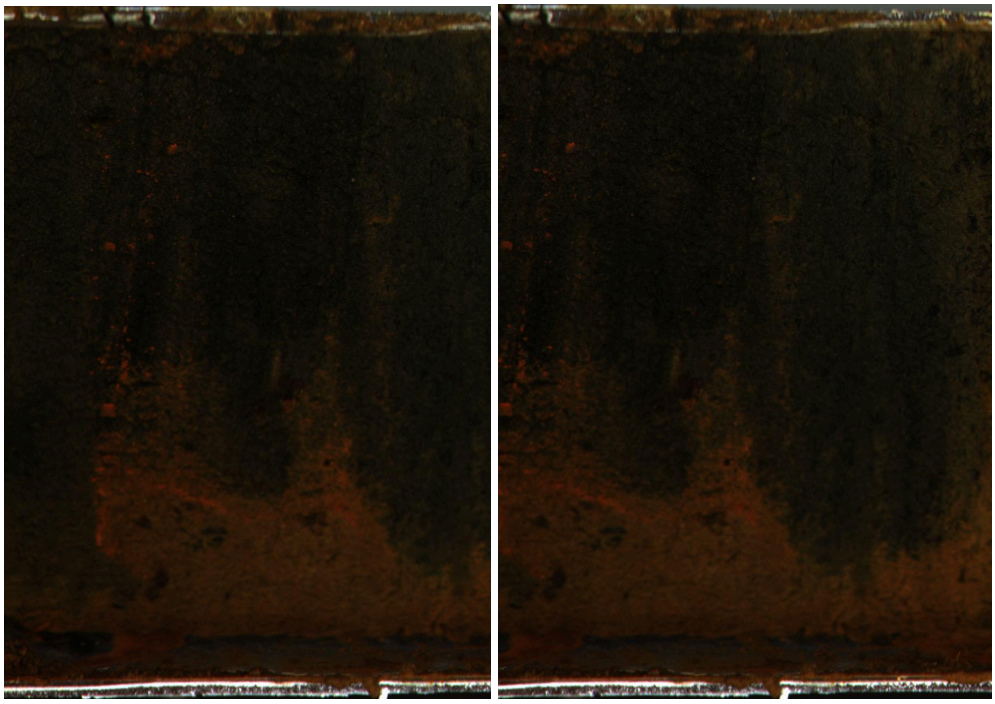
3.4.2.2 Sequential Disjoint Experiment

We presented our model with a similar challenge as before to check if it had greater or less success in the one-centimeter domain. In this experiment, the last 20% of the disjoint chunks were selected as the test set. This experiment was run for brGDGT and %TOC proxy values.

3.4.3 Moving Average (MA) Experiments

The following experiments were run by creating a five-centimeter chunk for each consecutive chunk of data by sliding along a one-centimeter window along the entirety of the dataset. This is a common practice in geoscience due to the difficulty of creating high-resolution data and the variability within even short time frames (77,78). To prevent data leakage, any centimeters placed within a test range were blacked out from the training set, and any centimeters placed within a training range were blacked out. To prevent the model from learning these breaks, 5% of unaffected chunks received a random one to three-centimeter removal of data from the sample. Without these additions, there would be an opportunity for the model to simply learn to predict certain values at the break or regurgitate information from adjacent training chunks. These checks alleviated these concerns and helped guide our model to learn what to predict only from the data at hand and the features in the data. At prediction, we predicted a one-centimeter interval of data by averaging the predictions for any five-centimeter chunk that included the target depth centimeter. For a given one-centimeter chunk, the model could average anywhere from one to five chunks that the target chunk was a part of. Not only did this approach almost expand our dataset to be five times the original size,

but the new format presented our model with the ability to learn at a five-centimeter level but predict at the high resolution of one centimeter.



Figures 9 & 10: Two consecutive chunks from LVID with 4 cm of overlap

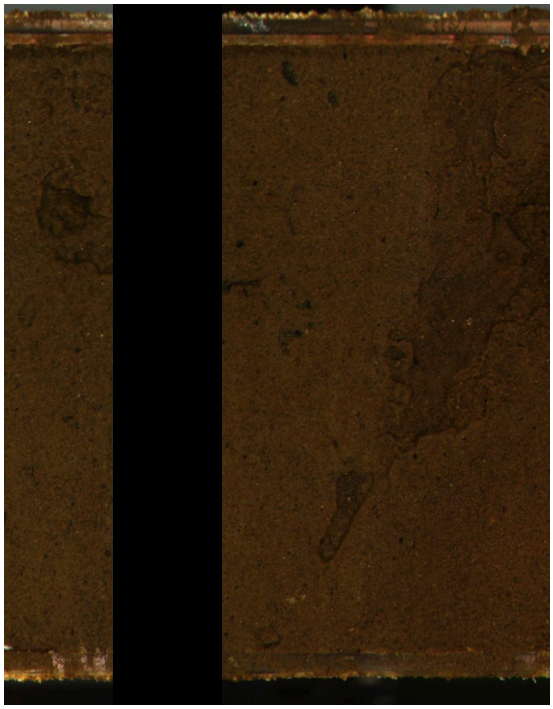
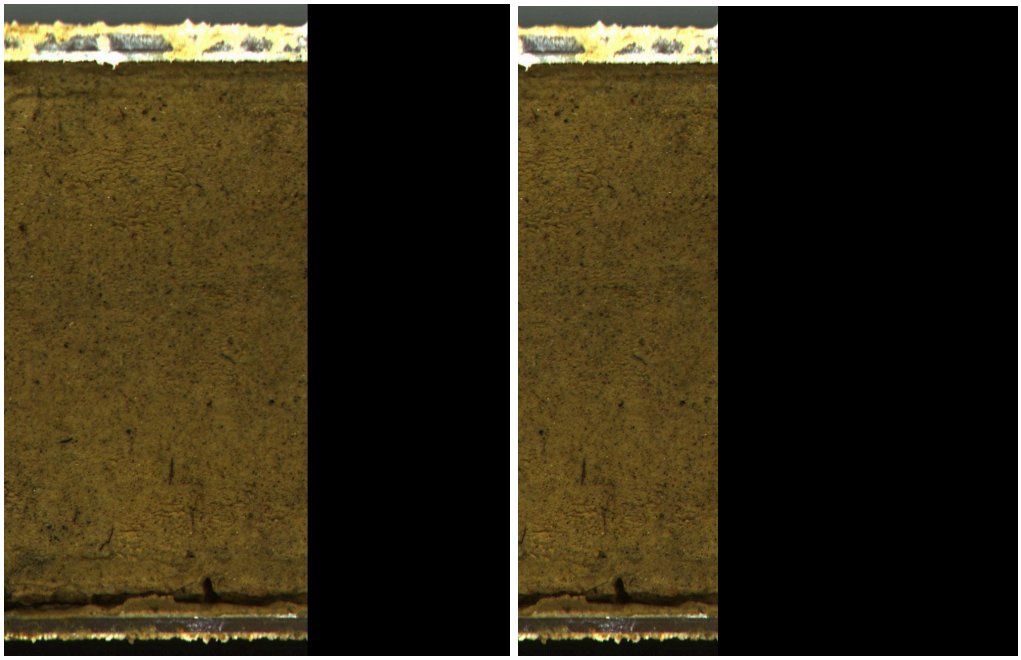


Figure 11: Chunk of SVID data with 1 cm randomly removed from the sample



Figures 12 & 13: Two chunks of SVID test data with 2 and 3 overlapping boundary cm removed

3.4.3.1 Randomized Experiment

To minimize the gaps between the two sets, ten-centimeter test blocks were randomly selected every fifty centimeters. This spacing was essential due to how we created our train and test chunks. If this process was instead completely random, many chunks would have few chunks to predict from, or chunks with a majority of the data around it blacked out. This experiment was run for all four geochemical variables of interest.

3.4.3.2 Five-Fold Experiment

In this experiment, we split the data into five sequential blocks for each lake. Then, we trained on four of the five folds, while withholding one fold for validation. With this experiment, we aimed to build off our findings of the sequential experiments laid out in 3.4.1.2 and 3.4.2.2 and test beyond only the last section of data. This allowed us to

examine the effectiveness of our model at different points in the sediment history, while at the same time checking its performance on completely unseen parts of the data. This experiment was run on both the %TOC and brGDGT proxy variables.

3.4.4 t-SNE Experiment

In this experiment, we used the embeddings from the trained model of predicting one-centimeter and five-centimeter disjoint random chunks to perform a t-SNE experiment on the data with the two lakes being our categories of interest. We first trained the model on the task of proxy prediction using both lakes as the train and test set. Then, rather than gathering the prediction for the test sets, we extracted the embedding of each sample from the model. The embedding was taken from the step immediately following the three convolutional layers. Then, we mapped the embedding in a 2-D space using the t-SNE method (90) to compare the distribution of the test examples across the lakes. This experiment was run for validation purposes. The sub-space of distinguishing between lakes allowed us to have qualitative backing for the model's learning ability. This allowed us to infer whether the model was learning something interesting or merely guessing based on similar examples in the training sets and getting lucky. Although a model could be able to learn the patterns within the data without distinguishing between the two lakes, the results of this experiment were important in verifying the potential for further generalization of the model and its techniques. We ran this experiment for both %TOC and brGDGT.

3.4.5 Manual Validation Experiment

In this experiment, the most “hard” and “straightforward” chunks from both our one-centimeter and five-centimeter disjoint random experiments were selected to create a dataset of core images. This was determined by selecting the five highest and five lowest RMSE values across experiments 2.4.1.1 and 2.4.2.1. The second stage involved a labeler to classify each image chunk as either “hard” or “easy” to manually label. The labeler, Jon, had a Ph.D. in Geology and was one of the collectors and labels of the initial data, so this process was well within his expertise. Every low RMSE sample marked as “hard” and every high RMSE sample marked as “easy” was noted as a difference in opinion between human and machine annotation. If the model was truly learning, it would be logical for it to struggle on conventionally difficult chunks of data and excel on simpler or repeated patterns. This also served as a manual check to verify that there wasn’t a certain “type” of sample that the model had consistently high error rates on. For example, if we examined this dataset and noticed that any example with significant volcanic ash deposits, led to a high error rate for our model, we would be able to better understand the struggles of our model from a more human and explainable perspective. At the conclusion of the test, we determined a pseudo-accuracy score by dividing the number of samples where the two opinions lined up over the total sample size.

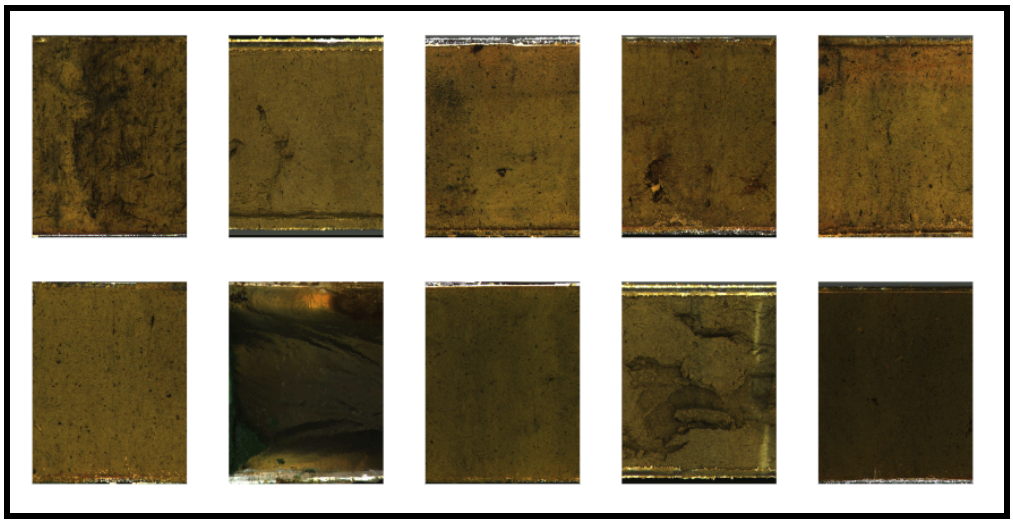


Figure 14: Five-centimeter chunk test dataset. Results laid out in section 4.3.2

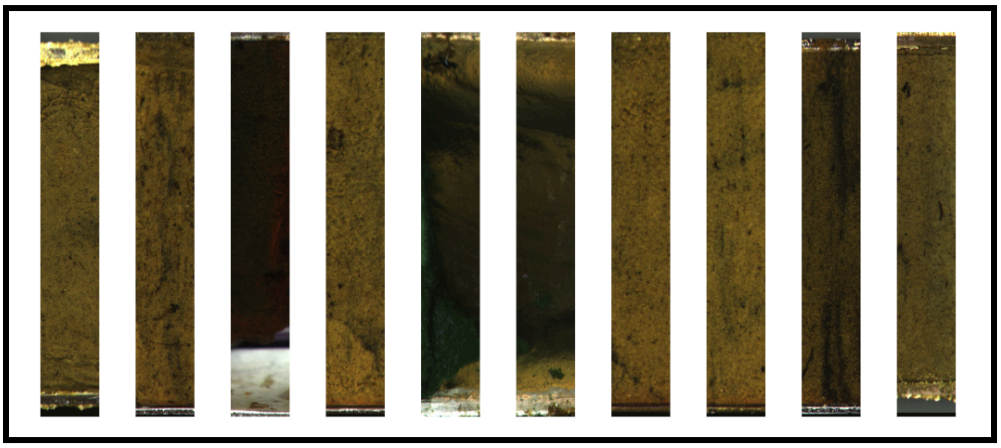


Figure 15: One-centimeter chunk test dataset. Results laid out in section 4.3.2

4 Results

4.1 Disjoint Experiments

4.1.1 Randomized Experiment Results

With our initial randomized disjoint experiments, we wanted to not only hone in on the best set of parameters for our model (section 2.3.3.1) but also to verify that learning was

possible within this domain. As seen in Figures 16 & 17, learning proceeded sufficiently through 100 epochs. There was a sharp decrease in loss early on with a plateauing close to zero. Further verifying our decision of 100 epochs, in Figure 17, loss within SVID after 100 epochs stayed extremely close to zero, leading us to disregard training the model beyond 100 epochs.

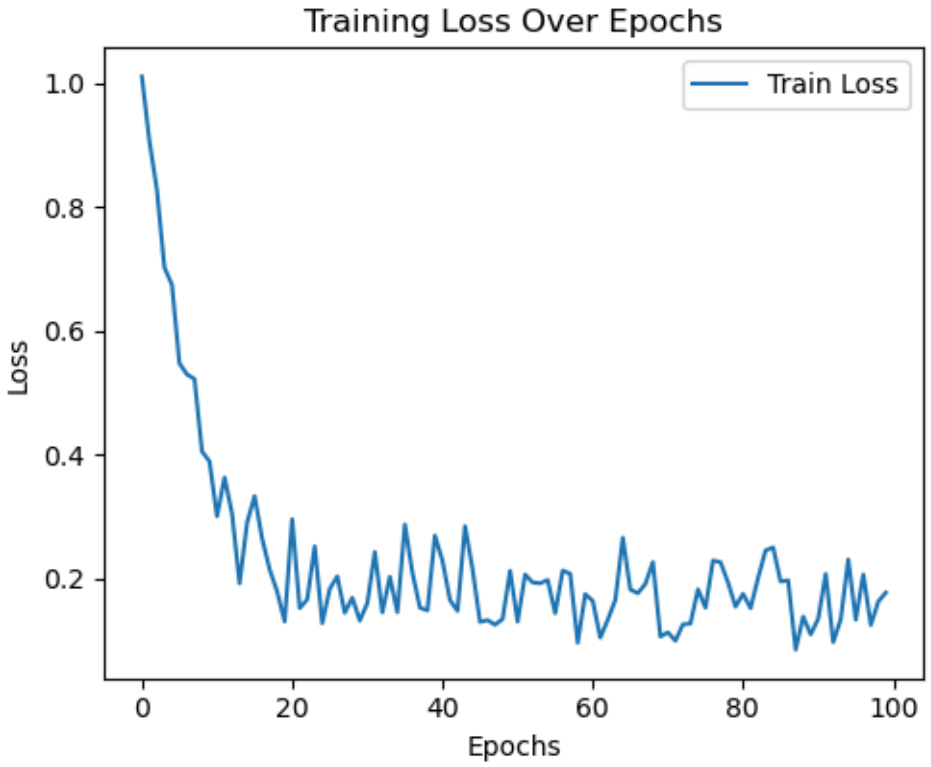


Figure 16: LVID %TOC train loss over 100 epochs within 5 cm experiment

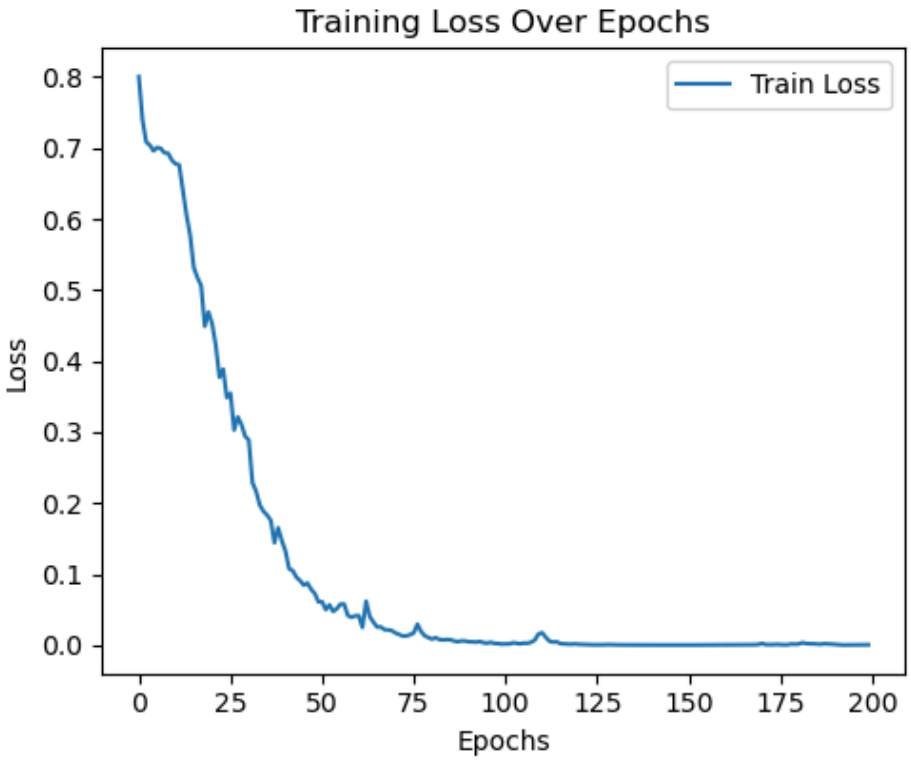
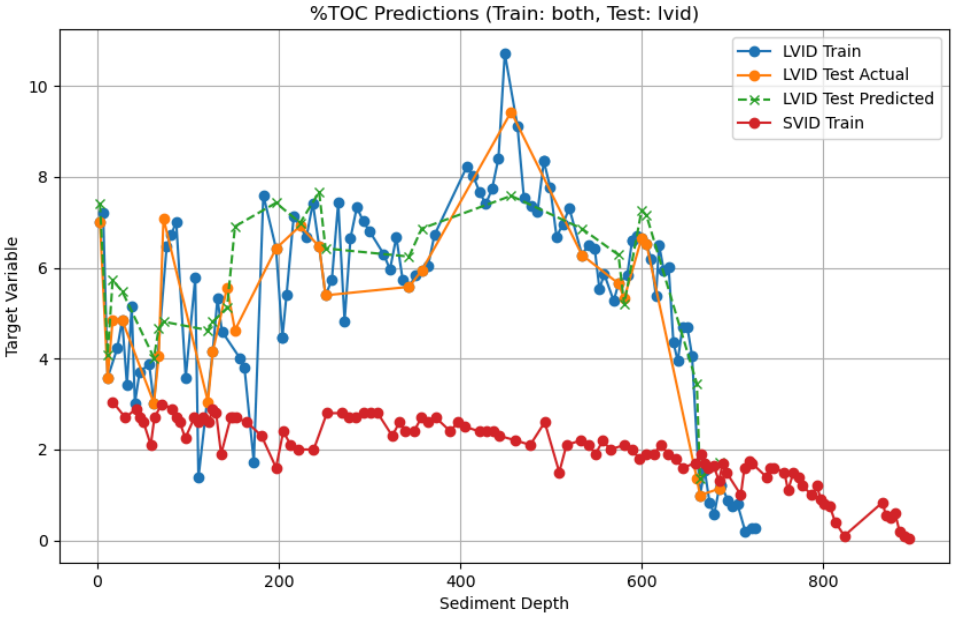


Figure 17: SVID brGDGT train loss over 200 epochs within 5 cm experiment

Promise was shown when using both lakes as a train set and predicting one entire lake, in both the one-centimeter and five-centimeter domains. Focusing on the LVID %TOC test set and the comparison of performance when using both lakes for training or only LVID, we see an improvement in the model's ability to capture the overall trend of the data. By comparing Figures 18 and 19, it is easy to see a better ability to capture the spike around 475 centimeters and more closely follow the trends in the first 100 centimeters. In the one-centimeter plots, Figures 20 and 21, the additional SVID data is seen to be most helpful within the first 400 centimeters. It is unable to follow it perfectly; however, the additional lake helps for a much closer fitting to the true trend in the data.



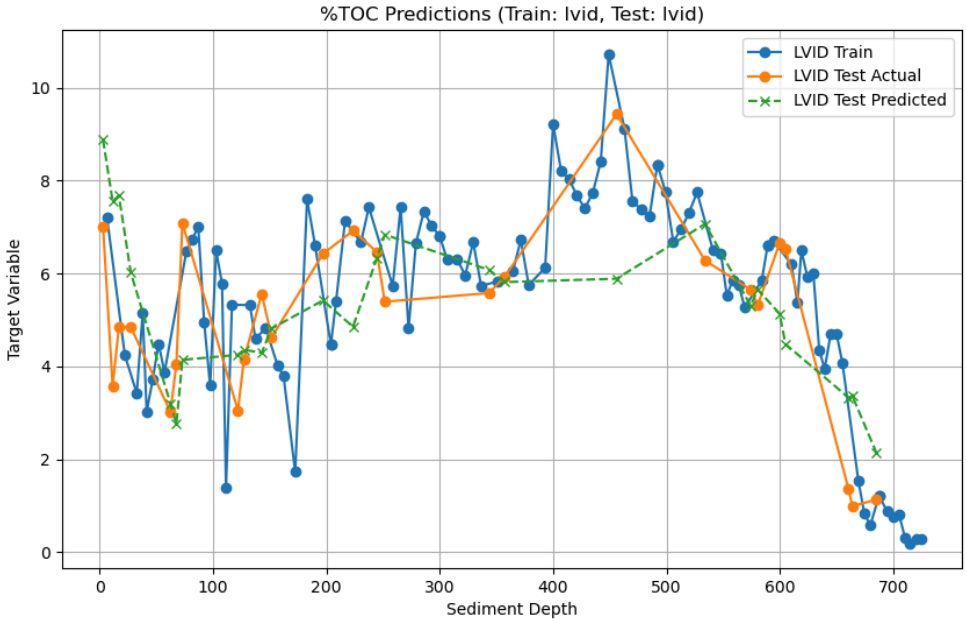
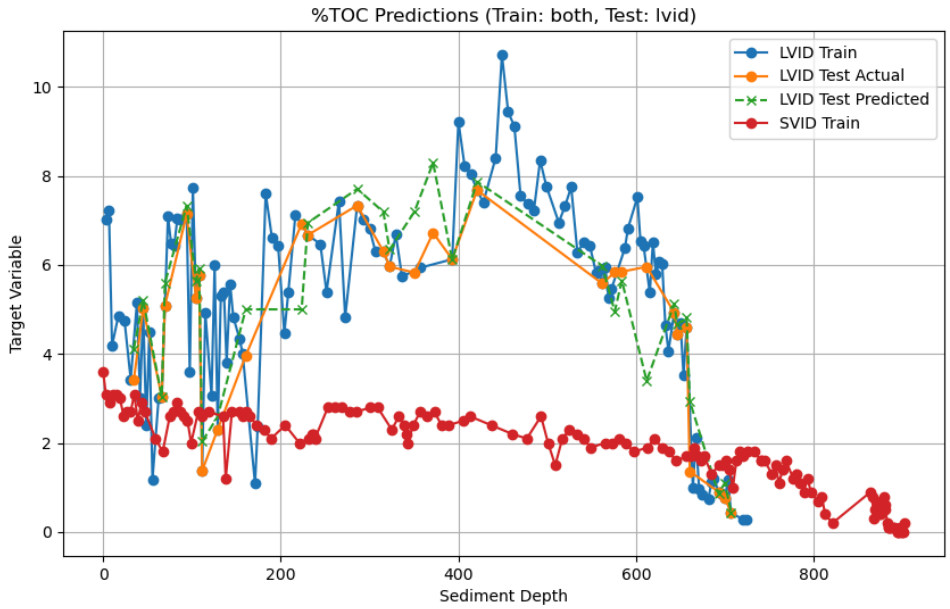


Figure 18: Random disjoint 5 cm LVID %TOC prediction over both lakes

Figure 19: Random disjoint 5 cm LVID %TOC prediction over LVID



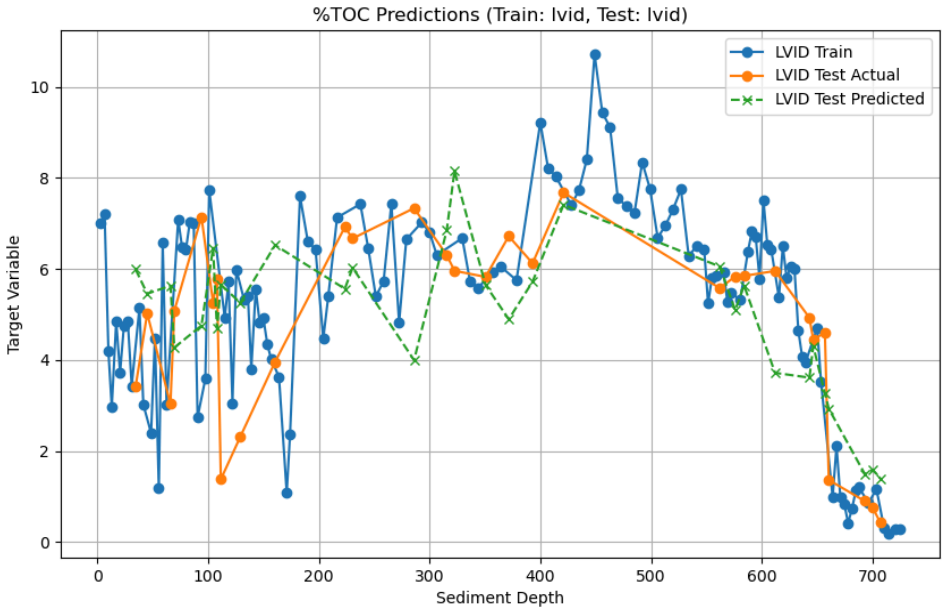


Figure 20: Random disjoint 1 cm LVID %TOC prediction over both lakes

Figure 21: Random disjoint 1 cm LVID %TOC prediction over LVID

For a more quantitative analysis of the addition of the second lake in the training sets, Tables 10-13 provide RMSE measures across all of the different combinations of train and test sets. As expected, in the majority of cases training on one lake and then testing on a completely unseen lake leads to the highest error measures. It is not always

the case that training on two lakes aids in performance. It is difficult to estimate the strength of the fits across the different variables. Simple R^2 analysis does not work in the presence of the non-linearity in our dataset. It is encouraging, however, that training across lakes either helps or leads to similar performance across lakes. The most surprising result is the success of training on LVID and testing on SVID for the brGDGT proxy values (Figure 22); however, due to the small amount of brGDGT data and the randomness of the selection of the test set, the conclusions that can be drawn from this result are limited, as emphasized in further experimental results. Regardless, adding in SVID to the train set helps overall model performance, as seen in the residual plot in Figure 23.

Train Set	Test Set	Target	RMSE
both	svid	MBT	0.013321049
lvid	svid	MBT	0.023391154
svid	svid	MBT	0.026589425
lvid	both	MBT	0.032437105
both	both	MBT	0.032825675
both	lvid	MBT	0.033068534
svid	both	MBT	0.035293996
lvid	lvid	MBT	0.037755854
svid	lvid	MBT	0.08935708
both	svid	%TOC	0.41131452
svid	svid	%TOC	0.63197637
both	lvid	%TOC	1.0936381
both	both	%TOC	1.4251784
lvid	lvid	%TOC	1.757412
svid	both	%TOC	2.4190009
svid	lvid	%TOC	3.2808833
lvid	both	%TOC	3.4468436
lvid	svid	%TOC	3.970675

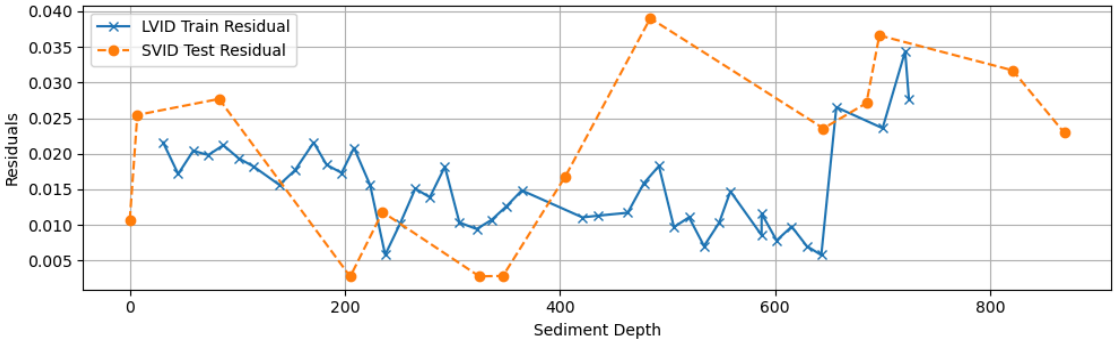
Train Set	Test Set	Target	RMSE
svid	svid	C-N	0.61255926
both	svid	C-N	0.77776265
both	lvid	C-N	0.9273869
both	both	C-N	0.98103905
lvid	lvid	C-N	1.1631758
svid	both	C-N	1.82114
svid	lvid	C-N	2.6804981
lvid	both	C-N	2.8288462
lvid	svid	C-N	3.2169452
both	svid	BSI	13.196177
both	both	BSI	14.478946
lvid	lvid	BSI	16.842072
svid	lvid	BSI	20.980274
both	lvid	BSI	23.790842
svid	both	BSI	29.907568
svid	svid	BSI	34.48197
lvid	both	BSI	38.13862
lvid	svid	BSI	67.558525

Tables 10-11: Random disjoint 5 cm proxy RMSE values, with brGDGT noted as MBT, over different combinations of train and test sets

Train Set	Test Set	Target	RMSE
both	svid	MBT	0.01893706
svid	svid	MBT	0.021227762
svid	both	MBT	0.027339462
lvid	both	MBT	0.029619541
both	both	MBT	0.03676841
lvid	lvid	MBT	0.040706873
lvid	svid	MBT	0.04389554
both	lvid	MBT	0.056721713
svid	lvid	MBT	0.11640741
svid	svid	%TOC	0.48452526
both	svid	%TOC	0.56187785
both	lvid	%TOC	0.85766
both	both	%TOC	1.2392733
lvid	lvid	%TOC	1.7514291
svid	both	%TOC	2.3254762
lvid	both	%TOC	3.3553202
svid	lvid	%TOC	3.7089012
lvid	svid	%TOC	4.9060416

Train Set	Test Set	Target	RMSE
svid	svid	C-N	0.5899284
both	svid	C-N	0.7340229
lvid	lvid	C-N	0.8748223
both	both	C-N	0.95827043
both	lvid	C-N	1.9040114
lvid	both	C-N	2.3194323
svid	both	C-N	2.5843735
lvid	svid	C-N	3.0294523
svid	lvid	C-N	3.2288086
both	lvid	BSI	12.87127
both	svid	BSI	13.867443
both	both	BSI	14.687435
svid	svid	BSI	19.197136
lvid	lvid	BSI	21.641983
svid	both	BSI	23.53028
svid	lvid	BSI	25.246235
lvid	svid	BSI	36.74999
lvid	both	BSI	45.700195

Table 12-13: Random disjoint 1 cm proxy RMSE values, with brGDGT noted as MBT, over different combinations of train and test sets



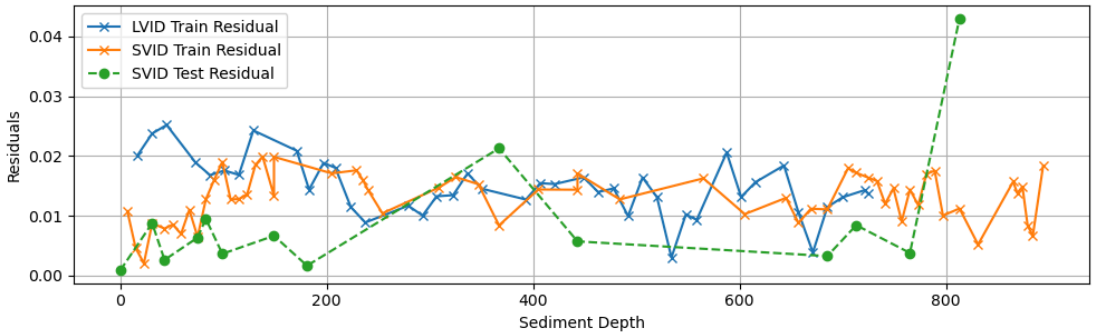


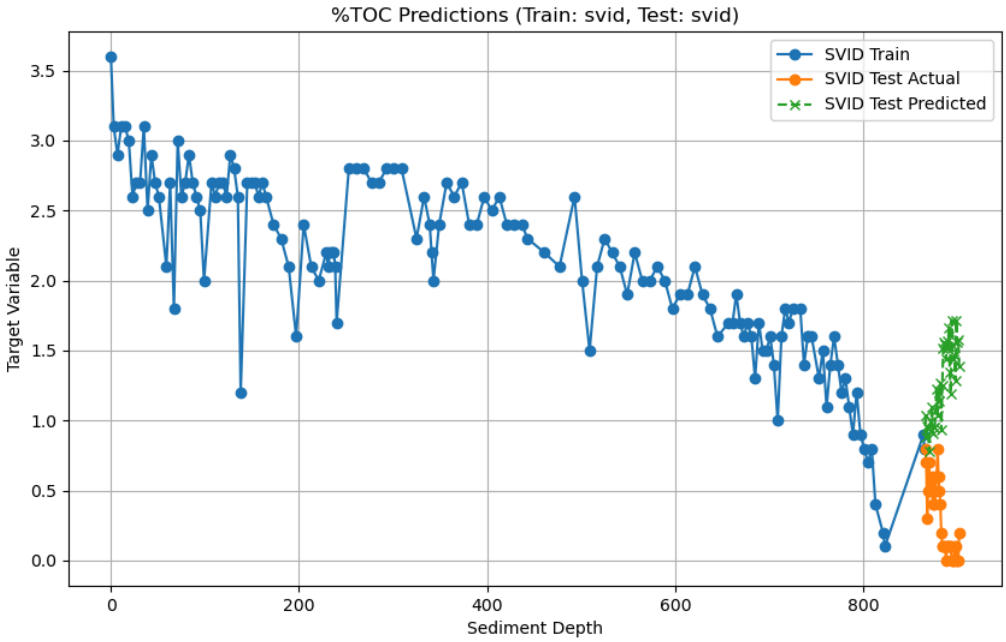
Figure 22: Random disjoint 5 cm SVID brGDGT residuals when training over LVID

Figure 23: Random disjoint 5 cm SVID brGDGT residuals when training over both lakes

4.1.2 Sequential Experiment Results

Challenging the model with a sequential breakup of our test and train splits proved to be, in no better words, a challenge. With this in mind, we decided to only explore %TOC and MBT for this stage. As shown in the one-centimeter predictions of Figures 24 & 25, even within lakes for %TOC – the more abundant proxy measurement, the fit left much to be desired. In both lakes, there is a strong overshooting in the predicted values, and in the case of SVID, values went entirely in the opposite direction following the last observation of the training set. Figure 26 shows the same task within SVID for five-centimeter chunks of data. The performance is more encouraging in the middle section of the test dataset; however, there is still strong divergence at the initial boundary

of the test set as well as in the last third of the test data, leading to an inability to trust the overall predictive power of the model within the task of out-of-sample intralake prediction.



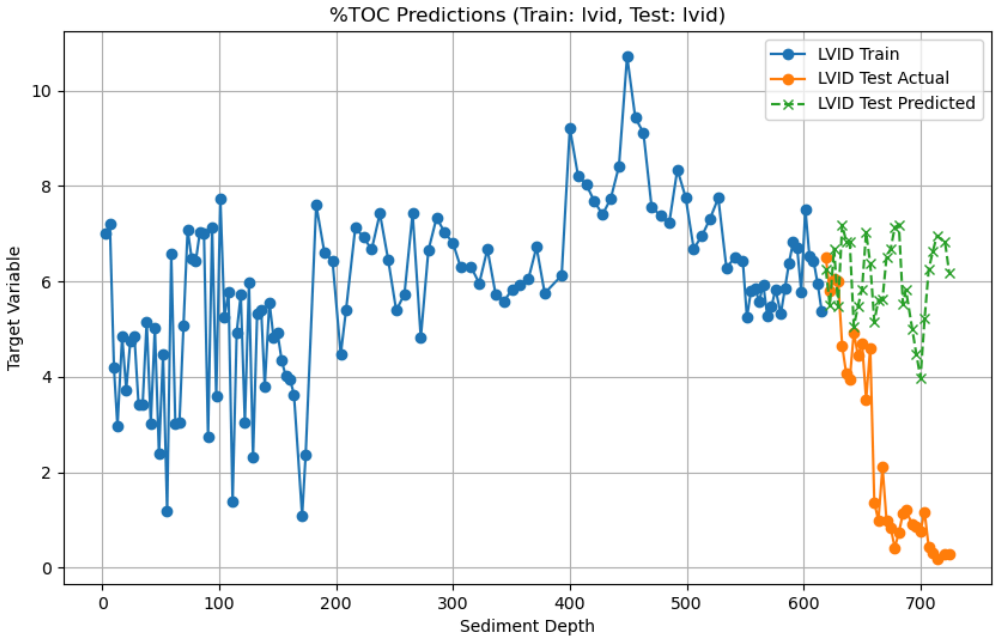


Figure 24: Sequential disjoint 1 cm SVID %TOC prediction over SVID

Figure 25: Sequential disjoint 1 cm LVID %TOC prediction over LVID

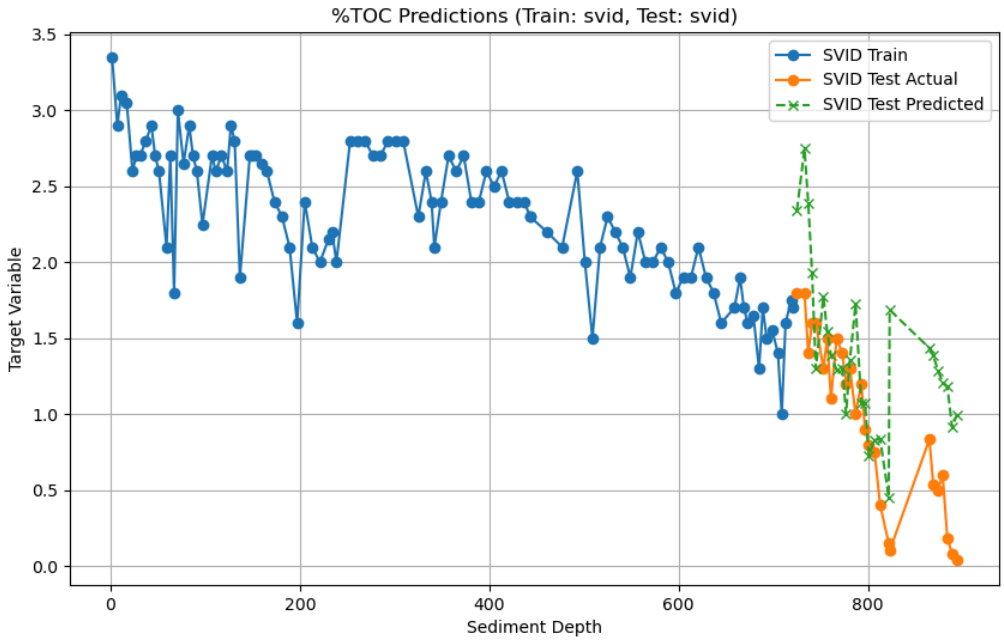


Figure 26: Sequential disjoint 5 cm SVID %TOC prediction over SVID

Merely by comparing Tables 10 & 12 with Tables 14-15, it is plain to see the decrease in performance in the new task. In regards to the comparison between Table 10 and Table 14, the highest brGDGT score in SVID now has an error on par with a middle-of-the-pack score in the previous experiment. Likewise, the higher RMSE value is almost double that of the corresponding task in the random split. Similarly, comparing Table 12 with Table 15, we see the same trends repeating themselves across the one-centimeter data. Moreover, the low RMSE value seen when training with both lakes and testing on SVID in Table 12 is over 300% larger when repeating the experiment in the sequential setup.

Train Set	Test Set	Target	RMSE
svid	svid	MBT	0.041075133
both	svid	MBT	0.047019716
lvid	svid	MBT	0.05132088
both	lvid	MBT	0.051337633
lvid	lvid	MBT	0.052078813
lvid	both	MBT	0.065972656
svid	both	MBT	0.06996567
both	both	MBT	0.07723051
svid	lvid	MBT	0.106680945
svid	svid	%TOC	0.6333953
both	svid	%TOC	1.1765397
svid	both	%TOC	1.7351246
svid	lvid	%TOC	2.348867
both	both	%TOC	3.2076359
lvid	svid	%TOC	3.5125787
lvid	both	%TOC	4.0897584
both	lvid	%TOC	4.2341113
lvid	lvid	%TOC	4.519088

Train Set	Test Set	Target	RMSE
both	svid	MBT	0.054537393
svid	svid	MBT	0.059529647
both	both	MBT	0.065888636
both	lvid	MBT	0.06709517
lvid	svid	MBT	0.06756277
svid	both	MBT	0.0706995
lvid	both	MBT	0.071065284
lvid	lvid	MBT	0.09319979
svid	lvid	MBT	0.1045979
svid	svid	%TOC	1.0870887
svid	both	%TOC	1.6057658
both	svid	%TOC	1.9475548
svid	lvid	%TOC	2.0866146
both	both	%TOC	2.872385
both	lvid	%TOC	3.696786
lvid	svid	%TOC	3.9732344
lvid	both	%TOC	4.170936
lvid	lvid	%TOC	4.2243485

Table 14-15: Sequential disjoint 5 cm & 1 cm proxy, RMSE values, respectively, over different combinations of train and test sets

4.2 Moving Average Experiments

4.2.1 Randomized Moving Average Results

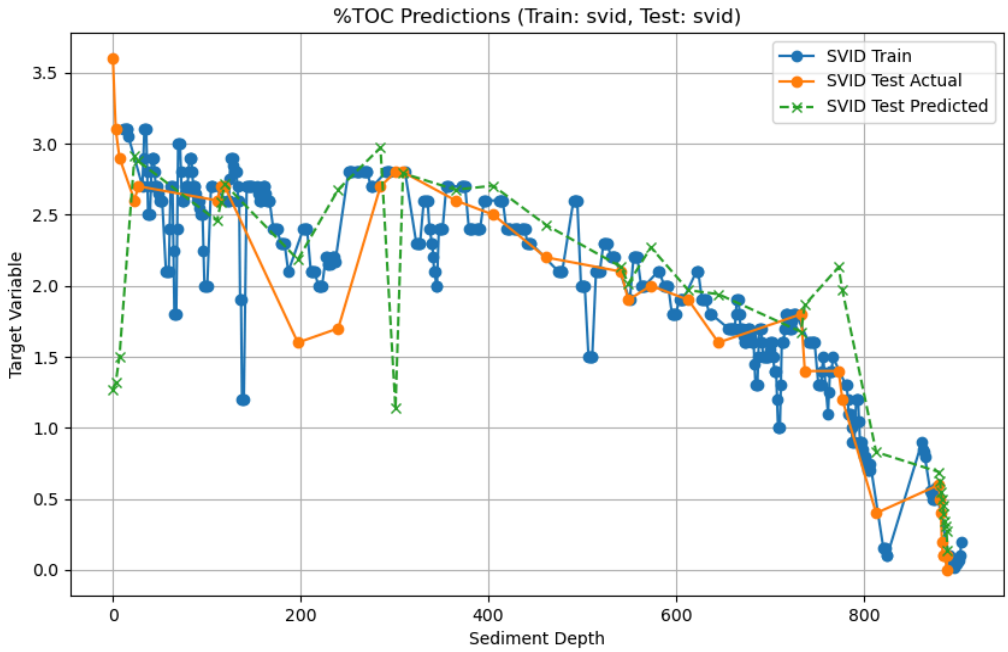
As stated in section 3.4.3, this experiment was different in that we completely remade our dataset by using moving average techniques and edited our prediction method for a one-centimeter chunk of data by averaging all of the five-centimeter predictions that the smaller sliver was a part of. The results were mixed when changing the dataset to be moving average one-centimeter predictions rather than only random one-centimeter predictions. By comparing Table 12 with Table 16, we can generate some comparisons

across the two methods. For %TOC, the model seems to be more stable, with a narrower spread than the random disjoint model; however, at the top, the random disjoint model outperforms the moving average model for %TOC. Something to note, as shown in Figures 27 and 28, is the difficulty of the model learning at the boundaries. Due to our blackout strategy, we would expect these predictions to be less accurate based on the lack of examples for averaging at the boundaries. This is most prominent in the first three test predictions, as well as the prediction for the 333rd centimeter. It's promising to note the alleviation of the hard divergence on centimeter 333 when training on both lakes in Figure 28.

Train Set	Test Set	Target	RMSE
svid	svid	%TOC	0.6919630202054130
both	svid	%TOC	0.8032980295810030
both	both	%TOC	0.9390736627192000
both	lvid	%TOC	1.1390255136480700
lvid	lvid	%TOC	1.220555961895240
lvid	both	%TOC	2.3040749622344500
svid	both	%TOC	2.667681606600470
lvid	svid	%TOC	2.6906820190941700
svid	lvid	%TOC	4.182629266774360

Train Set	Test Set	Target	RMSE
svid	lvid	MBT	0.01541411264547260
both	lvid	MBT	0.016883497348927100
both	both	MBT	0.022255829340664400
svid	svid	MBT	0.024713494060836400
lvid	lvid	MBT	0.03385651942283110
both	svid	MBT	0.04816465716691940
svid	both	MBT	0.06120463335454560
lvid	svid	MBT	0.07334084529779310
lvid	both	MBT	0.08531255414796050

Table 16-17: Random MA RMSE values over different train/test combinations for %TOC and brGDGT



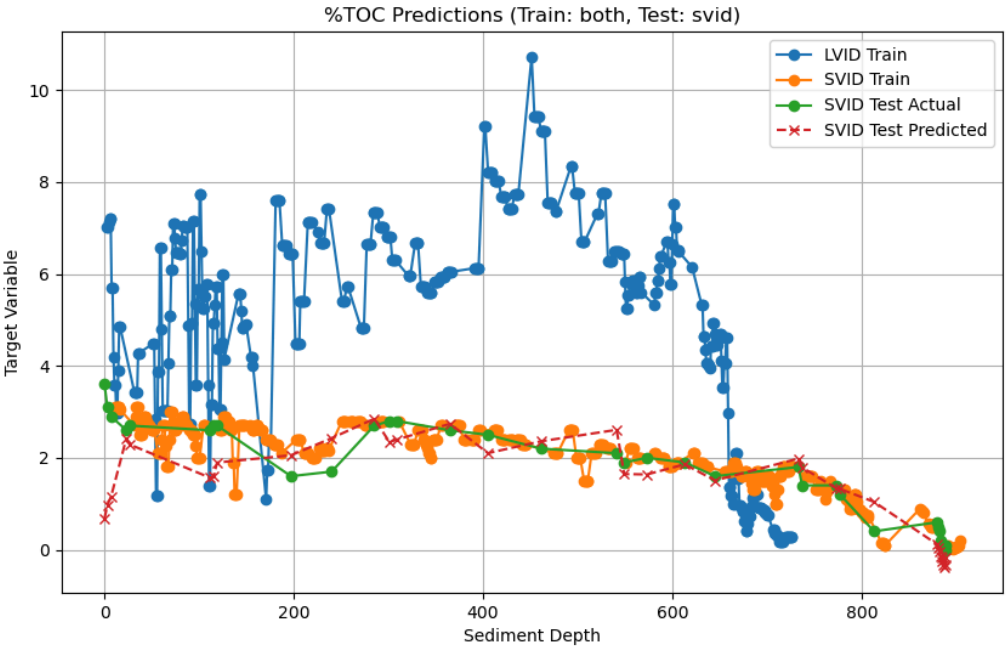


Figure 27: Random MA SVID %TOC prediction over SVID

Figure 28: Random MA %TOC prediction over both lakes

The brGDGT performance is better than the disjoint experiment; however, the baseline is also lower due to the higher difficulty in predicting brGDGT. We still see the same RMSE spike at the front of the test dataset; interestingly, in Figure 30, adding LVID

to the train set increases the initial error, whereas adding SVID to the LVID prediction brings down the error in Figure 32. The story continues throughout the plot as in the prediction of SVID, the model seems to get confused by adding in LVID, whereas adding in SVID allows the data to come much closer to the true LVID values. Nonetheless, even with the questions this inconsistency presents, it is promising that this new method allows for a harder task to become easier.

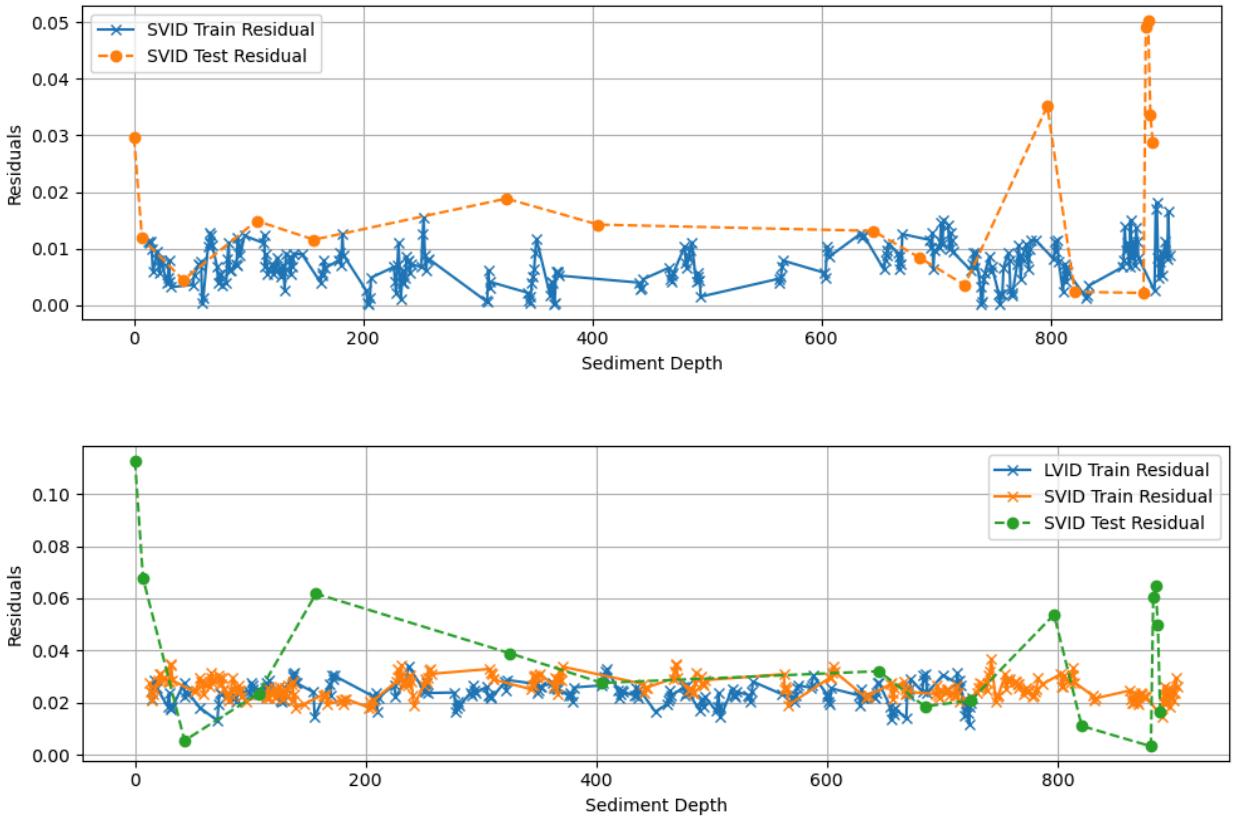
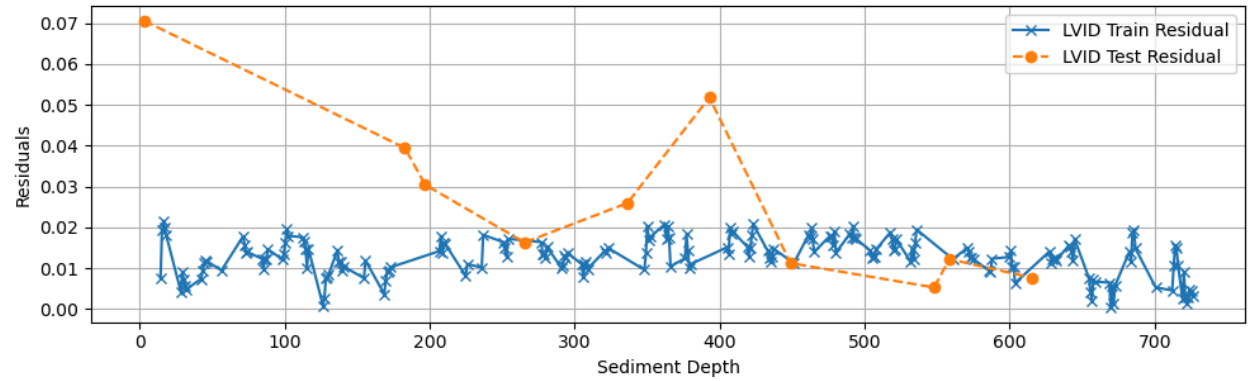


Figure 29: Random MA SVID brGDGT residuals when training over SVID

Figure 30: Random MA SVID brGDGT residuals when training over both lakes



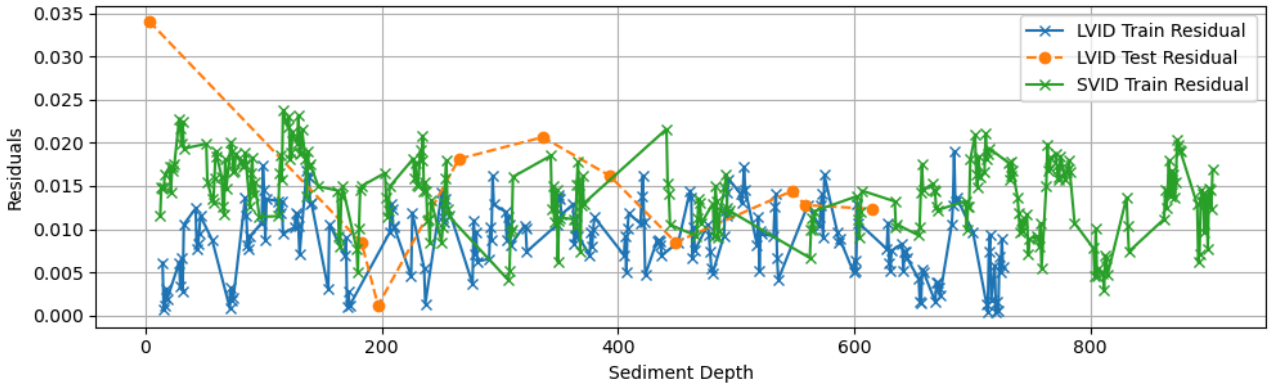


Figure 31: Random MA LVID brGDGT residuals when training over LVID

Figure 32: Random MA LVID brGDGT residuals when training over both lakes

4.2.2 Five-Fold Moving Average Results

This was the most challenging task. We hypothesized that the model would do better on the interior folds, and struggle at the boundary of the entire dataset. The performance was very hit-or-miss. Sometimes good, as shown in Figure 36, and sometimes completely erratic, as shown in the other four folds. This provided a challenge and a clear limit for our modeling. Interestingly, there were some folds where the variance seemed to be captured, as shown in Figure 33, but the values seemed to be offset. This presents interesting questions and warrants more discussion in section 5. Overall, as laid out in

Tables 18 and 19, the RMSE values were much worse and the results were not nearly as promising as our standard test set distributed throughout the entirety of the data.

Train Set	Test Set	Fold	Target	RMSE
both	svid	1	mbt	0.01724825966247140
both	svid	3	mbt	0.022838706043180500
both	both	3	mbt	0.02780431336527180
both	both	2	mbt	0.030718255289359900
svid	svid	1	mbt	0.03251834683608740
both	lvid	0	mbt	0.03357925322605680
both	lvid	2	mbt	0.037151912910069100
both	both	0	mbt	0.04019782948346100
svid	svid	4	mbt	0.04046074643969360
svid	svid	0	mbt	0.04056890468377300
both	svid	0	mbt	0.04691153092730170
both	both	1	mbt	0.051522271586190400
both	both	4	mbt	0.05802786242739050
svid	svid	2	mbt	0.06669987940910650
both	svid	4	mbt	0.07774365347506040
lvid	lvid	1	mbt	0.08073238727327380
lvid	lvid	4	mbt	0.09090213801131020
both	lvid	4	mbt	0.09659257612184180
lvid	lvid	3	mbt	0.1026437204649270
both	svid	2	mbt	0.10751428925293800
lvid	lvid	2	mbt	0.1283755825295720
svid	svid	3	mbt	0.13316240221746800
both	lvid	1	mbt	0.13471060358047500
lvid	lvid	0	mbt	0.21578083371259200
both	lvid	3	mbt	0.2571650991326310

Train Set	Test Set	Fold	Target	RMSE
svid	svid	3	%toc	0.4067783968705830
both	svid	3	%toc	0.43991967399597400
svid	svid	1	%toc	0.8734469332365700
both	svid	0	%toc	0.9501963062876790
both	svid	2	%toc	1.0404339491735100
svid	svid	2	%toc	1.0714220675959100
svid	svid	0	%toc	1.0816190958485000
both	both	1	%toc	1.2518074375943400
svid	svid	4	%toc	1.3330856989449900
lvid	lvid	2	%toc	1.6365533534669200
lvid	lvid	3	%toc	1.6953857649922200
both	svid	1	%toc	1.7645101683987100
lvid	lvid	0	%toc	1.9505157083625700
both	svid	4	%toc	2.126268497167800
both	both	4	%toc	2.413904877594110
both	lvid	4	%toc	2.8139147640631700
both	both	3	%toc	2.952563901555000
lvid	lvid	1	%toc	3.0179297270358400
both	both	0	%toc	3.020921794892520
both	both	2	%toc	3.026623104323340
both	lvid	0	%toc	3.0592836485746900
lvid	lvid	4	%toc	3.7520990143137500
both	lvid	1	%toc	4.000322544718600
both	lvid	3	%toc	4.9166414927007000
both	lvid	2	%toc	5.639897776217500

Table 18-19: Five-fold MA brGDGT and %TOC RMSE values across folds

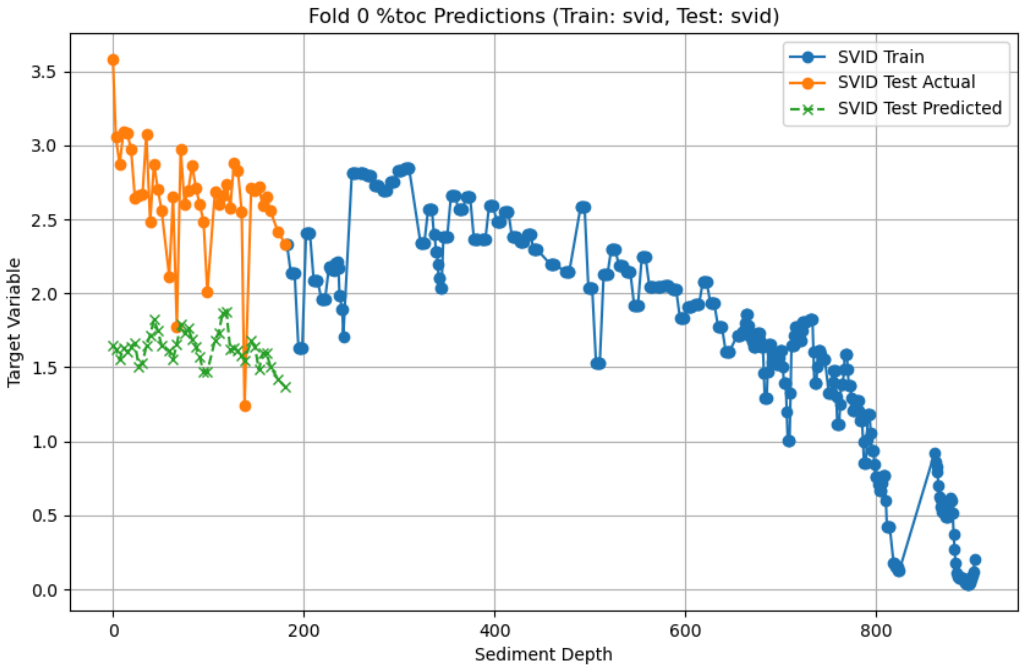


Figure 33: Fold 0 MA SVID prediction over SVID for five folds of the dataset

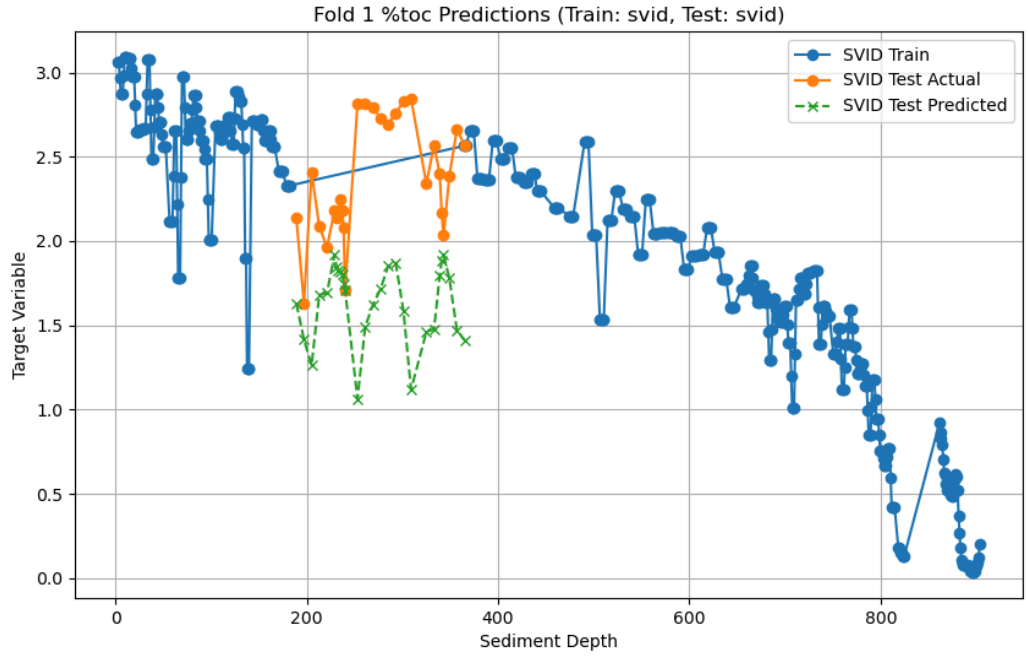


Figure 34: Fold 1 MA SVID prediction over SVID for five folds of the dataset

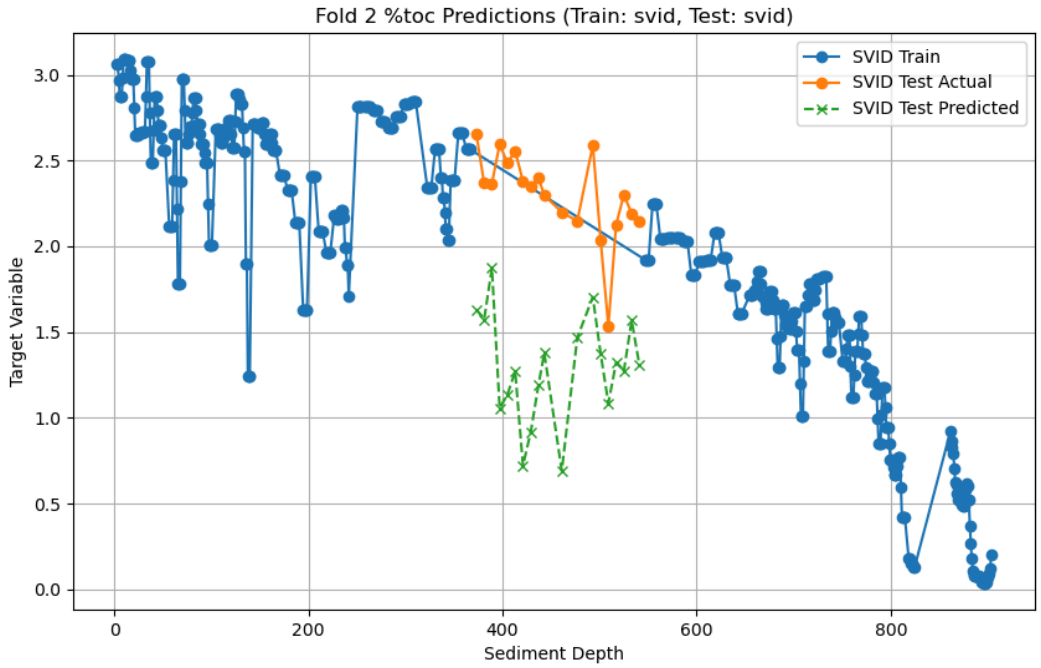


Figure 35: Fold 2 MA SVID prediction over SVID for five folds of the dataset

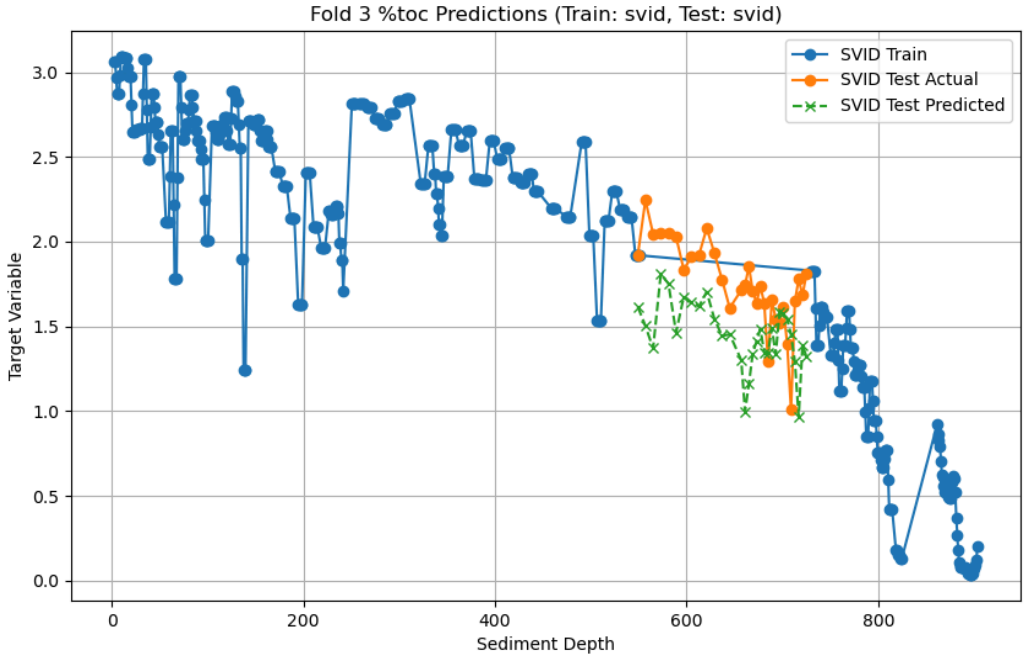


Figure 36: Fold 3 MA SVID prediction over SVID for five folds of the dataset

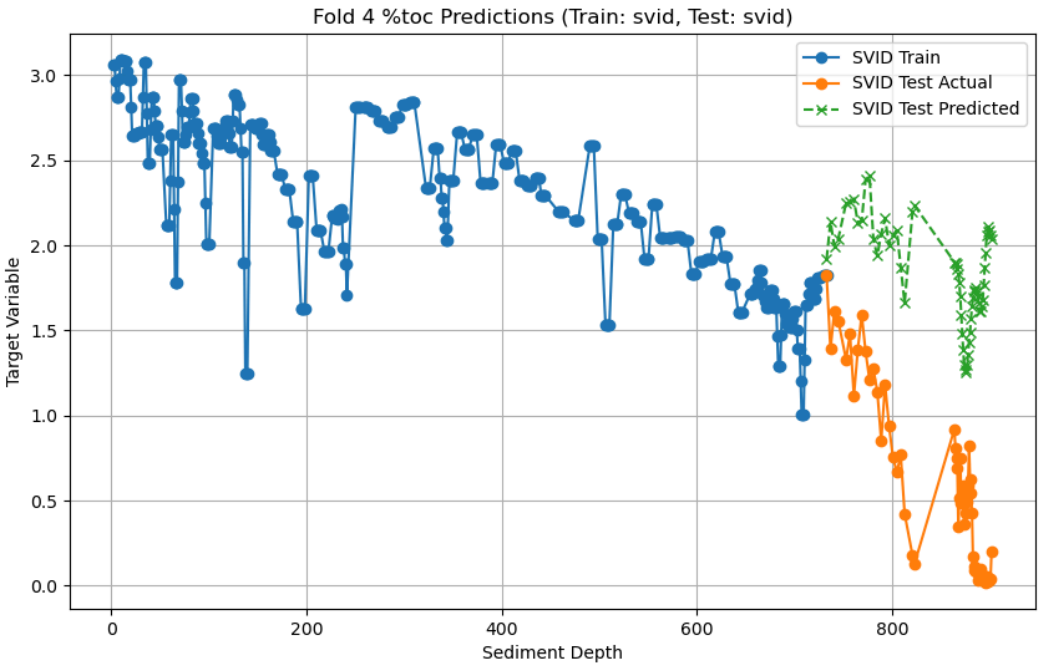


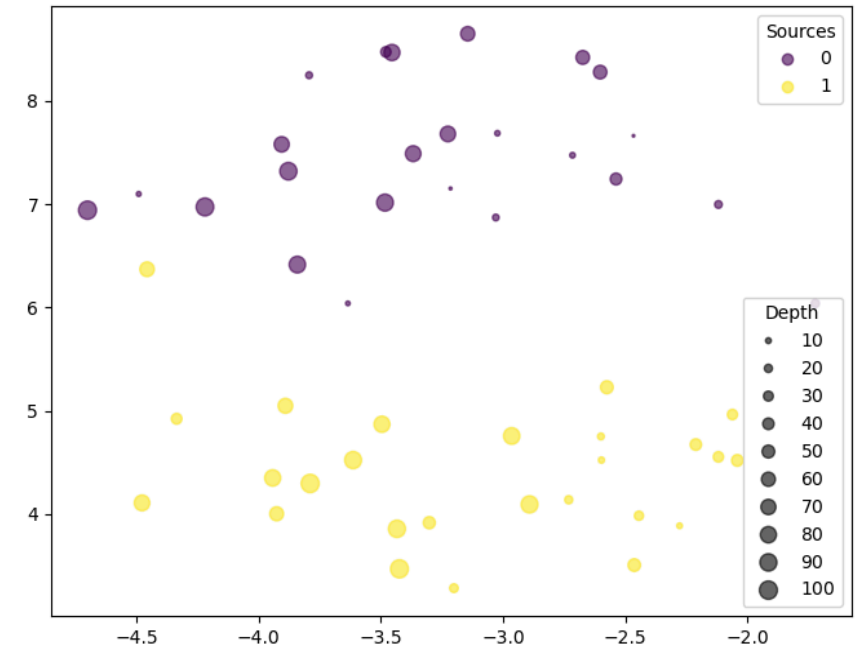
Figure 37: Fold 4 MA SVID prediction over SVID for five folds of the dataset

4.3 Validation Experiments

4.3.1 t-SNE

Within %TOC, the model did a very reasonable job of distinguishing between the two lakes. Within both chunk sizes, it was clear the two different clusters for the different lakes in both Figures 38 and 39. Outside of three misclassified examples, the lakes were linearly separable in both cases, although all three confused examples came within the shallow portions of both lakes. It would make sense for the model to learn less with less data, as in Figures 40 and 41, within the task of brGDGT prediction; however, the clusters are much harder to make out, and it is much less clear whether the model was able to distinguish between the two lakes. This would aid in our disregard for the surprising cross-lake training results mentioned in 4.1.1. Although the model's predictive

ability seemed relatively strong given the task, it proves difficult to trust a seemingly effective model if the methods for getting there are circumspect and unverified on a simpler task.



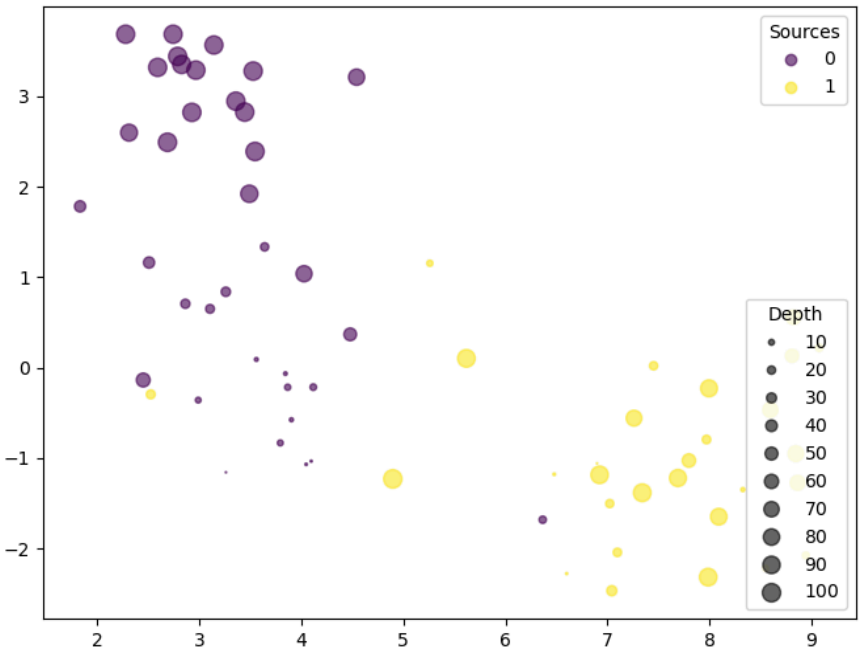
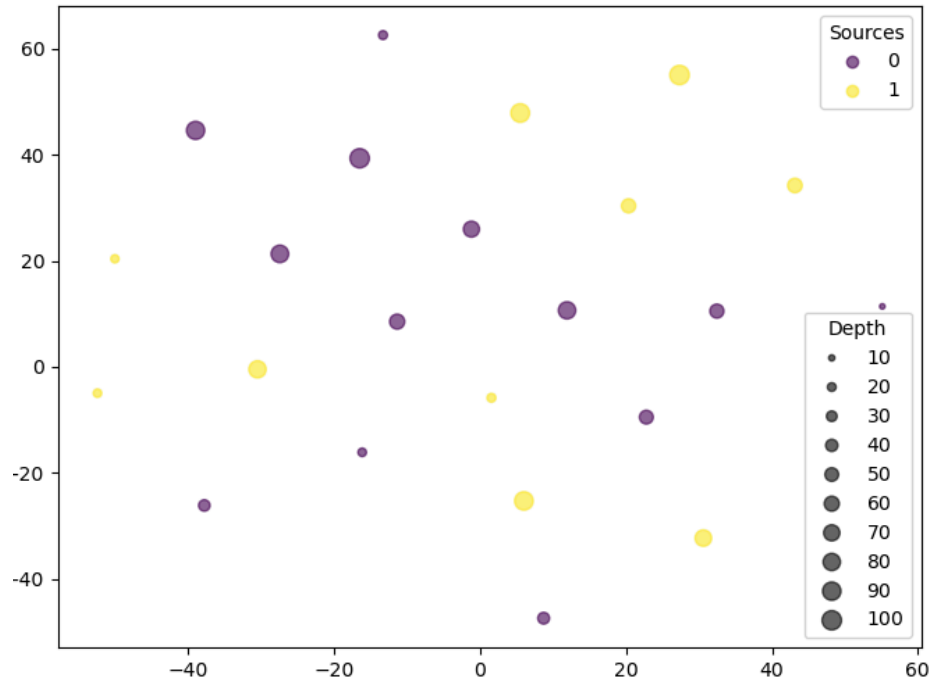


Figure 38: 5 cm %TOC t-SNE plot, with 0 denoting LVID and 1 denoting SVID

Figure 39: 1 cm %TOC t-SNE plot, with 0 denoting LVID and 1 denoting SVID



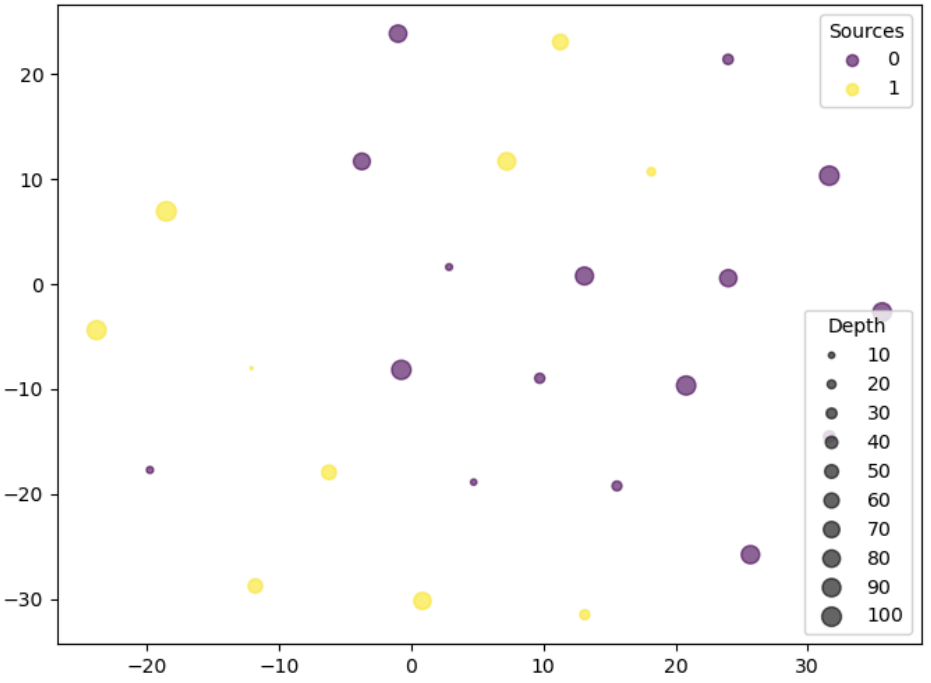
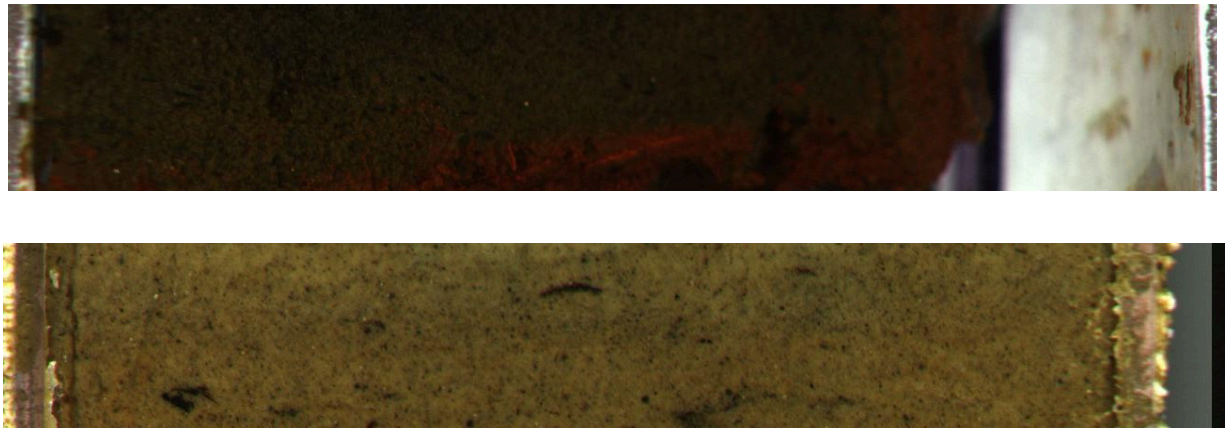


Figure 40: 5 cm brGDGT t-SNE plot, with 0 denoting LVID and 1 denoting SVID

Figure 41: 1 cm brGDGT t-SNE plot, with 0 denoting LVID and 1 denoting SVID

4.3.2 Manual Validation

The manual validation experiment was encouraging in that our model lined up with human opinion 95% of the time. All ten of the difficulties of the five centimeters chunks lined up with the corresponding RMSE levels. Eight out of the ten one-centimeter chunks were classified similarly. The two missed examples shared little in common visually. Figure 42 shows the first centimeter of LVID, one of two mislabeled samples. This measurement includes some obvious noise in the form of the filter at the bottom of the sample, which influenced the interpreter to label the sample as “hard;” nonetheless, the model was able to predict the amount of %TOC in this image with low RMSE. The second mislabeled sample comes from the 138th centimeter of SVID. In Figure 43, there is a noticeable spotty pattern in the sample. Before showing this example, three straight, somewhat similar examples were shown to be high, potentially influencing the interpreter to pick high again. This turned out to be a low RMSE example as well.



Figures 42 & 43 : First centimeter of LVID; 138th centimeter of SVID

(Displayed top to bottom)

This experiment was interesting. Even if we cannot definitively verify our model from this experiment, it indicates that the features being used by the model can to some extent be aligned with expert domain knowledge about the sediment cores. It is unclear exactly how the model is using these features, but we know it's doing something interesting. A manual labeler could not look at these image chunks and accurately predict %TOC from the image alone. The model can (sometimes), but the places where the model fails can be predicted by a researcher.

5 Discussion

When considering the results in their entirety, the experiments proved helpful in measuring the power of machine learning at understanding proxy amounts through image data. The results laid out in 4.1.1 and 4.2.1 point to the value of the model in an imputative task. The model showed promise at both the random disjoint and moving average prediction. The power in object detection and image comprehension leads to the possibility of an expansion of the use of the model for other image-related tasks in terms of either the recognition of large events or the extension of the current capabilities across more lakes.

Training across lakes seemed to be helpful, and this presents an opportunity to expand the dataset across more than only two lakes. It is exciting to consider the possibilities of the model if the dataset consisted of twenty lakes rather than two. Diminishing marginal returns would be expected from adding each successive lake; however, with an already seemingly effective model able to distinguish between the different lakes, it would be interesting and worthwhile to consider its performance over more lakes with varying distributions of data and images. This would present another

opportunity to examine the limits of our machine-learning approach and verify that our methods generalize outside of only Icelandic lakes.

The sequential and five-fold results covered in 4.1.2 and 4.2.2 emphasize the limit of the model with our current approach. A sequential task proves difficult in that the variation in both the image and proxy data means that by dividing the data into sequential chunks, the data can include values outside of the training domain area and images with events unseen anywhere outside of the test set. There is more to be done to attempt to surpass this limit. From a geoscience perspective, our model could be aided by the addition of other proxy values, depth measures, or other theory-based variables.

Future work could add to the model from a machine-learning perspective. When we consider the model as currently constructed, it is relatively straightforward. As mentioned, there is ample complexity to capture the general trends; however, there are still points, in both the sequential and random tasks where the model misses entirely. The model fluctuated between missing too low or too high but still captured the variance in some places. This presents the opportunity to use a method such as pinball loss or other techniques (91) to create a confidence interval for our predictions and determine which predictions we can have confidence in and which predictions we should disregard, as noted by the model.

Although the DeiT proved effective at the predictive task, it is logical to expect a more geoscience-specific ViT created for the task at hand would be much better at extracting the most important features without missing the nuance necessary to successfully solve our task. Rather than the few hundred-centimeter lakes, there are marine cores with hundreds of meters of data (92,93). By either implementing an

unsupervised approach (102) or predicting a related value from these cores, we could build a powerful embedding model more suited for our tasks using a dataset much richer than ImageNet and more abundant than our core dataset. In addition, this extra data could be used to pre-train our CNN to help it get in a better ballpark rather than starting from nothing and building up. It wouldn't be surprising to see the large spikes lessen or the performance at the boundaries improve after the stabilizing effect of the pre-training data (94).

6 Conclusion

This is not the pinnacle of image-related study for proxy prediction, but merely the frontier. Our methodology shows that proxy prediction is possible through using existing image data. Even with its limits, we found that the model could effectively increase the resolution of the data with sufficient accuracy. Most interestingly, the model seemed to be aided by the addition of the second lake, a fact that would not be expected through only geoscience theory due to the differences even between lakes close in physical distance. The power of the model could allow for greater core coverage by allowing researchers to spend less time meticulously labeling cores and rather only label the boundaries and some points throughout the core and allow the model to fill in the rest or increase analysis capabilities within a core due to the greater amount of data at a lower cost that this technique would provide. There is a plethora of data available on our Earth, and our research helps to go a little bit further in making the use of it possible and more accessible.

This methodology used limited data and complexity to reasonably tackle the problem at hand. With only a few hundred examples for learning, the model was able to

learn the trends in the data across both lakes. The model was not on the scale of some of the more advanced models we see in other domains of technology today and prediction for an individual sample only took milliseconds. In addition, the use of a pre-trained ViT limited the manual effort required to extract the features necessary and presented evidence for the use of domain-transfer techniques across other areas of geoscience and paleoclimatology.

Limits to machine learning within this domain were found with sequential training and testing. We challenged our model in prediction outside of its domain by presenting it with data outside of the data range within the lake and in a whole other lake entirely. There was some success, but overall, this presented a limit to the functionality of our model. There could be an improvement to be had through additional geoscience or machine learning techniques, and this problem provides an interesting playground to explore the synergy between machine learning and a grounded domain like geoscience.

The world around us is constantly changing and the techniques we use have to adapt alongside it to understand all of the changes our world is facing today. Paleoclimatology and geoscience provide an area for us to study the Earth and better understand the world and climate around us. Since we can't study it directly, the best way to understand the earth is through the resources and data we do have at our disposal. The better we can use the data we have, the better chance we have at maintaining a healthy world for our grandchildren and their grandchildren. Climate change is not going anywhere anytime soon and by understanding periods of temperature change in the past and how that affected ecosystems and biological life in the past, we can understand the effects of our actions today. This is by no means a solution to climate change but rather

provides the groundwork for a tool to understand our past to handle the unknown future in front of us.

7 Acknowledgements

A huge thanks needs to be given to those who helped create and compile the datasets used for this project. It would have been impossible to do any analysis without them. Thank you to the ILLUME Team: Gifford Miller, Áslaug Geirsdottir, Julio Sepúlveda, David Harning, Nicolò Ardenghi, Brooke Holmen, Thorvaldur Thordarson, and Jonathan Raberg, with funding from the National Science Foundation (grant no. OPP-1836981). In addition, this project would not have been possible without the help of my two advisors, Zach Wood-Doughty and Jonathan Raberg. Zach was essential in idea generation and a complete understanding of the machine learning methods used throughout the paper. Jon was instrumental in providing me with a paleoclimate domain expert to inform modeling decisions, verify the work being done, and provide input from a physical science perspective. Finally, this work was generously supported by the Office of Undergraduate Research at Northwestern University in the form of an AYURG grant.

8 Codebase

The code for the creation of this project can be found here. The codebase is currently incomplete as proprietary data is still to be published, and the scripts and plots will be updated once the data becomes available.

9 References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
2. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Beyer, L., Minderer, M., ... & Zhai, X. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
3. Chen, Junyu, et al. "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration." arXiv preprint arXiv:2104.06468 (2021).
4. Lin, F., Crawford, S., Guillot, K., Zhang, Y., Chen, Y., Yuan, X., ... & Tzeng, N. F. (2023). MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5774-5784).
5. Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine Learning for the Geosciences: Challenges and Opportunities. IEEE Transactions on Knowledge and Data Engineering, 31(8), 1544-1554.
6. Raberg, J. H., Miller, G. H., Geirsdóttir, Á., & Sepúlveda, J. (2022). Near-universal trends in brGDGT lipid distributions in nature. Science Advances, 8(20), eabm7625.
7. Fabijańska, A., Feder, A., & Ridge, J. (2020). DeepVarveNet: Automatic detection of glacial varves with deep neural networks. Computers & geosciences, 144, 104584.

8. Fang, M., & Li, X. (2019). An artificial neural networks‐based tree ring width proxy system model for paleoclimate data assimilation. Journal of Advances in Modeling Earth Systems, 11(4), 892-904.
9. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021, July). Training data-efficient image transformers & distillation through attention. In International conference on machine learning (pp. 10347-10357). PMLR.
10. Zimmermann, R. S., Klindt, D. A., & Brendel, W. (2024, March). Measuring Mechanistic Interpretability at Scale Without Humans. In ICLR 2024 Workshop on Representational Alignment.
11. Clarke, J., Gotoh, Y., & Goetze, S. (2023, December). Improving Audiovisual Active Speaker Detection in Egocentric Recordings with the Data-Efficient Image Transformer. In 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 1-8). IEEE.
12. Naafs, B. D. A., Inglis, G. N., Zheng, Y., Amesbury, M. J., Biester, H., Bindler, R., ... & Pancost, R. D. (2017). Introducing global peat-specific temperature and pH calibrations based on brGDGT bacterial lipids. Geochimica et Cosmochimica Acta, 208, 285-301.
13. Charbonnier, G., Adatte, T., Föllmi, K. B., & Suan, G. (2020). Effect of intense weathering and postdepositional degradation of organic matter on Hg/TOC proxy in organic‐rich sediments and its implications for deep‐time investigations. Geochemistry, Geophysics, Geosystems, 21(2), e2019GC008707.

14. Broadman, E., Kaufman, D. S., Henderson, A. C., Berg, E. E., Anderson, R. S., Leng, M. J., ... & Muñoz, S. E. (2020). Multi-proxy evidence for millennial-scale changes in North Pacific Holocene hydroclimate from the Kenai Peninsula lowlands, south-central Alaska. Quaternary Science Reviews, 241, 106420.
15. Geirsdóttir, Á., Harning, D. J., Miller, G. H., Andrews, J. T., Zhong, Y., & Caseldine, C. (2020). Holocene history of landscape instability in Iceland: Can we deconvolve the impacts of climate, volcanism and human activity?. Quaternary Science Reviews, 249, 106633.
16. Das, J., Woskie, L., Rajbhandari, R., Abbasi, K., & Jha, A. (2018). Rethinking assumptions about delivery of healthcare: implications for universal health coverage. Bmj, 361.
17. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347-1358.
18. Sayyadi, M., Collina, L., & Provitera, M. J. (2023). The End of Management Consulting as We Know it?. Management Consulting Journal, 6(2), 67-77.
19. Ransbotham, S., Kiron, D., Candelon, F., Khodabandeh, S., & Chu, M. (2022). Achieving individual—and organizational—value with AI. MIT Sloan Management Review.
20. Kamaruddin, R. I. (2023). ChatGPT and the Future of Management Consulting: Opportunities and Challenges Ahead (Doctoral dissertation, Massachusetts Institute of Technology).

21. Vijayakumar, H. (2021). The Impact of AI-Innovations and Private AI-Investment on US Economic Growth: An Empirical Analysis. Reviews of Contemporary Business Analytics, 4(1), 14-32.
22. Kouw, W. M., & Loog, M. (2018). An introduction to domain adaptation and transfer learning. arXiv preprint arXiv:1812.11806.
23. Ling, X., Dai, W., Xue, G. R., Yang, Q., & Yu, Y. (2008, August). Spectral domain-transfer learning. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 488-496).
24. Flavián, C., & Casaló, L. V. (2021). Artificial intelligence in services: current trends, benefits and challenges. The Service Industries Journal, 41(13-14), 853-859.
25. Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. Journal of Consumer Research, 46(4), 629-650.
26. Taylor, C., & Nakhaeizadeh, G. (1997). Learning in dynamically changing domains: theory revision and context dependence issues. In Machine Learning: ECML-97: 9th European Conference on Machine Learning Prague, Czech Republic, April 23–25, 1997 Proceedings 9 (pp. 353-360). Springer Berlin Heidelberg.
27. Camps-Valls, G., Verrelst, J., Munoz-Mari, J., Laparra, V., Mateo-Jimenez, F., & Gomez-Dans, J. (2016). A survey on Gaussian processes for earth-observation data analysis: A comprehensive investigation. IEEE Geoscience and Remote Sensing Magazine, 4(2), 58-78.

28. TENG Ji-Wen, SI Xiang, WANG Qian-Shen, ZHANG Yong-Qian, YANG Hui. Collation and stipulation of the core science problems and theoretical concept in the geoscience study on the Tibetan plateau[J]. Chinese Journal of Geophysics (in Chinese), 2015, 58(1): 103-124, doi: 10.6038/cjg20150109
29. Oliver, M., Webster, R., & Gerrard, J. (1989). Geostatistics in Physical Geography. Part I: Theory. Transactions of the Institute of British Geographers, 14(3), 259–269. https://doi.org/10.2307/622687
30. Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., ... & Zhou, B. (2021). Climate change 2021: the physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, 2(1), 2391.
31. De Lima, R. P., Marfurt, K., Duarte, D., & Bonar, A. (2019, June). Progress and challenges in deep learning analysis of geoscience images. In 81st EAGE Conference and Exhibition 2019 (Vol. 2019, No. 1, pp. 1-5). European Association of Geoscientists & Engineers.
32. P. M. Caldwell, C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer and B. M. Sanderson, "Statistical significance of climate sensitivity predictors obtained by data mining", Geophysical Res. Lett., vol. 41, no. 5, pp. 1803-1808, 2014.
33. D. Lazer, R. Kennedy, G. King and A. Vespignani, "The parable of Google flu: Traps in big data analysis", Science, vol. 343, no. 6176, pp. 1203-1205, Mar. 2014, [online] Available: http://www.ncbi.nlm.nih.gov/pubmed/24626916.
34. G. Marcus and E. Davis, "Eight (no nine!) problems with big data", New York Times, vol. 6, no. 4, 2014.

35. University of Minnesota, “Global surface water monitoring system,” 2017. [Online]. Available: http://z.umn.edu/watermonitor/
36. A. Khandelwal, V. Mithal, and V. Kumar, “Post classification label refinement using implicit ordering constraint among data instances,” in Proc. IEEE Int. Conf. Data Mining, 2015, pp. 799–804.
37. A. Karpatne, W. Watkins, J. Read, and V. Kumar, “Physicsguided neural networks (PGNN): An application in lake temperature modeling,” arXiv: 1710.11431, 2017.
38. J. Kawale et al., "A graph-based approach to find teleconnections in climate data", Statist. Anal. Data Mining, vol. 6, no. 3, pp. 158-179, 2013.
39. Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Transactions on knowledge and data engineering, 29(10), 2318-2331.
40. Madiajagan, M., & Raj, S. S. (2019). Parallel computing, graphics processing unit (GPU) and new hardware for deep learning in computational intelligence research. In Deep learning and parallel computing environment for bioengineering systems (pp. 1-15). Academic Press.
41. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics, 36(4), 193-202.
42. Oh, K. S., & Jung, K. (2004). GPU implementation of neural networks. Pattern Recognition, 37(6), 1311-1314.

43. Haykin (2008) Neural Networks and Learning Machines, 3rd edition
44. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318–362. MIT Press, Cambridge (1986)
45. Denker, J., Gardner, W., Graf, H., Henderson, D., Howard, R., Hubbard, W., ... & Guyon, I. (1988). Neural network recognizer for hand-written zip code digits. Advances in neural information processing systems, 1.
46. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.
47. Schmidhuber, J. (2022). Annotated history of modern ai and deep learning. arXiv preprint arXiv:2212.11279.
48. O'shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
49. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
50. Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society, 8(2), 20539517211035955.
51. Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., ... & Vapnik, V. (1994, October). Comparison of classifier methods: a case study in handwritten digit recognition. In Proceedings of the 12th IAPR International

Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5) (Vol. 2, pp. 77-82). IEEE.

52. LeCun, Y., Cortes, C., & Burges, C. (2010). MNIST handwritten digit database.
53. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.
54. Cai, Y., Zhou, Y., Han, Q., Sun, J., Kong, X., Li, J., & Zhang, X. (2022). Reversible column networks. arXiv preprint arXiv:2212.11696.
55. Liang, C. (2024). Lion. GitHub repository, https://github.com/google/automl/blob/master/lion/README.md
56. Gesmundo, A., & Dean, J. (2022). An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems. arXiv preprint arXiv:2205.12755.
57. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
58. Lee, G. G., Shi, L., Latif, E., Gao, Y., Bewersdorf, A., Nyaaba, M., ... & Zhai, X. (2023). Multimodality of ai for education: Towards artificial general intelligence. arXiv preprint arXiv:2312.06037.
59. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917.
60. Tayyab, B. U., & Chua, N. (2021). Pre-training transformers for domain adaptation. arXiv preprint arXiv:2112.09965.

61. Meng, J., Tan, Z., Yu, Y., Wang, P., & Liu, S. (2022). TL-med: A Two-stage transfer learning recognition model for medical images of COVID-19. biocybernetics and biomedical engineering, 42(3), 842-855.
62. Lyu, Y., Yu, X., Zhu, D., & Zhang, L. (2022, June). Classification of alzheimer's disease via vision transformer: Classification of alzheimer's disease via vision transformer. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments (pp. 463-468).
63. Tierney, J. E., Poulsen, C. J., Montañez, I. P., Bhattacharya, T., Feng, R., Ford, H. L., ... & Zhang, Y. G. (2020). Past climates inform our future. science, 370(6517), eaay3701.
64. Yan, H., Sun, L., Wang, Y., Huang, W., Qiu, S., & Yang, C. (2011). A record of the Southern Oscillation Index for the past 2,000 years from precipitation proxies. Nature Geoscience, 4(9), 611-614.
65. Scafetta, N. (2023). Empirical assessment of the role of the Sun in climate change using balanced multi-proxy solar records. Geoscience Frontiers, 14(6), 101650.
66. Guillet, S., Corona, C., Stoffel, M., Khodri, M., Lavigne, F., Ortega, P., ... & Oppenheimer, C. (2017). Climate response to the Samalas volcanic eruption in 1257 revealed by proxy records. Nature geoscience, 10(2), 123-128.
67. Von Gunten, L., D'Andrea, W. J., Bradley, R. S., & Huang, Y. (2012). Proxy-to-proxy calibration: Increasing the temporal resolution of quantitative climate reconstructions. Scientific reports, 2(1), 609.

68. Fabijańska, A., & Danek, M. (2018). DeepDendro–A tree rings detector based on a deep convolutional neural network. Computers and electronics in agriculture, 150, 353-363.
69. Raberg et al 2023 Arctic Data Center doi:10.18739/A26688K7F
70. Weijers, J. W., Schouten, S., Hopmans, E. C., Geenevasen, J. A., David, O. R., Coleman, J. M., ... & Sinninghe Damsté, J. S. (2006). Membrane lipids of mesophilic anaerobic bacteria thriving in peats have typical archaeal traits. Environmental Microbiology, 8(4), 648-657.
71. Weijers, J. W., Schouten, S., van den Donker, J. C., Hopmans, E. C., & Damsté, J. S. S. (2007). Environmental controls on bacterial tetraether membrane lipid distribution in soils. Geochimica et Cosmochimica Acta, 71(3), 703-713.
72. Harning, D. J., Curtin, L., Geirsdóttir, Á., D'Andrea, W. J., Miller, G. H., & Sepúlveda, J. (2020). Lipid biomarkers quantify Holocene summer temperature and ice cap sensitivity in Icelandic lakes. Geophysical Research Letters, 47(3), e2019GL085728.
73. Axford, Y., Miller, G. H., Geirsdóttir, Á., & Langdon, P. G. (2007). Holocene temperature history of northern Iceland inferred from subfossil midges. Quaternary Science Reviews, 26(25-28), 3344-3358.
74. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.
75. Wei, L., Xiao, A., Xie, L., Zhang, X., Chen, X., & Tian, Q. (2020, August). Circumventing outliers of autoaugment with knowledge distillation. In European

Conference on Computer Vision (pp. 608-625). Cham: Springer International Publishing.

76. Sahoo, S. (2018, August). Deciding optimal kernel size for CNN. Towards Data Science. https://towardsdatascience.com/deciding-optimal-filter-size-for-cnns-d6f7b56f9363

77. Goddéris, Y., Donnadieu, Y., Carretier, S., Aretz, M., Dera, G., Macouin, M., & Regard, V. (2017). Onset and ending of the late Palaeozoic ice age triggered by tectonically paced rock weathering. Nature Geoscience, 10(5), 382-386.

78. Abram, N. J., Gagan, M. K., Cole, J. E., Hantoro, W. S., & Mudelsee, M. (2008). Recent intensification of tropical climate variability in the Indian Ocean. Nature Geoscience, 1(12), 849-853.

79. Rothwell, R. G., & Croudace, I. W. (2015). Twenty years of XRF core scanning marine sediments: what do geochemical proxies tell us?. Micro-XRF Studies of Sediment Cores: Applications of a non-destructive tool for the environmental sciences, 25-102.

80. H. Shimodaira, “Improving predictive inference under covariate shift by weighting the log-likelihood function,” J. Stat. Plan. Inference, vol. 90, no. 2, pp. 227– 244, 2000.

81. A. Torralba and A. A. Efros, “Unbiased look at data set bias,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2011, pp. 1521–1528.

82. Patel, V. M.; Gopalan, R.; Li, R.; and Chellappa, R. 2015. Visual domain adaptation: A survey of recent advances.IEEE Signal Processing Magazine 32(3):53–69

83. Cirac, G., Farfan, J., Avansi, G. D., Schiozer, D. J., & Rocha, A. (2023). Cross-Domain Feature learning and data augmentation for few-shot proxy development in oil industry. Applied Soft Computing, 149, 110972.

84. Li, T., Zuo, R., Zhao, X., & Zhao, K. (2022). Mapping prospectivity for regolith-hosted REE deposits via convolutional neural network with generative adversarial network augmented data. Ore Geology Reviews, 142, 104693.

85. A. Mignan and M. Broccardo. A deeper look into ‘deep learning of aftershock patterns following large earthquakes’: Illustrating first principles in neural network physical interpretability. In International Work-Conference on Artificial Neural Networks, pages 3–14. Springer, 2019a.

86. Mignan, A., & Broccardo, M. (2019). One neuron versus deep learning in aftershock prediction. Nature, 574(7776), E1-E3.

87. Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. Advances in geophysics, 61, 1-55.

88. Puetz, S. J., Condie, K. C., Sundell, K., Roberts, N. M., Spencer, C. J., Boulila, S., & Cheng, Q. (2024). The replication crisis and its relevance to Earth Science studies: Case studies and recommendations. Geoscience Frontiers, 101821.

89. R. Shah and L. Innig. Aftershock issues, 2019. Github Repository https://github.com/rajshah4/aftershocks_issues.

90. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).
91. Chung, Y., Neiswanger, W., Char, I., & Schneider, J. (2021). Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. Advances in Neural Information Processing Systems, 34, 10971-10984.
92. IODP core repositories. IODP. (n.d.). https://www.iodp.org/resources/core-repositories
93. Columbia University. (n.d.). Collections | Lamont-Doherty Core Repository. Lamont-Doherty Earth Observatory. https://corerepository.ldeo.columbia.edu/content/collections
94. Huang, S. F., Chuang, S. P., Liu, D. R., Chen, Y. C., Yang, G. P., & Lee, H. Y. (2020). Stabilizing label assignment for speech separation by self-supervised pre-training. arXiv preprint arXiv:2010.15366.
95. Romero, D. W., Bruintjes, R. J., Tomczak, J. M., Bekkers, E. J., Hoogendoorn, M., & van Gemert, J. C. (2021). Flexconv: Continuous kernel convolutions with differentiable kernel sizes. arXiv preprint arXiv:2110.08059.
96. Forbus, K. D. (2011). Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(4), 374-391.
97. Hamilton, L., Elliott, D., Quick, A., Smith, S., & Choplin, V. (2023). Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. *International journal of qualitative methods*, *22*, 16094069231201504.
98. Faltings, B., & Sun, K. (1996). FAMING: Supporting innovative mechanism shape design. *Computer-Aided Design*, *28*(3), 207-216.

99. Qi, S., Lee, Z., Liu, J., Qin, Y., Du, Q., & Han, M. (2024). ResX: feature extraction block for medical image segmentation. *IEEE Access*.

100. Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., ... & Huang, T. (2011, June). Large-scale image classification: Fast feature extraction and SVM training. In *CVPR 2011* (pp. 1689-1696). IEEE.

101. De Jonge, C., Hopmans, E. C., Zell, C. I., Kim, J. H., Schouten, S., & Damsté, J. S. S. (2014). Occurrence and abundance of 6-methyl branched glycerol dialkyl glycerol tetraethers in soils: Implications for palaeoclimate reconstruction. *Geochimica et Cosmochimica Acta*, *141*, 97-112.

102. Li, J., Savarese, S., & Hoi, S. C. (2022). Masked unsupervised self-training for label-free image classification. *arXiv preprint arXiv:2206.02967*.