How ViTs scale on seismic data: a study on model and data trade-offs

Altay Sansal, Ben Lasscock, and Alejandro Valenciano, TGS*

Summary

Pre-trained seismic foundation models (SFMs) have shown promising performance in seismic interpretation tasks. They demonstrate effective generalization across various geographic areas. However, the impact of dataset size and model complexity on seismic data applications remains underexplored. Understanding these relationships is vital for optimizing model performance and improving geological interpretation accuracy in practical seismic applications. We systematically assess the effects of dataset size and model complexity on seismic data by training multiple Vision Transformer (ViT) variants using the Masked Auto Encoder (MAE) technique. We benchmark these models using a few-shot facies classification task with the established LANDMASS1 dataset. Our analysis reveals clear scaling metrics, demonstrating performance enhancements with larger models and datasets consistent with scaling laws noted in other domains. These insights offer actionable guidelines for larger 3D models and datasets.

Introduction

Vision Transformer (ViT) models (Dosovitskiy, 2020) are the latest advancement in computer vision algorithms. He et al. (2022) demonstrate that by utilizing a Masked Auto Encoder (MAE) architecture, we can train ViTs in a self-supervised manner to learn meaningful representations of natural images for use in object detection, segmentation, and classification tasks with minimal fine-tuning.

Since the original Transformer (Vaswani et al., 2017) was introduced, researchers have studied the so-called “Scaling Laws” for these models across various domains and modalities. Hoffmann et al. (2022) investigated these laws in the language modeling domain, while Zhai et al. (2022) examined the scaling laws for ViT models with up to 2 billion parameters. Both studies demonstrate that they achieve state-of-the-art results by increasing data size and model parameters. Building on that research, Zhai et al. (2023) trained a ViT model with 22 billion parameters and discussed their engineering challenges.

Lasscock et al. (2024), Gao et al. (2024), and Sheng et al. (2025) explored SFMs and downstream task applications. These efforts were conducted on a small scale. The largest and most data-intensive published SFMs are 3D ViTs trained on proprietary data with 632 million and 1.8 billion parameter models (Sansal et al., 2025a and 2025b). However, whether we are training optimal foundation models for given seismic data remains unclear.

In this study, we explore aspects of scaling laws on SFM by training various ViT size variants using the MAE technique and benchmarking the performance of facies classification tasks with the LANDMASS1 (Alaudah and AlRegib, 2015) dataset under a {one, two, five, ten, twenty}-shot linear probing setting. This approach helps us understand how model and dataset size influence downstream task performance. We conduct this study using 2D images (cross-sections of 3D seismic data) to transfer these insights into training 3D ViTs in future research. Conducting a similar study in 3D increases each model's training time and, consequently, the computational cost.

Dataset

We utilize a large corpus of proprietary data to pre-train our variant of Seismic Foundation Models, termed *SeisFM*. The datasets consist of depth-migrated seismic data, covering a surface area of approximately 420,000 km² (19 terabytes). We sample non-overlapping 224×224 seismic images from carefully crafted regions of interest (ROIs) in both inline and crossline directions. The ROIs are chosen to minimize or avoid empty patches (e.g., water) and uninformative, deep, noisy areas of seismic data. In total, our dataset contains 164 million samples. We decimate the dataset in both inline and crossline directions by a fixed factor for dataset size tests. Figure 1 illustrates the geographic coverage of the pre-training data, and Table 1 details the dataset sizes and the decimation used to achieve the desired pre-training dataset size variants.

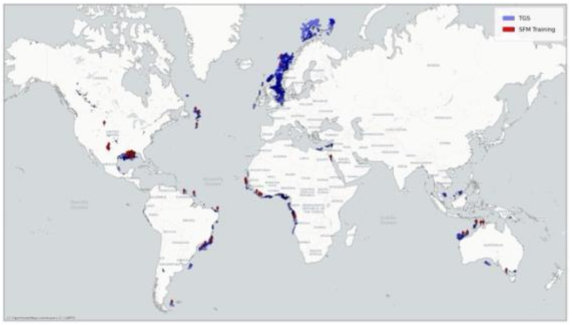

Figure 1: Map showing worldwide available data (blue) and the ones used in *SeisFM* training (red).

For the downstream task of facies classification, we utilized the LANDMASS1 dataset. This dataset forms part of the Large North Sea Dataset of Migrated Aggregated Seismic Structures (LANDMASS) and consists of 17,667 small seismic image patches (99×99 pixels) from post-migrated seismic volumes. Developed by the Center for Energy and Geo Processing (CeGP), it includes four classes: 9,385 Horizon, 5,140 Chaotic, 1,251 Fault, and 1,891 Salt Dome patches. The LANDMASS dataset is derived from the Dutch F3 dataset. While the dataset exhibits limitations in geographic coverage and ambiguity in the classifications of “chaotic” and “fault,” it remains valuable for benchmarking our *SeisFM* models. Figure 2 displays example images.

Table 1: Datasets and corresponding dataset decimation factors in inline and crossline directions. Decimation subsamples the original lines by taking every nth line.

# Samples	IL/XL Decimation
0.1 million	1500
1 million	160
10 million	16
41 million	4

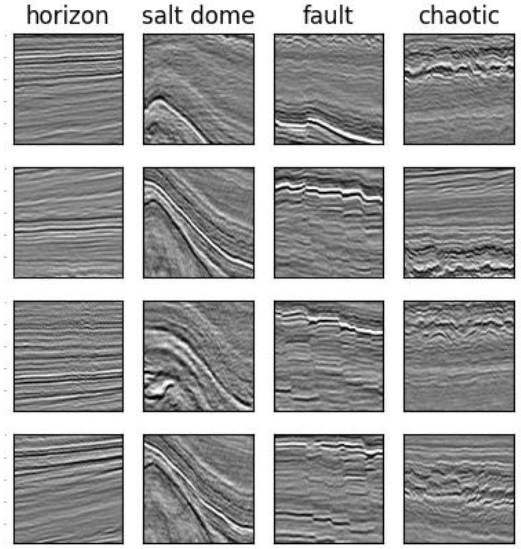

Figure 2: Example tiles and labels from LANDMASS 1 dataset for all four classes in columns.

Method

The ViT-MAE method involves training a vision encoder model using unlabeled data. It works by masking about 75% of the images and using the unmasked patches to reconstruct the remaining parts of the image in pixel space. This task is quite challenging, allowing the model to learn meaningful representations of the data through patch and image features. For our experiments, we trained the models listed in Table 2. We follow the sizing convention of ViT models and pre-train all models for 1600 epochs. The hyperparameters for pre-training are described in Table 3.

For linear probing, we took the pre-trained encoder and froze its weights. On top of the encoder, we add a batch normalization to the embeddings and then a linear head that maps the image class **[CLS]** token embeddings to four classes in the LANDMASS1 dataset. The batch normalization helps select stable hyperparameters (e.g., learning rate) for various k-shot configurations. For the loss function, we utilize simple single-label cross-entropy.

Table 2: Pre-trained model configurations for scaling analysis. Each model configuration is trained with 0.11 million, 1 million, 10 million, and 41 million dataset size variants with 12 models.

Model Name	Num. Param	Hidden Size	MLP Size	ViT Layers	Atten. Heads
SeisFM-Ti	5.7M	192	768	12	3
SeisFM-S	21M	384	1536	12	6
SeisFM-B	85M	768	3072	12	12

We utilized few-shot fine-tuning to measure performance. For instance, one-shot means we take one example from each class (4 images) and train the head. Twenty-shot means we take 80 total images for fine-tuning the classification head. As mentioned earlier, the LANDMASS1 dataset has over 17 thousand samples. We first divide the dataset into training (3%), validation (10%), and testing (87%) subsets. Then we sample the training split for {1, 2, 5, 10, 20}-shot datasets. The validation and testing datasets are kept consistent between every model evaluation. Depending on the number of examples (N-shot) in fine-tuning, we adjust the number of epochs to between 100 and 150 and the batch size to between 4 and 16. The rest of the hyperparameters are given in Table 4. We adjust the learning rate and batch size for stable training based on the number of examples.

Table 3: Pre-training setting.

config	***value***
optimizer	AdamW
base learning rate	1e-4
weight decay	0.05
batch size	2048
learning rate schedule	cosine
warmup epochs	20
augmentation	Flip, RandomResizedCrop

Table 4: Linear probing setting.

config	***value***
optimizer	AdamW
base learning rate	Varies ~ [0.0005, 0.001]
weight decay	0
batch size	varies ~ [2, 16]
learning rate schedule	cosine
warmup epochs	Varies ~ [8, 14]
augmentation	Flip, RandomResizedCrop

Our performance metrics for linear probing are the F1 score and Top-1 Error Rate, calculated on LANDMASS1 test data facies predictions achieved through the few-shot fine-tuning methodology described above. We use positional encoding interpolation to enable training on 99×99 images with a ViT encoder that was pre-trained on 224×224 data patches (Chen et al., 2023).

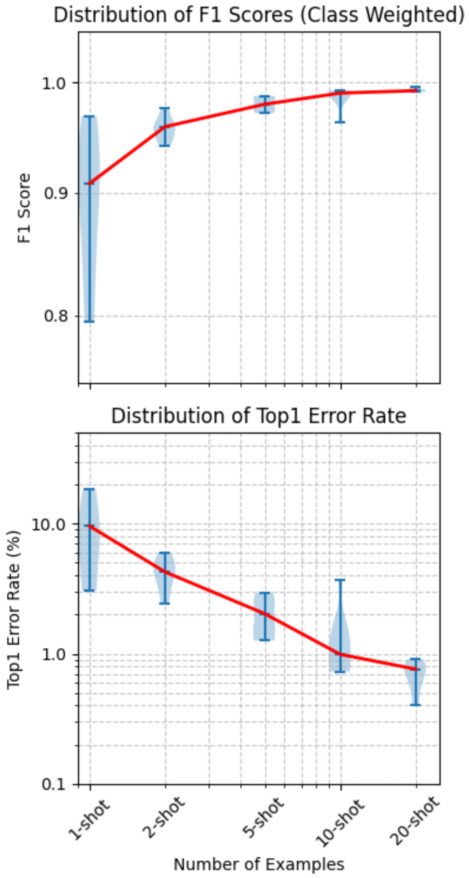

Figure 3: Few-shot classification F1 score (top) and Top-1 Error Rate (bottom) and their distribution across five runs. In this chart, we keep the model size at *SeisFM-B-1M*.

Results

Similar to other domains, we find that increasing both the dataset size and the model parameter count (and thus the computational requirements) enhances the few-shot performance of the model. Figure 3 illustrates two metrics used for benchmarking: the F1 score (which measures how well a model predicts classes, higher=better) and the Top-1 Error Rate (the percentage of the model incorrectly selecting the top answer, lower=better). We conducted each experiment seven times to minimize sensitivity to model initialization and data sampling to reduce scoring sensitivity to data sampling and training dynamics. The blue-shaded areas depict the distributions for the *SeisFM-B-1M* model for all experiments. As expected, performance improves with more examples per class. Nonetheless, the one-shot performance is impressive. The broader distribution (in blue) for one- and two-shot linear probing suggests that the performance of the supervised classification head depends on the samples it received during training and is more sensitive to the quality of examples. Small parameter counts and less training data make the models more prone to this issue.

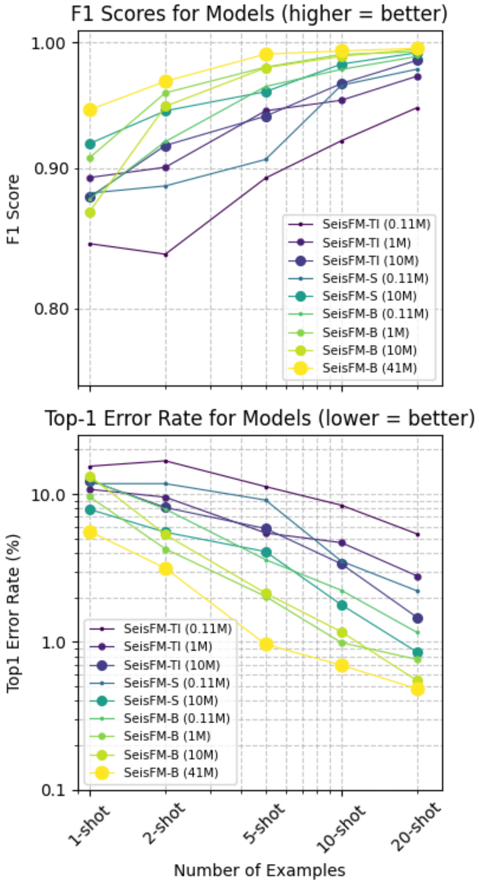

Figure 4: Few-shot classification performance (median of runs) with various model and dataset size configurations.

We also benchmarked the effect of model size using the other SeisFM variants with the same method. Figure 4 presents a selection of results alongside the F1 and Top-1 Error Rate. The color change represents model size (from blue: small to yellow: large). The size of the markers reflects the amount of data used in pre-training each specific model highlighted in the legend. We notice that few-shot classification performance improves as both the model and dataset sizes increase. Evidence suggests that the models can scale further with additional data and larger sizes.

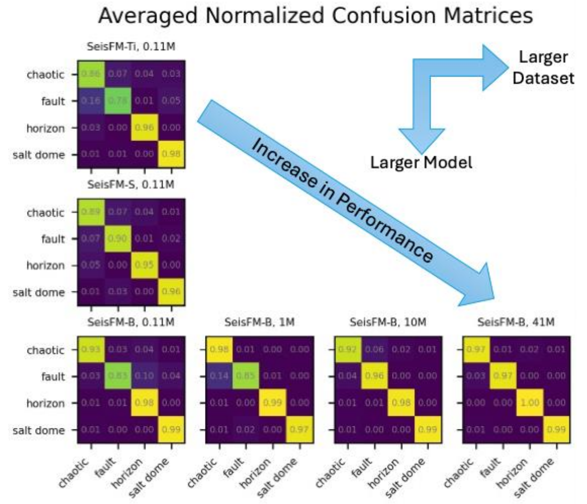

Figure 5: Confusion matrix for facies classification on test data. The model size increases top-to-bottom, and the dataset size increases left-to-right.

To validate the per-class results, the confusion matrices (which show how accurately a model predicts facies, both correctly and incorrectly) for different model and data combinations are presented in Figure 5. To address the class imbalance in the LANDMASS1 dataset, the metrics from the few-shot runs were normalized based on the true labels. This normalization is essential for averaging the results across various few-shot samples and creating aggregated confusion matrices. The results suggest that increasing the dataset and model sizes positively impacts performance.

Figure 6 shows good and bad predictions in the testing dataset. The predictions are from the largest model (*SeisFM-B-41M).* Bad examples are rare (F1: 0.998) and explainable (e.g., class mixtures or faults/chaotic being similar).

Conclusions

We have demonstrated that the size of the model and dataset are crucial for scaling seismic foundation models (SFMs). This was illustrated through systematic training and evaluation of pre-trained *SeisFM* models using the Masked AutoEncoder (MAE) technique and by benchmarking their linear probing performance on a facies classification task with the LANDMASS1 dataset. Our findings validate that the "scaling laws" observed in other fields, such as language modeling and computer vision, also apply to foundation models trained on seismic data.

The *SeisFM* models achieved notable effectiveness even in one-shot classification settings, with the most-scaled model (*SeisFM-B* trained on 41M samples) reaching an impressive F1 score of 0.998. This capability highlights a valuable use case, as limited labeled data can generate additional labels, which, upon human validation, can subsequently be utilized for further supervised training. However, we observed increased sensitivity to the quality of samples chosen during one-shot training, evidenced by broader error distributions across multiple runs. Adopting strategies involving five-shot learning or higher significantly mitigates this sensitivity, yielding consistently high performance. Therefore, we recommend an iterative approach to label curation, progressively enhancing training data quality through repeated cycles of model inference and human validation.

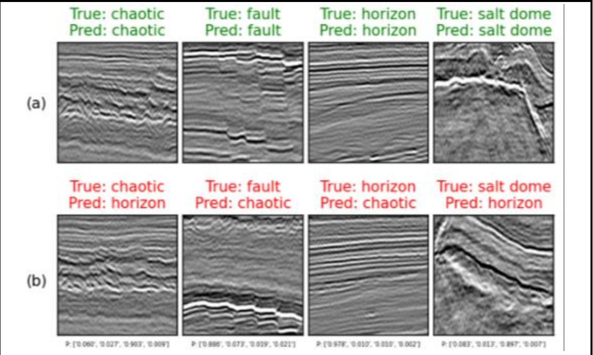

Figure 6: Correct (a) and erroneous (b) results in test data.

Our analysis, supported by Figure 4, shows consistent performance gains up to 41M samples and larger models, with no clear evidence of saturation, suggesting that further improvements are attainable with increased scale. This finding closely aligns with scaling laws observed in other domains. While these conclusions directly apply to our current 2D seismic analysis, they emphasize the significance of ongoing investment in larger-scale datasets and advanced model architectures to drive future innovations in artificial intelligence for seismic interpretation. Future research should focus on extending these insights to 3D seismic data contexts to confirm and refine these scaling trends, potentially providing more profound and comprehensive geological insights.

Acknowledgments

We thank TGS for providing the pre-training data and computational resources. We also thank the Center for Energy and Geo Processing (CeGP) and the OLIVES lab for making the LANDMASS1 dataset publicly available.