Scaling seismic foundation models

Altay Sansal[1]*, Ben Lasscock[1] and Alejandro Valenciano[1] address the complexities of large-scale training of a seismic foundation model on a global dataset of 63 seismic volumes and leverage a cloud-native, digitalised seismic data infrastructure to address the data engineering challenges, avoiding duplication.

Abstract

Traditional workflows using machine learning interpretation of seismic data rely on iterative training and inference on single datasets, producing models that fail to generalise beyond their training domain. Self-supervised training and scaling of 3D vision transformer (ViT) architectures enables seismic interpretation with improved generalisation across diverse datasets. We address the complexities of large-scale training on a global dataset of 63 seismic volumes using the masked autoencoder (MAE) architecture with the ViT-H model consisting of 660 million parameters. We leverage a cloud-native, digitalised seismic data infrastructure to address the data engineering challenges, avoiding duplication. For a downstream task, a salt segmentation model trained using interpretation labels from the Gulf of Mexico and Brazil demonstrated zero-shot generalisation on a West African survey. These findings underscore the potential of pre-trained foundation models to overcome the limitations of iterative approaches and extend seismic interpretation across diverse basins, marking a significant advancement in scalable machine learning for subsurface challenges.

Introduction

The pre-trained ViT-MAE model is an emerging technology in seismic processing and interpretation [Lasscock 2024, Sheng 2023]. Much like how large language models have been a step change in natural language processing, there is potential for this new way of approaching AI to disrupt geophysical applications. Until now, these studies have been applied to small, open-source datasets with synthetic data and older seismic imaging and processing techniques. [Ordonez 2024] reported an expansive study that high-graded a subset of 60,000 2D crops for pretraining from a larger 20 survey dataset. In each case, these studies have demonstrated the efficacy of pre-training a seismic foundation model (SFM) and then using or fine-tuning it on various downstream tasks, including seismic salt and facies classification.

The highly scalable characteristics of the ViT-MAE technology, mainly when applied in 3D [Lasscock 2024], have yet to be explored in geophysical literature. In computer vision, it has been established [Zhai 2022] that larger models pre-trained on large datasets (ImageNet-21k and JFT-300M) achieve better performance in image classification tasks. This study aims to tackle the problem of scaling ViT-MAE models trained on seismic data to a global corpus of 63 seismic surveys. And evaluate if a subsequent downstream task can be efficiently fine-tuned from these large pre-trained models to outperform existing AI methods regarding their generalisation capacity.

As we train large models, data management becomes a crucial enabling technology, both in need of exploring and curating such a large corpus of data and efficiently saturating large clusters of GPU computing required to train them in a timely manner. Tracking this problem on seismic data presents unique challenges. We will explain how cloud object storage and the MDIO seismic data format [Sansal 2023] are used efficiently in pretraining a 660M parameter 3D seismic ViT-H model. We will address the model's usefulness by fine-tuning it for salt interpretation. The salt interpretation model builds on our SaltNet dataset, consisting of interpretation from 23 seismic volumes, and we will compare model IoU scores with existing state-of-the-art 2D and 3D U-Net models [Warren 2023, Roberts 2024].

Methodology

Model Architecture

The model architecture shown schematically in Figure 1 is based on a Masked Autoencoder (MAE) with a Vision Transformer (ViT) backbone, as described by He et al. [2021], modified to process 3D seismic volumes. Input seismic data is divided into overlapping mini-cubes, which undergo augmentations such as inline/crossline flips. The model adapts the ViT-MAE design initially created for 2D images to 3D, projecting 16^3 patches (visual tokens) to a collection of 1280-length vector embeddings. At each training step, a batch of mini-cubes is selected from the global dataset, 90% of the patches are masked, and the remaining 10% is used to reconstruct the original mini-cube. The learning objective is the pixel space reconstruction accuracy of the masked patches using the mean-squared error (MSE) metric.

One advantage of this self-supervised training is that it is memory-efficient since all the data is used only in the small decoder of the model loss. The large encoder only has to propagate 10% of the patches. This approach is highly scalable to large model sizes without complex distributed training techniques. An advantage of working in the 3D domain of the data over 2D is that a more significant percentage of the data, 90% in this case

[1] TGS
* Corresponding author, E-mail: altay.sansal@tgs.com
DOI: 10.3997/1365-2397.fb2025016

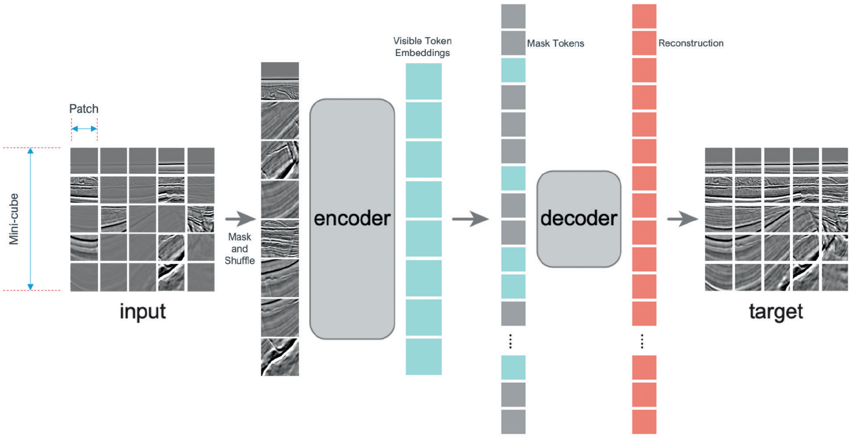

Figure 1 A modified schematic view that explains the ViT-MAE pre-training concept [He 2021] is shown in the picture. Large 3D data patches are loaded in batches, 90% of the data is discarded, and the remaining 10% is used to reconstruct the original data from the mask tokens.

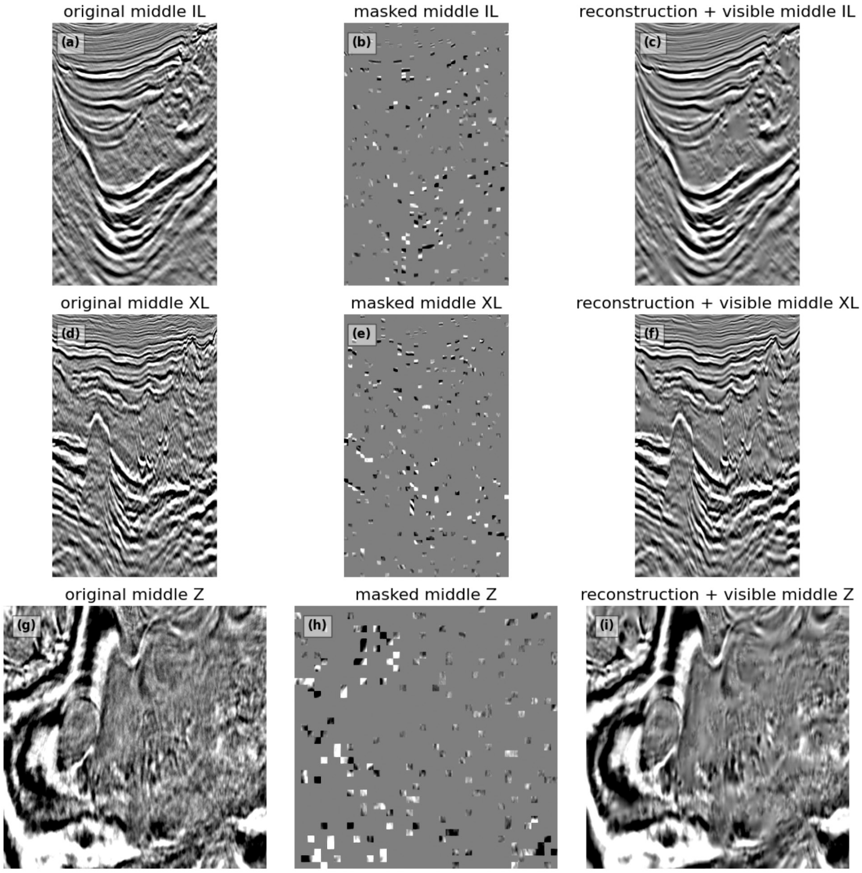

Figure 2 A specific example of a sampled 640x640x1024 mini-cube and its reconstruction. (a-c) A mid-point inline slice through the 3D patch showing the original data, the data used in reconstruction, and the reconstructed 3D patch. (d-f) and (g-i) show the equivalent crossline and depth slices, respectively.

[Feitchtenhofer 2022], can be masked during training. Masking 90% of the data reduces the memory overhead of training and thereby makes the model more scalable. An example of the pre-training is shown in Figure 2; the left column shows a set of inline, crossline, and depth sections from an input mini-cube. The middle column shows a random collection of 16^3 patches input to the model, and the right column shows the reconstruction. We see qualitatively that the model can reconstruct fine details, including faults and truncations, even from a minimal subset of the input data.

The encoder trained in this study is a generic transformer architecture with a depth of 32 layers, 16 attention heads, an embedding vector size of 1280, and a feedforward dimension of 5120, equivalent to ViT-H model in the literature. This model has a total of 660M trainable parameters. On the other hand, the decoder is a smaller transformer with eight layers, 16 attention heads, and a feedforward network size of 2048.

The context size of the model is the number of 16^3 patches (visual tokens) the model can attend to in a mini-cube. The larger the context size, the greater the geological context the model can see, which is important for local and global features. Once pre-training is complete, we fine-tune the model's context size to accommodate larger seismic mini-cubes. Pre-training is done with 512^3 mini-cubes, equivalent to a context size of 32,768. Once pre-training is complete, we fine-tune the model using 640x640x1024 seismic mini-cubes to achieve a context size of 102,400. This means that, based on the bin spacing of the seismic data, the model sees approximately 8-16 km in the lateral direction and 5-10 km in the depth direction. The context fine-tuning has been done on the same hardware.

Pre-training dataset

This study aims to scale the ViT-MAE concept to a global geological context. For this reason, we assembled a corpus of 63 seismic surveys, sampled from around the world, to use in pretraining. The spatial region, expanding the surveys in the pre-training dataset, is shown in Figure 3. Table 1 summarises the size and contribution to the training data from each region. We train on depth-migrated final stacks, which have been imaged with either reverse time migration (RTM) or Kirchoff depth migration (KPSDM). This dataset has 1.8 billion 16^3 patches (visual tokens) without overlap. For comparison, our dataset contains an equivalent of 293 million 224x224 2D inline and crossline subset images without augmentation and decimation. We also overlap mini-cube sampling by 50% to achieve 12 billion visual tokens that augment positional information.

Although the scaling laws of ViT models are not explored in seismic data, studies into vision transformers on natural images suggest that larger models achieve higher accuracy when fine-tuned on image classification tasks and that larger datasets are beneficial when training large models. However, even with limited data, the large models, although requiring more compute resources, perform better than smaller models [Zhai 2022]. For reference, as of the time of writing, the largest published ViT model is ViT-22B, a 22 billion parameter ViT model [Dehghani 2023], which was trained on a proprietary dataset of approximately 4 billion images with 256 visual tokens per image.

The computational cost of pre-training the 3D ViT-MAE model with our configuration is approximately 976 A100 core days, which is significant. For large context fine-tuning, 244 more

Region	File Size (GB)	Survey Area (sq km)	Number of Surveys
Africa	3,603	91,438	11
Asia	668	12,285	3
Australasia	1,194	36,622	3
Canada	1,910	17,900	3
Europe	1,284	14,713	2
Gulf of Mexico	4,894	106,227	26
South America	6,765	164,394	9
Onshore USA	124	1,130	5
Total	**20,444**	**444,710**	**63**

Table 1 The dataset size by region by file size in GB and project area in sq km.

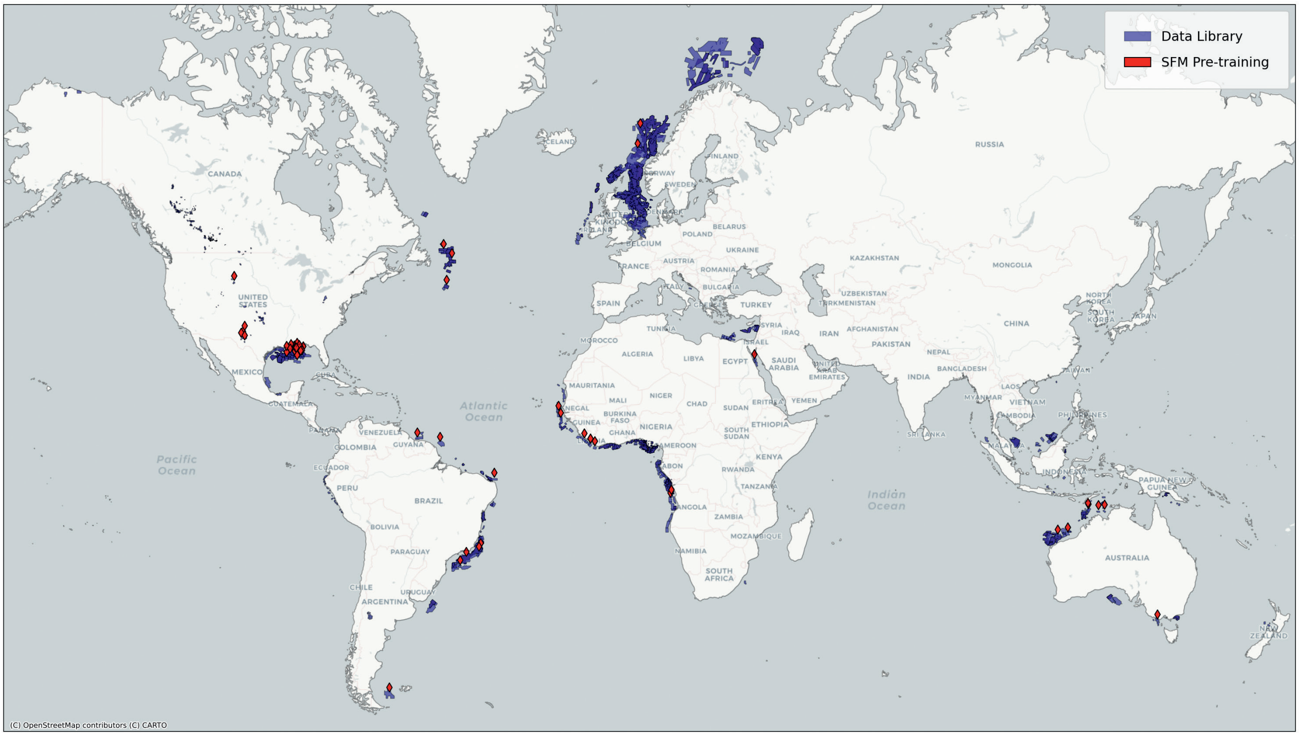

Figure 3 A view of 3D post-stack seismic data in our data library. (red) The 63 surveys we sampled from around the world are included in pre-training.

A100 core days are added, totaling 1220 A100 core days. We pre-trained the model using a cluster of A100 GPUs for this study. Keeping the GPUs sufficiently utilised is a critical requirement that makes training a large seismic foundation model on a global scale feasible. Another crucial requirement is that we can co-locate the data with the computer. Otherwise, repeated sampling of this data throughout training would also be inefficient.

This study was made possible by accessing an extensive library of multi-client seismic data hosted on the cloud. A key feature of the data library is that all data is accessible in place, which means that any data in the library can be utilised in training without duplication or any additional overhead of discovery or preprocessing. A key enabling technology is the MDIO open-source format for seismic data [Sansal 2023]. MDIO has the advantage of providing lossless data compression, which minimises network traffic, and more importantly, it is a chunked data format compatible with native cloud storage. Each stack is arranged as a collection of 128^3 chunks on a cloud bucket.

The network architecture of the model was chosen to align with the 3D domain of the post-stack data, removing the need to sample data in 3D and then make an arbitrary 2D slice, consequently reducing the I/O overhead. When the model training steps iterate, it samples batches of large amounts of data in desired chunks from surveys across the globe. For example, to fit three 512^3 mini-cubes on an 8 GPU node (24 total batch size per node), we are fetching 12GB of seismic data samples at each iteration. Since we are working with 3D data, chunked data formats like MDIO are significantly more capable than sequential formats like SEG-Y, which require indexing and orders of magnitude more requests to read a mini-cube.

We use Dask [Dask Development Team 2016] for multi-process read operations, MDIO as a file format, and the MDIO library to access the data with relevant metadata. This is integrated with PyTorch [Paszke 2019] datasets to provide high-performance I/O, which allows us to keep the GPU cluster fully utilised during training.

In summary, combining these technologies means that all the independent mini-cubes in our data corpus can be randomly sampled into batches and used in training without creating an I/O bottleneck.

Downstream task: Salt interpretation

To demonstrate the usefulness of the pre-trained model, we train a new decoder for the pre-trained foundation model for salt segmentation. Examples of downstream tasks relevant to geophysics are summarised by [Sheng 2023]. In this study, we fine-tune using an expanded version of the salt interpretation dataset previously used for training salt segmentation U-Net models in 2D and 3D in Roberts [2024] and Warren [2023]. This dataset consists of salt annotations from 20 reverse time-migrated depth stacks from the Gulf of Mexico. For this study, we have added four new interpreted RTM stacks from South America for training and another interpreted RTM stack from Africa for testing (out-of-domain). A ground truth salt label is a binary mask derived from interpretations carried out by expert geophysicists.

As in pretraining, the salt labels are stored in MDIO format. The survey geometry and other metadata are consistent between the seismic labels and underlying seismic data, which is essential for correct training. Both labels and data are accessible in place, removing the need for data duplication.

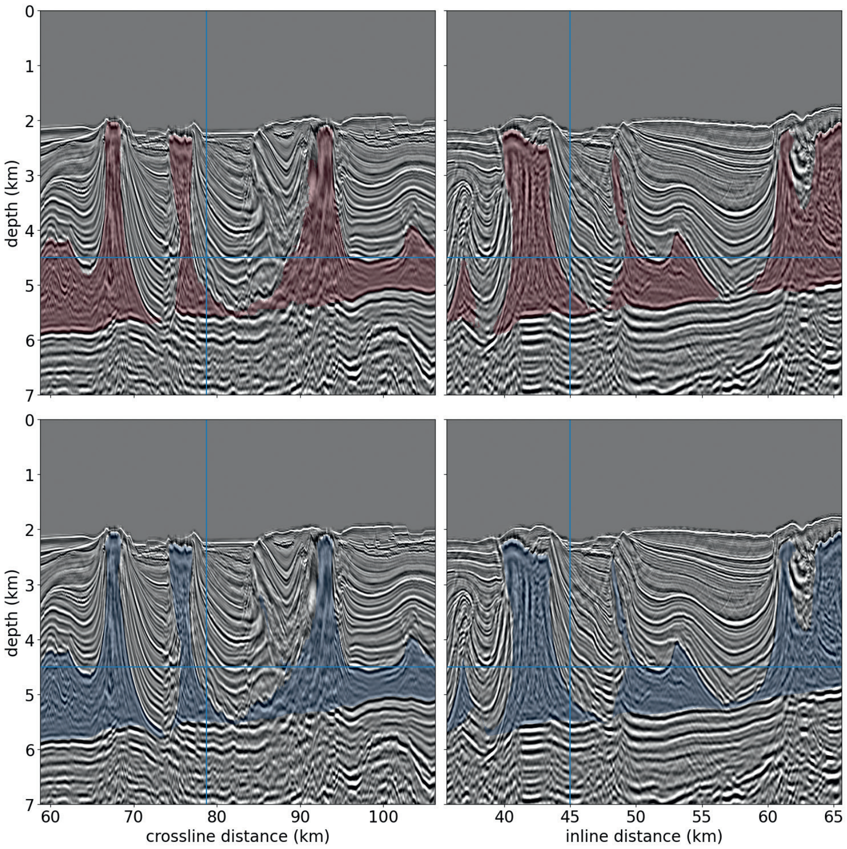

Figure 4 Offshore Africa: (top red) the raw and unprocessed salt label prediction masks for an inline and crossline section, respectively. (bottom blue) The ground truth labels. Guidelines indicate the location of the other orthogonal slices shown for this volume.

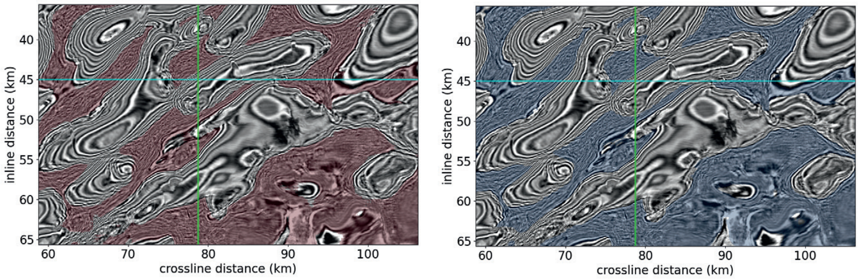

Figure 5 Offshore Africa: (left red) is the predicted depth slice salt mask. (right blue) The associated ground truth. (cyan and green lines) The location of the inline and crossline sections is shown in Figure 4.

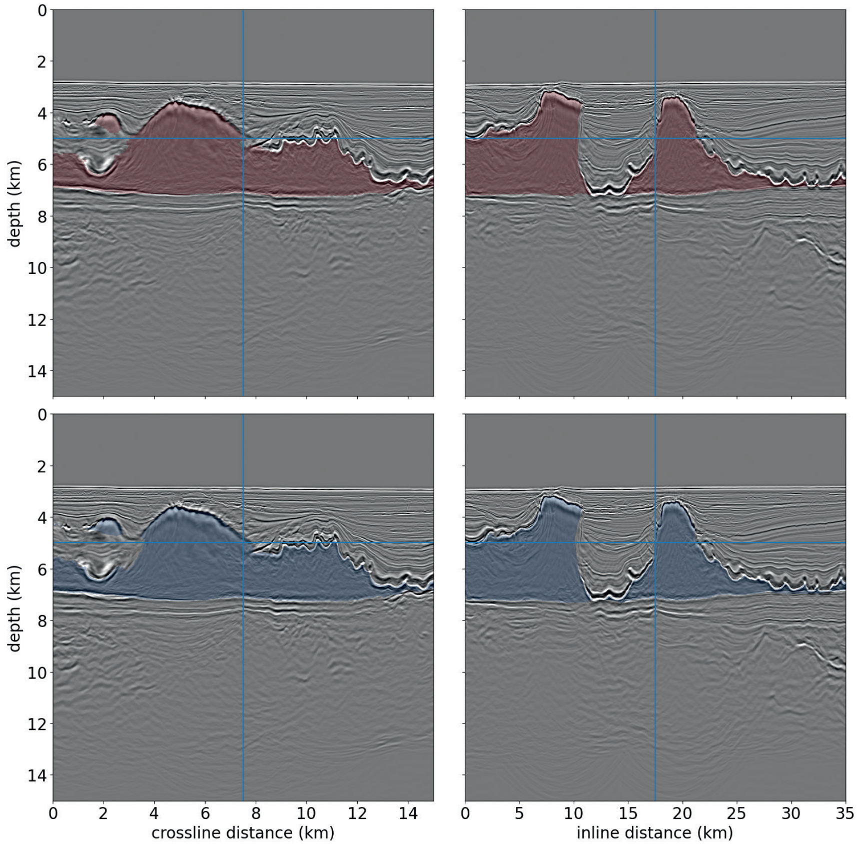

Figure 6 As in Figure 4, but for Inline and crossline sections from the held-out South American dataset.

The salt classification network architecture consists of a frozen pre-trained encoder (weights do not need to be updated) with a transformer decoder and a single layer as a classification head. The model's performance is evaluated in terms of intersection over union (IoU); this gives an immediate comparison with metrics used in previous studies [Roberts, 2024; Warren, 2023; Sheng, 2023].

Results

To evaluate the performance of the fine-tuned model on salt interpretation, an RTM stack offshore South America is held out in both pre-training and for salt classification. This provides a performance comparison analogous to [Roberts 2024 and Warren 2023], where data is held out in the region where the model is trained. To evaluate the potential for the SFM to aid the geologic region generalisation of the salt model, an interpreted stack offshore Africa is used in pre-training but held out in creating the salt model. Table 2 shows the intersection over union (IoU) metrics used to score the model's performance.

Metric	Africa - hold out	South America - hold out
IOU (mean)	0.90	0.96
IOU (foreground)	0.83	0.93
IOU (background)	0.97	0.99

Table 2 Performance metrics for Africa and South America hold out datasets.

With the held-out African dataset, we can test the efficacy of applying the model outside the basin in which it was trained. In both previous ML salt interpretation studies [Roberts 2024 and Warren 2023], the prediction is evaluated on held-out volumes but within the area where the model is trained. Figures 4 and 5 show example prediction versus ground truth salt masks for the African dataset. The IoU score of 0.83 is consistent with the state-of-the-art results of (0.84, 0.96) for the two GoM datasets evaluated by [Roberts 2024] using 3D U-Nets. The result indicates that we can realise excellent generalisation of the salt model outside of the basins where it has been trained in the case where the underlying dataset was included in pretraining. We also expect strong few-shot generalisation

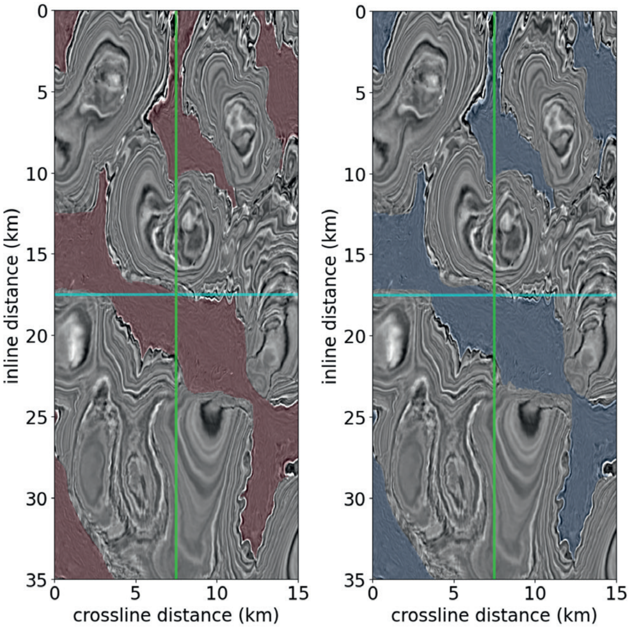

Figure 7 As in Figure 5, but for the depth slices for the held-out South American dataset.

across new areas, achievable with minimal labels and fine-tuning.

Figures 6 and 7 show analogous examples of salt classification in South American data. This survey was not included in the pre-training or the salt model training. The IoU of 0.93 is at the high end of the range obtained by [Roberts 2024]. This indicates that the ViT-based self-supervised model achieves state-of-the-art performance as seen in the GoM-only U-Net models but is trained across two basins, the GoM, and South America.

Conclusions

This study demonstrates the transformative potential of scaling the Vision Transformer architecture with the Masked Autoencoder training technique (ViT-MAE) to seismic data, achieving state-of-the-art performance in salt segmentation tasks. We highlight the advancements made possible by pretraining a 660-million-parameter model on a global dataset of 63 seismic surveys through efficient data handling and model scalability. The MDIO format enabled high-throughput access to large seismic datasets stored on the cloud, ensuring efficient large-scale data delivery to utilise the power of A100 GPUs during pretraining. This infrastructure is a key enabler of scaling, allowing efficient data management for training large-scale models.

Working in 3D allowed us to use a 90% masking ratio, further enhancing scalability by reducing memory overhead in pre-training and enabling the larger models for a given GPU footprint. This approach effectively reconstructs fine geological details from sparse inputs, showcasing its power in handling 3D seismic data.

Achieving an IoU of 0.83 on a held-out African data and 0.93 on the held-out South American data, the salt segmentation task model demonstrates exceptional generalisation beyond the basins where it was trained. This aligns with state-of-the-art CNN-based approaches when the ML models are trained and applied to the same basin.

We have demonstrated a highly scalable method of training a seismic foundation model. This work establishes a framework for leveraging large-scale data and cutting-edge architectures for training seismic foundation models, which is scalable beyond a 1-billion parameter model.

References

He, K., Chen, X., Xie, S., Li, Y., Dollár, P. and Girshick, R. [2021]. Masked Autoencoders Are Scalable Vision Learners. *ArXiv*. https://arxiv.org/abs/2111.06377.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. [2021]. Masked Autoencoders Are Scalable Vision Learners. *ArXiv*. https://arxiv.org/abs/2111.06377.

Zhai, X., Kolesnikov, A., Houlsby, N. and Beyer, L. [2021]. Scaling Vision Transformers. *ArXiv*. https://arxiv.org/abs/2106.04560.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A., Caron, M., Geirhos, R., Alabdulmohsin, I., Jenatton, R., Beyer, L., Tschannen, M., Arnab, A., Wang, X., Riquelme, C., Minderer, M., Puigcerver, J., Evci, U. and Houlsby, N. [2023]. Scaling Vision Transformers to 22 Billion Parameters. *ArXiv*. https://arxiv.org/abs/2302.05442.

Sheng, H., Wu, X., Si, X., Li, J., Zhang, S. and Duan, X. [2023]. Seismic Foundation Model (SFM): A new generation deep learning model in geophysics. *ArXiv*. https://arxiv.org/abs/2309.02791.

Ordonez, A., Wade, D., Ravaut, C. and Waldeland, A.U. [2024]. *Towards a Foundation Model for Seismic Interpretation*. 85th EAGE Annual Conference & Exhibition, 2024, 1-5. DOI: https://doi.org/10.3997/2214-4609.2024101119.

Roberts, M., Warren, C., Lasscock, B. and Valenciano, A. [2024]. *A Comparative Study of the Application of 2D and 3D CNNs for Salt Segmentation*. 85th EAGE Annual Conference & Exhibition, 1-5.

Warren, C., Kainkaryam, S., Lasscock, B., Sansal, A., Govindarajan, S. and Valenciano, A. [2023]. Toward generalized models for machine-learning-assisted salt interpretation in the Gulf of Mexico. *The Leading Edge*, **42**(6), 390-398.

Lasscock, B.G., Sansal, A. and Valenciano, A. [2024, September 17-20]. *Encoding the Subsurface in 3D with Seismic [Paper presentation*]. IMAGE 2024, Houston, TX, United States.

Sansal, A., Kainkaryam, S., Lasscock, B. and Valenciano, A. 2023. MDIO: Open-source format for multidimensional energy data. *The Leading Edge,* **42**(7), 465-470. https://doi.org/10.1190/tle42070465.1.

Dask Development Team [2016]. *Dask: Library for Dynamic Task Scheduling*. Available at: http://dask.pydata.org [Accessed: 9 December 2024].

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. [2019]. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, **32**, 8024-8035.