2025 The Middle East Oil, Gas and Geosciences Show

Paper Number (SPE-227123-MS)

Advances in Scaling and Architecture of 3D Foundation Models for Seismic Data

T.A. Sansal[1]*, B.G. Lasscock[1], A. Valenciano[1], J. Brittan[2], and M.T. Raafat[3]

[1]TGS, Houston, TX, USA
[2]TGS, Weybridge, UK
[3]TGS, Cairo, Egypt

Abstract

Traditional machine learning approaches for seismic data interpretation have historically depended on iterative training and inference processes applied to individual datasets, resulting in models lacking robustness and failing to generalize beyond their specific training domains. To address these limitations, we introduce a novel methodology that leverages self-supervised training and the scaling of 3D Vision Transformer (ViT) architectures to significantly enhance seismic interpretation capabilities significantly, enabling improved generalization across diverse geological datasets. This study focuses on the complexities associated with large-scale training, utilizing a comprehensive global dataset comprising 63 seismic volumes. We employ the Masked Autoencoder (MAE) architecture, integrated with the ViT-H model, which encompasses an impressive 660 million parameters, to achieve unprecedented scalability and performance in seismic data processing.

Our approach is underpinned by a cloud-native, digitalized seismic data infrastructure that effectively tackles data engineering challenges, eliminating the need for data duplication and streamlining access to large-scale datasets. This infrastructure, combined with the MDIO seismic data format, facilitates efficient data management and high-throughput access, ensuring that computational resources, such as A100 GPU clusters, are fully utilized during training. We fine-tuned a salt segmentation for a practical demonstration of the model's capabilities.

Introduction

Emerging research highlights the potential of the pre-trained ViT-MAE model to revolutionize seismic processing and interpretation [Lasscock 2024, Sheng 2023]. Similar to the transformative impact of large language models in natural language processing, this AI approach could significantly advance geophysical applications. Previous studies have focused on small, open-source datasets with synthetic data and outdated seismic imaging techniques. A comprehensive study by [Ordonez 2024] curated 60,000 2D crops from a 20-survey dataset for pretraining, demonstrating the effectiveness of developing a seismic foundation model (SFM) and fine-tuning it for downstream tasks like seismic salt and facies classification.

The scalability of ViT-MAE, particularly in 3D applications [Lasscock 2024], remains underexplored in geophysical research. In computer vision, studies like [Zhai 2022] show that larger models trained on extensive datasets (e.g., ImageNet-21k, JFT-300M) excel in image classification. This project investigates scaling ViT-MAE models on a global dataset of 63 seismic surveys to assess whether fine-tuning these large pre-trained models for downstream tasks can surpass current AI methods in generalization performance.

Effective data management is critical for training large models, requiring both the curation of vast datasets and the optimization of GPU clusters for efficient training. Seismic data poses unique challenges in this context. We describe how cloud object storage and the MDIO seismic data format [Sansal 2023] enable efficient pretraining of a 660M-parameter 3D ViT-H model. The model's utility is evaluated by fine-tuning it for salt interpretation using the SaltNet dataset, which includes interpretations from 23 seismic volumes. We compare its IoU scores against state-of-the-art 2D and 3D U-Net models [Warren 2023, Roberts 2024].

Methodology

Model Architecture

The architecture of the model, as illustrated in Figure 1, is built upon a Masked Autoencoder (MAE) framework integrated with a Vision Transformer (ViT) backbone, as initially proposed by He et al. [2021], but specifically tailored to handle 3D seismic volumes. The input seismic data is systematically partitioned into overlapping mini-cubes, which are then subjected to data augmentations, such as flips along the inline and crossline directions, to enhance model robustness. This design modifies the ViT-MAE framework, initially developed for 2D image processing, to accommodate the complexities of 3D seismic data by projecting 163 patches, referred to as visual

tokens, into a series of 1280-dimensional vector embeddings. During each training iteration, a subset of mini-cubes is randomly sampled from the global dataset, with 90% of the patches intentionally masked. The remaining 10% of unmasked patches serve as the input for reconstructing the original mini-cube. The primary training objective is to optimize the reconstruction accuracy of the masked patches in pixel space, quantified using the mean-squared error (MSE) metric.

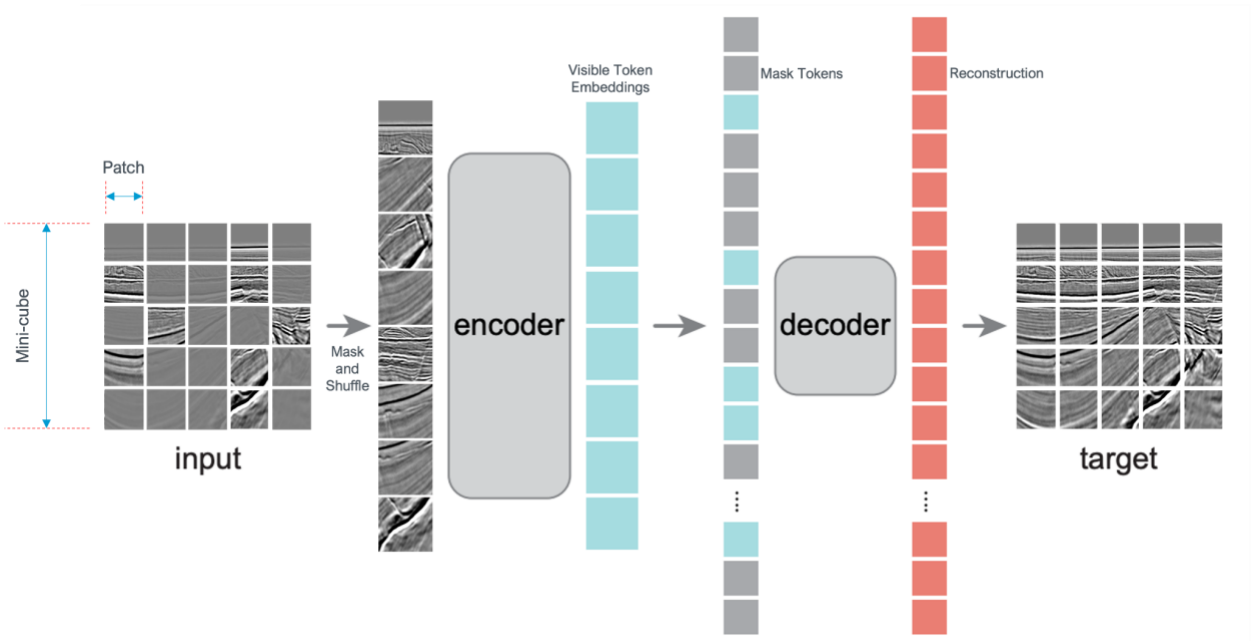

Figure 1. The picture shows a modified schematic view that explains the ViT-MAE pre-training concept [He 2021]. Large 3D data patches are loaded in batches. 90% of the data is discarded, and the remaining 10% is used to reconstruct the original data from the mask tokens.

One of the key strengths of this self-supervised training strategy is its memory efficiency, which stems from the fact that the entire dataset is only processed by the lightweight decoder during the computation of the loss function. In contrast, the significantly larger encoder only needs to process 10% of the patches, substantially reducing memory demands. This approach allows the model to scale effectively to large architectures without the need for intricate distributed training techniques. Furthermore, operating in the 3D domain, as opposed to 2D, permits a higher masking ratio—90% in this instance, as highlighted by Feitchtenhofer [2022]—which further lowers the memory footprint during training. This high masking ratio makes the training process more computationally efficient and enhances the model's scalability, enabling it to handle large seismic datasets effectively. An example of the pre-training process is showcased in Figure 2, where the left column presents a set of inline, crossline, and depth sections extracted from an input mini-cube. The middle column displays a randomly selected subset of 163 patches provided as input to the model, while the right column reveals the reconstructed mini-cube. Qualitatively, the reconstruction demonstrates the model's capability to recover intricate geological features, including faults and truncations, with remarkable fidelity, even when relying on a minimal subset of the input data.

The encoder employed in this study adheres to a generic transformer architecture, characterized by a depth of 32 layers, 16 attention heads, an embedding vector size of 1280, and a feedforward network dimension of 5120. This configuration aligns closely with the ViT-H model described in the literature and comprises 660 million trainable parameters. In contrast, the decoder is designed to be more compact, featuring a transformer with eight layers, 16 attention heads, and a feedforward network dimension of 2048, making it significantly less resource-intensive. The context size of the model, defined as the 163 patches (visual tokens) it can process within a single mini cube, plays a critical role in determining the extent of geological context the model can capture. This context is essential for accurately modeling both local and global geological features, such as faults, horizons, and other subsurface structures.

During the pre-training phase, the model is trained on mini cubes with dimensions of 512^3, corresponding to a context size of 32,768 patches. Following the completion of pre-training, the model undergoes fine-tuning to accommodate larger seismic mini cubes with dimensions of 640x640x1024, resulting in an expanded context size of 102,400 patches. This larger context size enables the model to analyze seismic data over a spatial extent of approximately 8-16 kilometers in the lateral directions (inline and crossline) and 5-10 kilometers in the depth direction, depending on the bin spacing of the seismic dataset. This capability is particularly valuable for capturing large-scale geological structures and ensuring the model can generalize across diverse seismic datasets. Notably, the fine-tuning process is conducted on the same hardware used for pre-training, ensuring consistency in computational resources and facilitating a seamless transition from pre-training to fine-tuning. This approach underscores the model's adaptability and potential for practical deployment in seismic data analysis tasks.

Pre-Training Dataset

This study seeks to expand the Vision Transformer with Masked Autoencoder (ViT-MAE) methodology to encompass a global geological framework to capture a broad spectrum of subsurface characteristics. To achieve this, we meticulously curated a pretraining dataset comprising 63 seismic surveys, carefully selected from diverse regions across the globe to ensure comprehensive geographical representation. The spatial extent of these surveys, which collectively form the pretraining corpus, is visually depicted in Figure 3. Table 1 provides a detailed breakdown of the dataset, summarizing the data volume contributed by each region and highlighting their respective roles in the training process.

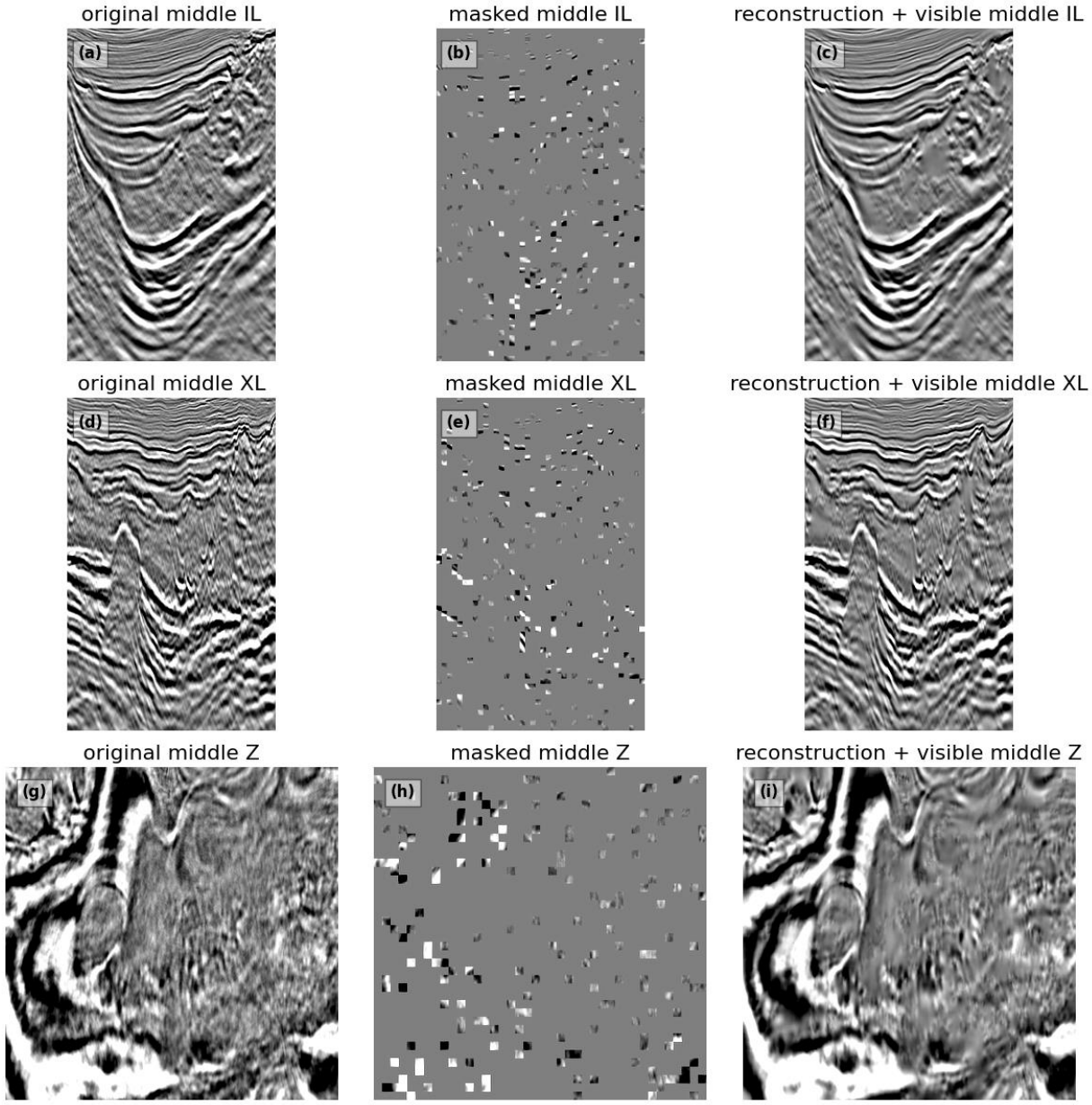

Figure 2. A specific example of a sampled 640x640x1024 mini-cube and its reconstruction. (a-c) A mid-point inline slice through the 3D patch showing the original data, the data used in reconstruction, and the reconstructed 3D patch. (d-f) and (g-i) show the equivalent crossline and depth slices, respectively.

For this research, we utilized depth-migrated final stacks, which were processed using one of two advanced imaging techniques: reverse time migration (RTM) or Kirchhoff pre-stack depth migration (KPSDM). These methods ensure high-fidelity subsurface imaging, which is critical for the accuracy of our model. The resulting dataset is substantial, containing 1.8 billion non-overlapping 163 patches, referred to as visual tokens, which serve as the fundamental units for our pretraining process. To provide a sense of scale, this dataset is equivalent to approximately 293 million 224x224 2D images, derived from inline and crossline subsets, without the application of data augmentation or decimation techniques.

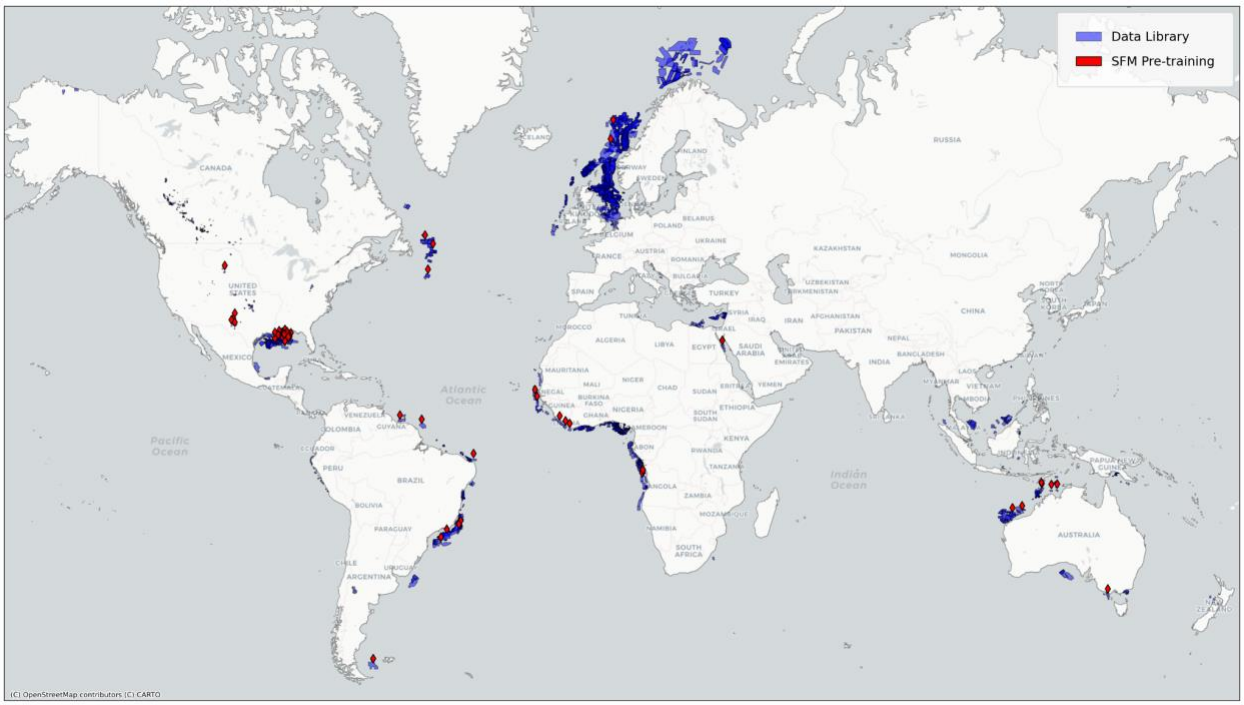

Figure 3. A view of 3D post-stack seismic data in our data library. (red) The 63 surveys we sampled from around the world are included in pre-training.

To further enhance the dataset's robustness and enrich the positional information available to the model, we implemented a 50% overlap in mini-cube sampling. This approach significantly increases the dataset's density, yielding 12 billion visual tokens. This augmentation strategy amplifies the dataset's volume and improves the model's ability to capture intricate spatial relationships within the seismic data, thereby strengthening the pretraining process for global geological applications.

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

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**

Research on vision transformers (ViTs) applied to natural images provides valuable insights into their scaling properties, indicating that larger models achieve superior accuracy when fine-tuned for image classification tasks. Additionally, these studies highlight the advantage of using larger datasets to train expansive models. Notably, even when data is limited, larger ViT models outperform their smaller counterparts despite demanding greater computational resources [Zhai 2022]. As a point of reference, the largest ViT model published to date is ViT-22B, which boasts 22 billion parameters and was trained on a proprietary dataset comprising roughly 4 billion

images, each represented by 256 visual tokens [Dehghani 2023]. Although scaling laws for ViT models have not been extensively studied in the context of seismic data, these findings from natural image applications offer a promising foundation for seismic research.

The computational demands of pre-training our 3D ViT-MAE model are substantial, requiring approximately 976 A100 GPU core days for the selected configuration. Fine-tuning the model for large-context applications further increases the computational load by 244 A100 core days, bringing the total to 1,220 A100 core days. This pre-training was conducted on a cluster of A100 GPUs, where maintaining high GPU utilization was paramount to ensuring the feasibility of training a large-scale seismic foundation model. Equally critical was the co-location of data with computational resources. Without this, repeated data sampling during training would introduce significant inefficiencies, hampering the process.

This study leveraged an extensive cloud-hosted library of multi-client seismic data, a resource that proved instrumental to its success. A standout feature of this library is its universal accessibility, allowing any dataset within it to be seamlessly utilized for training without duplication, additional discovery, or preprocessing. Central to this capability is the MDIO open-source format for seismic data [Sansal 2023]. MDIO offers lossless compression, which reduces network traffic, and its chunked data structure is optimized for native cloud storage compatibility. In this setup, each seismic stack is organized into 128^3 chunks stored in a cloud bucket, facilitating efficient data access.

The model's network architecture was designed to accommodate the 3D nature of post-stack seismic data. This approach eliminates the need to sample 3D data and then extract arbitrary 2D slices, which would otherwise increase I/O overhead. During training, the model iterates through steps that involve sampling large batches of data in chunks from seismic surveys worldwide. For instance, to process three 512^3 mini-cubes on an 8-GPU node (resulting in a batch size of 24 per node), approximately 12GB of seismic data is retrieved per iteration. Using a chunked data format like MDIO is a significant improvement over traditional sequential formats such as SEG-Y, which require extensive indexing and exponentially more requests to access a mini-cube, making them less efficient for 3D data handling.

To ensure high-performance data access, we employed Dask [Dask Development Team 2016] for multi-process read operations, combined with the MDIO file format and its associated library to retrieve data and metadata efficiently. These components were integrated with PyTorch datasets [Paszke 2019], creating a robust I/O pipeline that utilized the GPU cluster throughout training. This integration is critical for avoiding bottlenecks that could otherwise arise from the intensive data demands of large-scale model training.

In conclusion, the synergy of these advanced technologies enables the random sampling of independent mini-cubes from our global seismic data corpus into training batches without creating I/O bottlenecks. This study demonstrates a scalable and efficient approach to training large 3D ViT models for seismic applications by leveraging cloud-native storage, optimized data formats, and high-performance computing frameworks.

Downstream Task: Salt Interpretation

To showcase the effectiveness of the pre-trained foundation model, we developed a new decoder specifically tailored for salt segmentation, leveraging the pre-trained model's capabilities. The significance of downstream tasks in geophysical applications, such as salt segmentation, is well-documented in Sheng [2023], which provides a comprehensive overview of relevant tasks. For this study, we performed fine-tuning using an enhanced version of the salt interpretation dataset initially employed for training 2D and 3D U-Net models for salt segmentation, as described in Roberts [2024] and Warren [2023]. This dataset comprises salt annotations derived from 20 reverse time-migrated (RTM) depth stacks sourced from the Gulf of Mexico. We incorporated four additional interpreted RTM stacks from South America to enrich the dataset for training purposes. We also included one interpreted RTM stack from Africa as an out-of-domain test set. The ground truth for these salt annotations consists of binary masks meticulously crafted by expert geophysicists based on their interpretations.

Consistent with the pretraining phase, the salt labels are stored in the MDIO format, ensuring compatibility and streamlined access. The survey geometry and associated metadata are carefully aligned between the seismic data and the corresponding salt labels, a critical factor for ensuring accurate model training. By enabling in-place access to both the seismic data and labels, the need for redundant data duplication is eliminated, enhancing efficiency.

The architecture of the salt classification network is designed to capitalize on the pre-trained model's strengths. It features a frozen encoder, where the weights remain fixed and are not updated during training, paired with a transformer-based decoder. A single-layer classification head is appended to the decoder to produce the final output. To evaluate the model's performance, we adopted the intersection over union (IoU) metric, which facilitates direct comparisons with the evaluation metrics used in prior studies, including Roberts [2024], Warren [2023], and Sheng [2023]. This approach ensures that our results are contextualized within the existing body of research, highlighting the model's effectiveness in salt segmentation tasks.

Results

To assess the effectiveness of the fine-tuned model for salt interpretation, a Reverse Time Migration (RTM) stack from offshore South America was excluded from both the pre-training phase and the salt classification process. This approach enables a performance evaluation comparable to the methodologies outlined in [Roberts 2024] and [Warren 2023], where data from the same region used for training was withheld for testing. To explore the capability of the Self-Supervised Foundation Model (SFM) to enhance the generalization of the salt model across different geologic regions, an interpreted seismic stack from offshore Africa was incorporated during pre-training but omitted during the development of the salt classification model. The performance metrics, specifically the Intersection over Union (IoU) scores, are presented in Table 2 to quantify the model's accuracy.

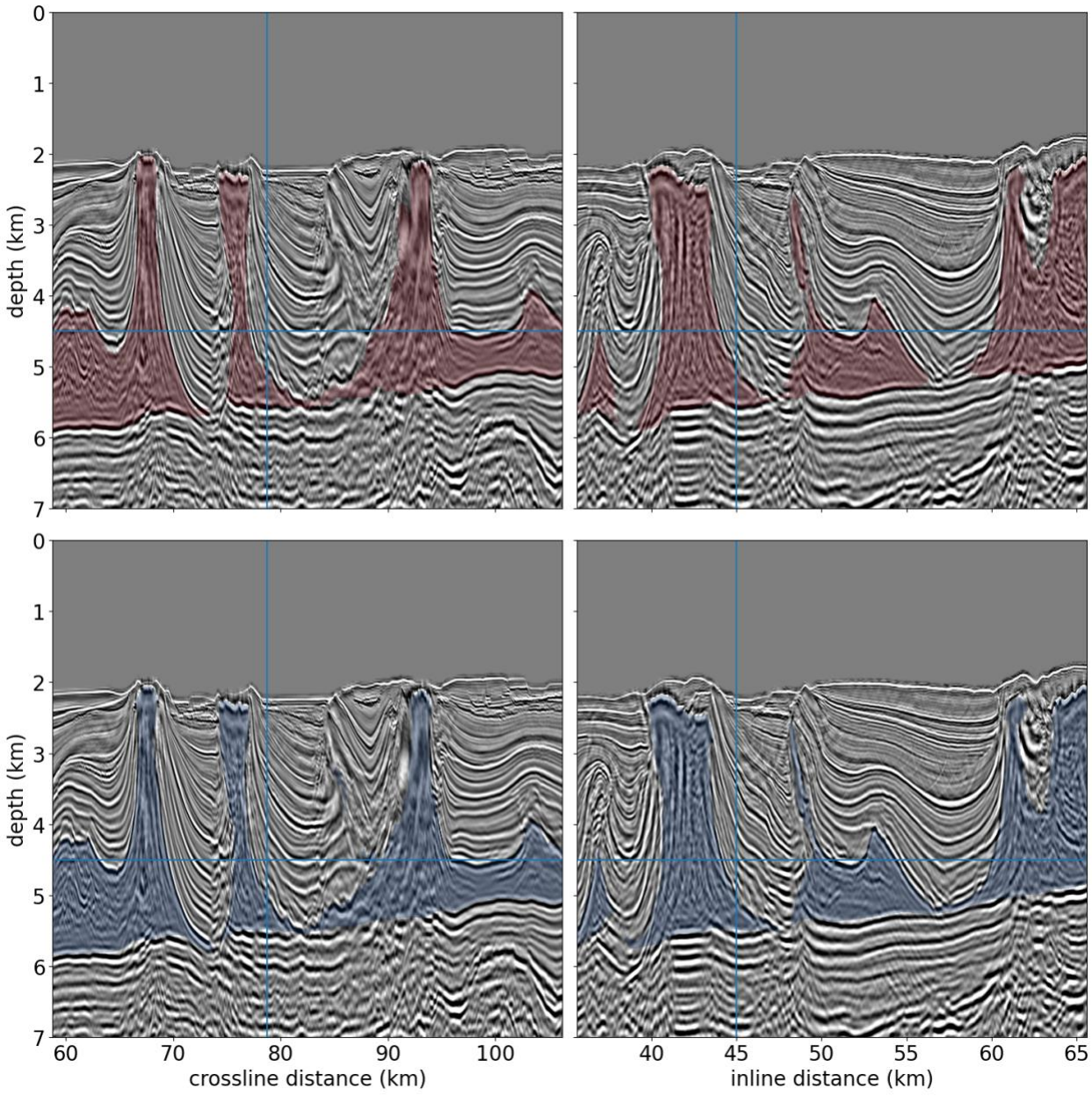

Figure 4. Offshore Africa: (top red) the raw and unprocessed salt label prediction masks for an inline and crossline section, respectively. (bottom blue) The ground truth labels. Guidelines indicate the location of the other orthogonal slices shown for this volume.

By utilizing the withheld African dataset, we can evaluate the model's ability to perform effectively in a geologic basin distinct from the one used for training. In prior machine learning studies on salt interpretation, such as those by [Roberts 2024] and [Warren 2023], model predictions were tested on held-out data volumes. However, these were still within the same geographic region as the training data. Visual comparisons of the model's predictions against ground truth salt masks for the African dataset are depicted in Figures 4 and 5. The IoU score of 0.83 aligns closely with the state-of-the-art results reported by [Roberts 2024], who documented IoU scores ranging from 0.84 to 0.96 for two Gulf of Mexico (GoM) datasets using 3D U-Net architectures. This outcome suggests that the salt model

exhibits robust generalization to regions outside the training basins, particularly when the underlying dataset was included in the pre-training phase. Furthermore, the model is expected to demonstrate strong few-shot generalization in new regions, requiring only minimal labeled data and fine-tuning to achieve high performance.

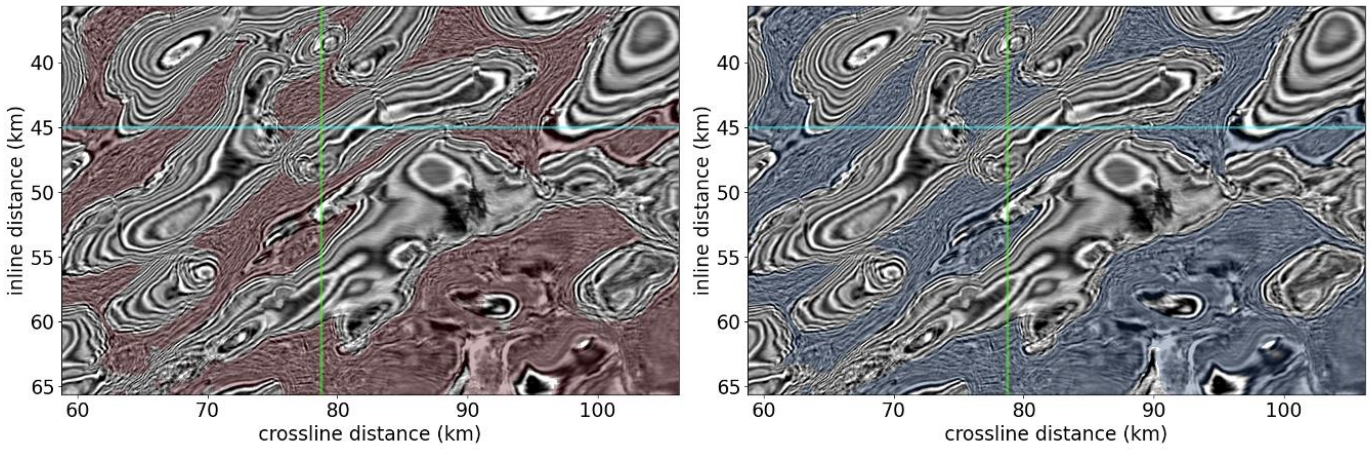

Figure 5. Offshore Africa: (left red) is the predicted depth slice salt mask. (right blue) The associated ground truth. (cyan and green lines) The location of the inline and crossline sections is shown in Figure 4.

For the South American dataset, which was entirely excluded from both pre-training and salt model training, Figures 6 and 7 provide illustrative examples of salt classification performance. The IoU score of 0.93 for this dataset is notably high, positioning it at the upper end of the performance spectrum reported by [Roberts 2024] for GoM datasets. This result underscores the capability of the Vision Transformer (ViT)-based self-supervised model to deliver state-of-the-art performance, matching or surpassing the results of GoM-focused 3D U-Net models. Importantly, this performance is achieved despite the model being trained across two distinct geologic basins—namely, the Gulf of Mexico and South America—highlighting its ability to generalize across diverse seismic environments.

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

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

Conclusions

This study showcases the transformative capabilities of scaling the Vision Transformer (ViT) architecture with the Masked Autoencoder (MAE) training approach for seismic data interpretation, achieving state-of-the-art performance in salt segmentation tasks and setting a new benchmark for geophysical applications. By pretraining a 660-million-parameter model on a diverse global dataset of 63 seismic surveys, we have demonstrated the feasibility of constructing highly scalable seismic foundation models that can effectively generalize across varied geological contexts. Integrating the MDIO seismic data format with a cloud-native infrastructure has been pivotal, enabling high-throughput access to large-scale seismic datasets and ensuring optimal utilization of A100 GPU clusters during pretraining. This robust data management framework eliminates inefficiencies such as data duplication and minimizes network traffic, thereby facilitating the training of large models in a computationally efficient manner.

Adopting a 3D approach, coupled with a 90% masking ratio, significantly enhances the scalability of the ViT-MAE framework by reducing memory overhead during pretraining. This allows the model to reconstruct intricate geological features, such as faults, truncations, and other subsurface structures, from sparse inputs, demonstrating its ability to capture local and global geological contexts with high fidelity. The salt segmentation task, fine-tuned on interpretation labels from the Gulf of Mexico and South America, achieved exceptional intersection over union (IoU) scores of 0.83 on a held-out African dataset and 0.93 on a held-out South American dataset. These results align with state-of-the-art convolutional neural network (CNN)-based approaches and highlight our model's superior generalization capabilities, particularly when applied to basins outside the training domain. The ability to achieve strong few-shot

generalization with minimal additional labels further underscores the practical utility of this approach in real-world geophysical workflows.

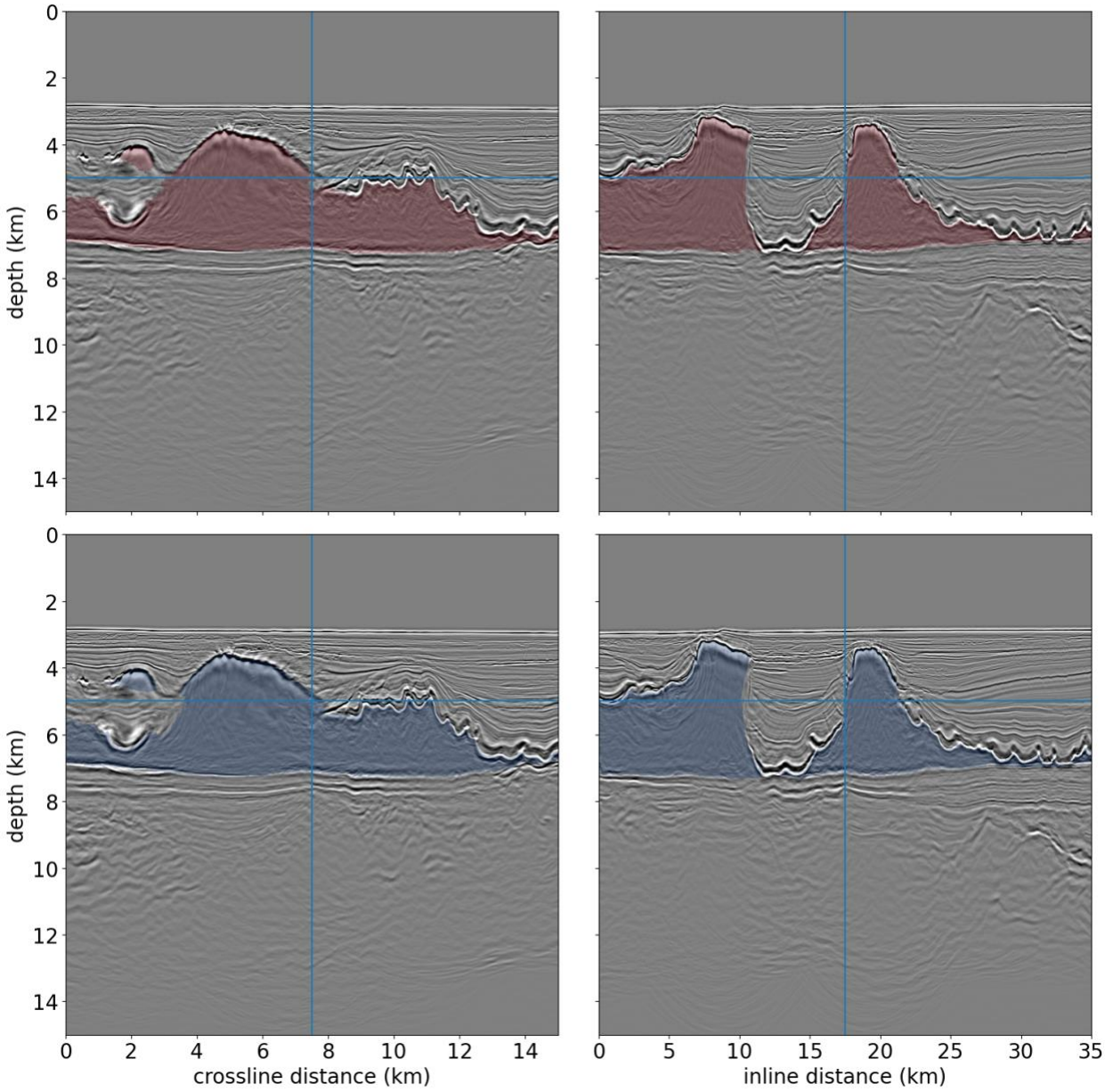

Figure 6. As in Figure 4, but for Inline and crossline sections from the held-out South American dataset.

This work establishes a comprehensive framework for leveraging large-scale seismic datasets and advanced machine learning architectures to develop scalable seismic foundation models beyond a one-billion-parameter threshold. By addressing the challenges of data management, computational efficiency, and model scalability, our methodology paves the way for future advancements in seismic interpretation, offering a pathway to more accurate, efficient, and generalizable subsurface analyses. The demonstrated success in salt segmentation suggests that similar approaches could be extended to other geophysical tasks, such as facies classification, fault detection, or reservoir characterization, potentially revolutionizing the field of subsurface imaging. As the geophysical community continues to embrace large-scale data and cutting-edge computational techniques, this study provides a blueprint for harnessing the power of foundation models to tackle complex subsurface challenges, ultimately contributing to more informed exploration and resource management decisions across global basins.

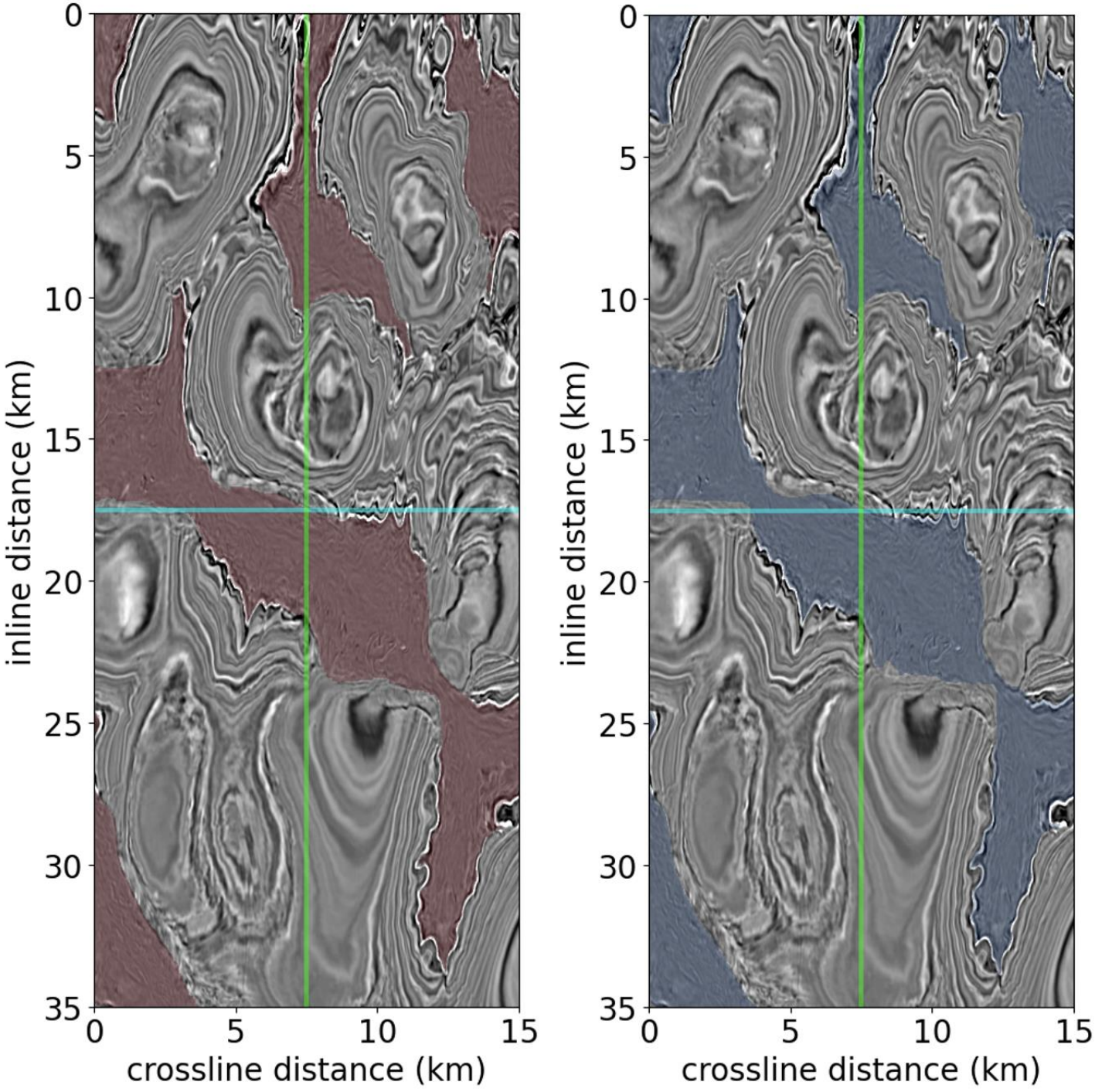

Figure 7. As in Figure 5, but for the depth slices for the held-out South American dataset.

References

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

Zhai, X., Kolesnikov, A., Houlsby, N., & 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., . . . 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., & 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., & 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., & 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., & 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.