Towards a cross-modal masked autoencoder for subsurface data

Ben Lasscock, Meher Gajula, Keyla Gonzalez, Brian Michell, and Alejandro Valenciano, TGS*

Summary

Fusing spatially dense, band-limited seismic data with sparse, high-resolution well logs remains a fundamental challenge in quantitative interpretation. We present a Cross-Modal Masked Autoencoder (CM-MAE), a self-supervised foundation model that learns a shared latent representation of 3D seismic and 1D well-log data. CM-MAE uses a single Vision Transformer encoder-decoder architecture to process a joint sequence of seismic and well tokens, enabling cross-modal reconstruction within a unified framework. The model is pretrained on a heterogeneous mixture of paired seismic-well, seismic-only, and well-only samples, allowing it to learn relationships across modalities**.** Ablation experiments show that the learned transfer is strongly asymmetric: seismic context substantially improves well-log reconstruction, reducing well reconstruction error by 52%, whereas well-log context provides negligible benefit for seismic reconstruction. Further analysis of well reconstruction reveals that it relies on seismic context and aligns with reflector-scale structures. We also demonstrate the practical value of the learned representation by supervised fine-tuning the pretrained model for seismic-only well reconstruction, where it predicts spatially coherent pseudo-logs directly from seismic data. Collectively, these ablation and downstream prediction results show that early-fusion self-supervised pretraining provides an effective framework for seismic-well integration and seismic-guided prediction of subsurface properties.

Introduction

Subsurface characterization draws on multiple geophysical and geological measurements. In this work, we focus on integrating two core modalities: 3D seismic data and well logs. This data is complementary but highly asymmetric in both coverage and information content. Seismic data provide dense lateral coverage and capture structural and stratigraphic context over large spatial extents, but they are band-limited and primarily sensitive to contrasts in elastic properties rather than to absolute rock and fluid properties. Well logs, in contrast, provide high-resolution measurements of petrophysical and elastic properties, but only along sparse 1D boreholes. As a result, predicting well properties away from borehole control remains a strongly non-unique inverse problem.

Recent work has examined foundation models for these modalities separately. The effectiveness of the vision transformer masked autoencoder (ViT-MAE) approach (He et al., 2022) for 3D seismic data was first demonstrated on seismic data (Lasscock et al., 2024; Sheng, 2025). Subsequent work explored scaling ViT-based seismic foundation models to larger datasets and architectures, demonstrating improved downstream performance and generalization across basins (Sansal et al., 2025a, 2025b). Meanwhile, the same method can be applied to well data, (Lasscock et al., 2025) demonstrated that a 60M-parameter foundation model pretrained on over one million well logs perform well on interpretation tasks, including log imputation and formation top picking.

While recent models can learn useful representations for seismic data and well logs separately, an important question is how these domains should be combined. One common architectural strategy is to analyze each data type using its own specialized model, with interaction between modalities introduced only at later stages (often referred to as *late fusion*). This modular design is widely adopted in modern multimodal architectures, including CLIP (Radford et al., 2021) and VL-JEPA (Chen et al., 2025). Late fusion can be computationally efficient because pre-trained, domain-specific models can be reused, while preserving the resolution and physical characteristics of each dataset prior to integration. As a result, cross-modal interaction occurs only after specialized models have independently learned representations of each data type.

An alternative approach is to combine different data types by learning a shared representation and process them jointly from the earliest layers of the model (commonly termed *early fusion*). This type of "natively multimodal" design underpins recent large-scale systems such as Google's Gemini (2023) and Meta's Chameleon (2024). In the case of Gemini, the model is trained from the beginning on interleaved multimodal data, enabling a single architecture to learn relationships across text, images, audio, and video in an integrated manner. The effectiveness of such approaches is typically assessed in two ways: first, by measuring performance on tasks within individual data domains, and second, by evaluating the ability to solve problems that require the combined interpretation of multiple data sources. Motivated by this integrated learning paradigm, the approach presented here adopts a similar early-fusion strategy for seismic and well-log data, while enabling the integration of additional subsurface datasets in future work.

The model presented here builds on our existing network designs and training infrastructure, introducing a new Cross-Modal Masked Autoencoder (CM-MAE) network. Based on the original ViT-MAE algorithm (He et al., 2022), the ViT architecture (Dosovitskiy et al., 2020), and multimodal extensions (Bachmann et al., 2022), CM-MAE learns a shared latent space between seismic and well-log data through a joint self-supervised reconstruction task. We demonstrate that this model learns an effective, asymmetric mapping in which dense seismic context enhances the prediction of sparse well-log properties.

CM-MAE architecture

The CM-MAE architecture (Figure 1) comprises a single shared ViT encoder and a single shared decoder with modality-specific linear prediction heads. The model is trained to reconstruct heavily masked input data, encouraging it to learn compact latent representations. Although the architecture is extensible to multiple modalities, including post-stack seismic, wells, offset vector tiled (OVT) gathers, angle stacks, and velocity cubes, this study focuses on the fusion of seismic and well data.

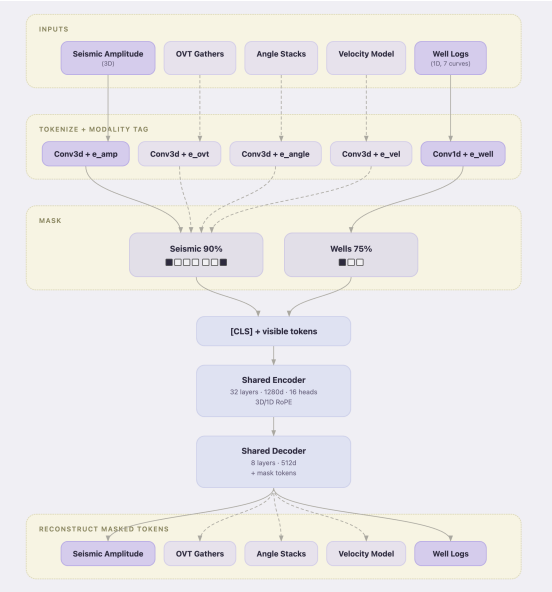

Figure 1: An extensible early fusion network design, taking in multiple modalities and combining them with a single encoder. Seismic amplitude and well modalities are explored in this study.

Tokenization and shared embedding space

To enable joint processing, both seismic and well data are converted into token sequences. The 3D seismic input is partitioned into non-overlapping 16^3voxel patches, producing 8,192 tokens from an $128 \times 128 \times 2048$ input window. Well data is tokenized on a per-curve basis into patches of 64 depth samples, producing up to 5,250 tokens from seven log curves (GR, NPHI, RHOB, DT, RESD, SP, and DTS) over 48,000 depth samples.

Patches from both modalities are then linearly projected into a shared latent space of dimension $d = 1280$. To preserve positional and modality-specific information, each token is augmented with axial Rotary Position Embeddings (RoPE) and a learned modality embedding, e_{seis}or e_{well}.

Masking and encoder design

A high masking ratio is applied during pretraining, with 90% masking for seismic tokens and 75% masking for well tokens. The shared ViT encoder, with 32 layers, 16 heads, and 663M parameters, processes only the visible subset of tokens, approximately 2,131 of the 13,442 tokens. This asymmetric design makes pretraining computationally efficient. Seismic and well tokens are projected into a shared embedding space, concatenated together with a CLS token, and processed jointly by the encoder, following multimodal masked autoencoder architectures (Bachmann et al., 2022). To encourage cross-modal learning, we additionally employ modality dropout, in which an entire modality is occasionally masked during training, forcing it to be reconstructed from the other modality. This encourages the model to learn relationships across modalities rather than relying only on intra-modal structure.

Decoder, loss, and inference

The shared lightweight transformer decoder, with 8 layers, 16 heads, and a latent dimension $d = 512$, takes the encoded visible tokens together with learned mask tokens to reconstruct the full data sequence. This encoder-decoder design is comparable in scale to the seismic model of (Sansal et al. 2025a). The total loss, L, is defined as the sum of the reconstruction losses for each modality; in this study, it includes only seismic and well-log reconstruction terms. Training minimizes the normalized per-patch mean squared error between reconstructed and original patches. At inference time, the model can operate on any available subset of input modalities. The model is pretrained for 267,000 steps using x8 H200 GPUs.

We then fine-tune (He et al., 2022) the pretrained CM-MAE foundation model to predict well logs from seismic data. In this example, we freeze the bottom 28 encoder layers and train only the top 4 layers, all Layer Norms, and the decoder. All well tokens are 100% masked, forcing the model to predict wells entirely from seismic. The seismic mask ratio is annealed from 90% to 0% over 15,000 steps, followed by 5,000 consolidation steps at full context. The resulting model produces a single best-estimate pseudo-log from unmasked seismic input.

Dataset and validation split

The dataset comprises three sample types drawn from marine Gulf of America and onshore Permian Basin data. First, paired seismic-well samples are extracted from 33

seismic surveys, including 30 overlapping marine surveys and 3 onshore Permian surveys, yielding 80,041 spatially matched well-to-seismic patch pairs across 34,361 unique wells. Second, seismic-only patches from the same surveys are added at 80% of the paired count, allowing the model to learn seismic reconstruction independently of well control. Third, 80,000 well-only samples are drawn from an unpaired pool of 144,199 offshore and Permian wells with no collocated seismic, providing additional supervision for the well modality. Together, these sources produce approximately 480,000 training samples spanning more than 114,000 unique wells.

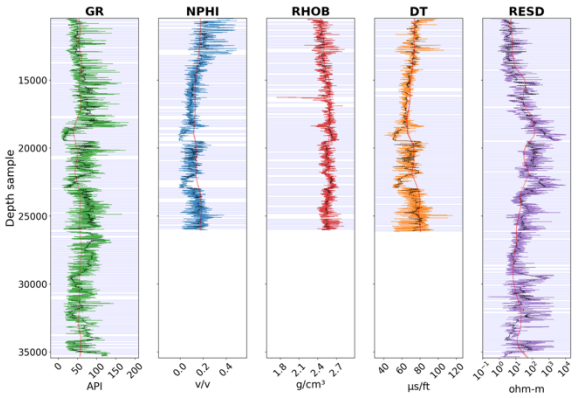

Figure 2 Gamma ray, neutron porosity, density, sonic and deep resistivity reconstruction using (black) seismic and well data (red) using only seismic data. (Blue background) Represents data masked in reconstruction

To prevent information leakage, the train-validation split is performed at the well level rather than the sample level. From 34,361 unique UWIs, 10% are randomly sampled and assigned to the validation set. All samples associated with held-out wells are placed entirely in validation, ensuring that no well-log curves appear in both training and validation. Seismic-only patches have no associated well identifier and are therefore assigned only to training. All evaluation and visualization scripts reproduce this same split.

Qualitative well-log reconstruction from the pretrained model

Figure 2 illustrates the latent capabilities of the pretrained CM-MAE model. Here, the model takes either "only seismic" samples or combinations of seismic and well data. The model is then trained to reconstruct both the well and seismic data. The modality dropout trains the model to make inference with missing information. The figures show (red) the reconstruction made from a random 10% sample of seismic data alone, and (black) the reconstruction made from both seismic and 25% of the available well data. We see that the model has learned to infer a smooth background trend from the seismic alone.

Evidence of cross-modal fusion

Evidence that the model can reconstruct well logs from seismic alone does not establish that the model is genuinely fusing modalities. To test that directly, we perform an ablation study on a 500 held-out validation wells by selectively withholding seismic and well inputs and measuring reconstruction loss. This design allows us to assess whether information learned from one modality can improve reconstruction in the other modality without compromising the accuracy of the original modality. Table 1 presents an ablation study comparing reconstruction loss when one modality is available versus both. We also compare to a baseline reconstruction ("Neither") in which no seismic or well tokens are provided to the encoder, and the model receives only the learned CLS token, forcing predictions to rely solely on the learned prior. We additionally compare against a non-learned, depth-aligned, averaged-curve baseline (Survey Mean). In each case, the MSE loss is computed on standardized curve values (Lasscock et al., 2025) and on standardized seismic data. The results show that the cross-modal transfer is strongly asymmetric

When seismic context is available, the well reconstruction improves over the well-only modality, demonstrating that the model has learned to use spatial seismic context to better reconstruct sparse well-log responses. This gain is not explained by regional priors, and the model outperforms a depth-averaged survey-mean baseline on every curve.

*Table 1: Cross-modal ablation over a collection of 166 land and 334 marine validation wells. A lower loss indicates better reconstruction, loss is calculated on standardized seismic and curves (*Lasscock et al., 2025*). As a baseline "Survey Mean" represents non-learned depth averaged log values across the training dataset, and "Neither" relies only on a learned class token.*

Scenario	Seismic	GR	NPHI	RHOB	DT	RESD	SP	DTS
Survey Mean	1.0773	0.0125	0.0213	0.0078	0.0194	0.0133	0.0020	0.0349
Neither	0.8656	0.0292	0.0746	0.0193	0.0653	0.0211	0.0049	0.0315
Both Visible	**0.1710**	**0.0038**	**0.0055**	**0.0024**	**0.0041**	**0.0019**	**0.0011**	**0.0030**
Seismic Only	0.1713	0.0092	0.0129	0.0066	0.0094	0.0062	0.0012	0.0206
Wells Only	0.8753	0.0097	0.0084	0.0057	0.0094	0.0033	0.0023	0.0058

Attention analysis supports this interpretation. Figure 3 shows that well reconstruction attends to the full 3D seismic volume rather than only to the trace collocated with the well, and that the strongest attention is concentrated at reflector-consistent depths. This behavior suggests that the model has learned a spatially distributed seismic-to-well relationship rather than a purely local regression rule.

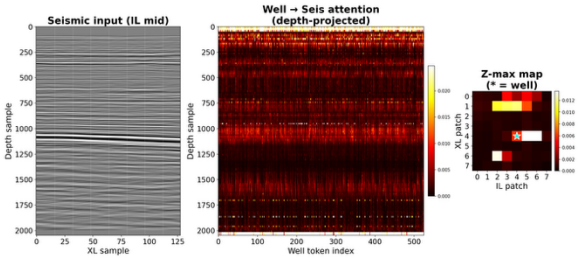

Figure 3: Cross-modal attention from a well log (right) to a seismic volume (left). The seismic encoder focuses its attention (bright horizontal bands) on specific depths in the well log that correspond to major seismic reflectors, indicating a learned physical alignment.

The seismic reconstruction does not meaningfully benefit from the addition of well data. One plausible explanation for the asymmetry is that seismic already provides a strong prior for its own reconstruction, while wells contribute only sparse local information. A second contributing factor may be the per-patch normalization (He et al., 2022) used in the seismic reconstruction loss, which emphasizes relative structure within each patch rather than absolute amplitudes. Under this loss, the model can reconstruct seismic effectively from seismic context alone, reducing the incremental value of well information.

Fine-tuned seismic-to-well prediction

Pseudo-log (Banchs et al., 2002) prediction from seismic is not new; prior approaches based on shallow multilayer perceptron and hand-crafted seismic attributes have long been used for this purpose. However, such methods typically depend on explicit feature engineering, operate on individual traces or small local windows, and require retraining for each survey. In contrast, CM-MAE learns cross-modal structure through large-scale self-supervised pretraining and is designed to capture broader spatial context. Figure 4 shows pseudo-log predictions from the Permian using a fine-tuned model. These predictions are smooth and consistent with interpolated 3D well-log attribute volumes (Gonzalez, 2024) and with direct well overlays where available, though with less vertical resolution than the interpolated reference volume. The interpolated model uses wells individually corrected for depth-elevation. The predicted section (center, measured depth) and the well-log model (right, true vertical depth) use different depth references, so exact correspondence is not expected. Each representation is indexed in its own depth domain, and no explicit well tie is enforced, so alignment should be interpreted as approximate and intended for visual comparison.

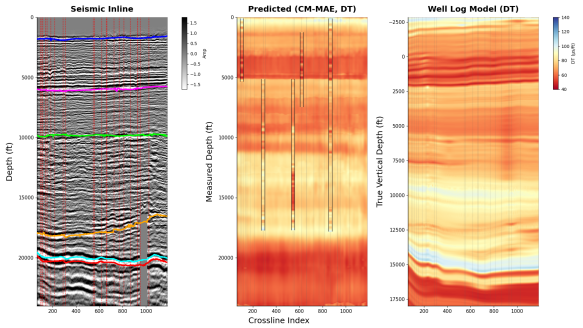

Figure 4: A pseudo-log section (center) along an inline (left). The model predicts spatially coherent well logs from seismic data, capturing lateral geological variations that align with seismic reflectors. As a comparison, interpolated well logs are shown (right).

Conclusions

We introduce a cross-modal masked autoencoder that combines 3D seismic data with 1D well logs into a shared representation. A five-way ablation study across 500 wells (both land and marine) shows that the model effectively transfers information between modalities: full fusion cuts well reconstruction error by 79% compared to a non-learned depth-aligned average baseline, and adding seismic data to partial well input halves the remaining error (a 52% reduction), improving both individual curves and overall results. The information transfer is notably asymmetric: seismic data significantly enhances well prediction, but wells do not provide measurable benefits for seismic reconstruction. This asymmetry highlights a fundamental dimensionality mismatch between the sparse 1D trace and the dense 3D volume, worsened by using patch-normalized loss for seismic reconstruction, which stabilizes training but decouples seismic targets from absolute amplitudes. Fine-tuning the pretrained encoder enables pseudo-log predictions from seismic data, producing spatially consistent log responses across seismic sections.

Several limitations remain. First, the current experiments focus on seismic-to-well transfer, and its ability to generalize to other modalities (e.g., angle stacks, OVT gathers, velocity models) is still untested. Second, evaluation is limited to the Permian and Gulf of America, and further validation is needed across different basins and acquisition geometries. Third, the asymmetric transfer of information suggests the training objective could be improved to better utilize cross-modal data. Future work will expand the architecture to include more modalities, test across different basins, and develop alternative objectives to enhance the balance of information exchange.