OpenSeisML: Open Large-Scale Real Seismic and well-log Dataset for Generative AI

Ipsita BharHuseyin Tuna ErdincThales SouzaCharles JonesFelix J. Herrmann

SUMMARY

The advent of machine learning (ML) and computer vision has significantly accelerated seismic inversion workflows by reducing the computational cost of traditionally expensive iterative methods. However, the development and evaluation of ML methods remains limited by the scarcity of realistic velocity models, as most high-quality data are privately owned by oil and gas companies. To address this gap, we present *OpenSeisML*, a collection of real seismic datasets designed to support generative AI (Gen-AI) workflows for seismic inversion. The datasets are curated from publicly available surveys in the UK National Data Repository (NDR). When seismic volumes are in the time domain and wells are in depth, a time-to-depth conversion is required. We use checkshot data to establish the time–depth relationship and construct a velocity model through interpolation for accurate conversion of post-stack seismic data. Here, we present an automated data curation pipeline that enables seismic data preparation while ensuring reproducibility. The objective is to train a generative model that captures the statistical distribution of subsurface properties, enabling the synthesis of multiple statistically consistent realizations for uncertainty quantification which can act as a prior for seismic inversion.

INTRODUCTION AND RELATED WORK

Recent advances in machine learning show that progress depends not only on model improvements, but also on standardized benchmarks, shared datasets, and reproducible workflows that enable fair comparison of different ML models under consistent data, metrics, and protocols, ensuring reproducible and trustworthy results (Donoho, 2024). The lack of publicly accessible high-quality seismic datasets, especially realistic velocity models, limits the development and validation of Gen-AI workflows, as also emphasized by (Jin et al., 2024) and (Alaudah et al., 2019). To mitigate this bottleneck, we present an automated seismic data curation pipeline that provides large scale real datasets consisting of imaged seismic volumes and well-log data. In addition to velocity related information, the well dataset also includes petrophysical measurements such as gamma-ray, neutron, density, sonic, and resistivity logs, which provide complementary constraints on lithology, porosity, and fluid properties.

Synthetic datasets have played a major role in developing new algorithms to train ML models for a myriad of subsurface applications such as underground energy storage (Gahlot et al., 2025); (Mekonnin et al., 2025), reservoir characterization (Chen et al., 2025), etc. These models rely on large-scale good quality training datasets to achieve reliable inference. Some of these synthetic datasets are discussed below.

Compass model and SEAM open data: The Compass model was constructed to evaluate acquisition strategies and velocity inversion methods (Yin et al., 2024); (Yin et al., 2025); (Orozco et al., 2025), by mimicking real geological complexity by incorporating features such as faults, folds, channels, and gas clouds derived from real seismic and well-log data from the North Sea (Jones et al., 2012). The model was built through a manual workflow by integrating well logs, horizons, and seismic attributes. Although realistic, the authors constructed only a single subsurface realization, thus limiting variability, so a generative model trained on it may overfit to that specific geology and struggle to capture the broader distribution of subsurface structures across different regions (Jin et al., 2024). SEAM (Fehler and Cheng, 2007) and (Pangman, 2014), is another realistic subsurface benchmark, developed for research advancement in seismic imaging and inversion. They incorporated realistic geological features such as salt bodies, stratigraphy, and structural complexity based on field data, however, the simulations still remain synthetic and controlled. This dataset also represents a limited number of deterministic realizations, whereas ML methods require diverse datasets with multiple realizations to adequately capture subsurface variability.

OpenFWI Benchmarks: To support the development and comparison of data-driven inversion methods, the OpenFWI benchmark provides a large-scale, open-access collection of synthetic seismic datasets (Deng et al., 2022). These were created by generating large-scale synthetic datasets using numerical wave equation solvers with diverse velocity models. The authors designed multiple geological scenarios with varying structures and acquisition settings to produce paired seismic data and ground-truth models. This controlled setup enables supervised training and systematic evaluation of data-driven inversion methods; however, these datasets do not fully capture the complexity and variability of real field data (Jin et al., 2024). As a result, models trained on these datasets tend to learn mappings specific to the synthetic setup and may struggle to generalize to more complex or realistic subsurface conditions (Consolvo, 2023).

OUR CONTRIBUTION

From the above discussion, it is evident that while some models are geologically realistic, they are limited to single realizations and fail to capture stochastic variability, whereas others, despite their scale, lack sufficient geological complexity to represent real subsurface conditions. Hence, we wanted to address this challenge by curating a training set based on real-field seismic data from the UK National Data Repository (NDR) (North Sea Transition Authority, 2026), a government-supported platform structured access to subsurface data across the UK energy sector. The curated dataset is intended to support the training of generative models that produce statistically consistent and geologically plausible subsurface realizations

Properties	Units
Gamma	API
Neutron	dimensionless
Density	g/cc
Sonic	$\mu s/ft$
Resistivity	ohm-m

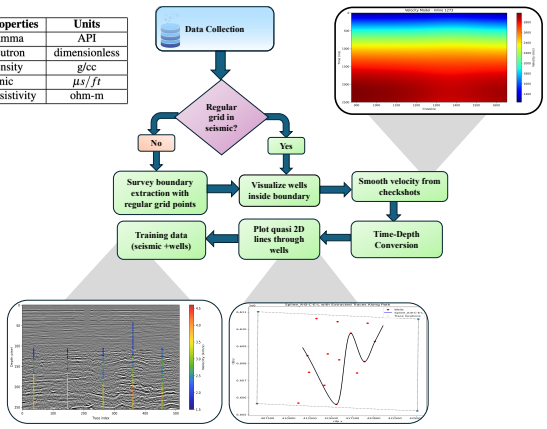


Figure 1: The flow diagram for data curation pipeline and the table shows the well logs along with their units present in the las files.

representative of the North Sea basin, while maintaining the ability to generalize across diverse marine environments. We have automated our data curation pipeline to produce structured and consistent datasets that can be used directly for training without additional preprocessing.

MACHINE LEARNING DATA CURATION PIPELINE

The data curation workflow, illustrated in figure 1, consists of four components, which are described below.

Data Collection

The UK-NDR portal enables bulk download of 3-D migrated seismic data, as illustrated in figure 2. For quality control, we prioritized surveys acquired after the 1990s, as earlier datasets often contain inconsistent SEG-Y headers, nonstandard formatting, and uncertain well coordinates that complicate automated geometry extraction and well–seismic alignment. A small number of pre-1990 surveys were included only after careful verification. For each selected survey, we downloaded SEG-Y volumes directly, and retrieved associated well data (LAS and checkshot files) using automated scripts that access the repository through APIs. This provides a scalable and reproducible workflow for downloading large seismic and well datasets.

Preprocessing of SEG-Y Data

After collecting the 3D seismic datasets, preprocessing is required to make the data suitable for generative model training. The first step involves aligning seismic volumes and well data within a common coordinate reference system (CRS) (Janssen, 2009), ensuring that wells are correctly positioned within their corresponding survey boundaries. To represent 3D seismic data on a regular grid, we first estimate survey boundaries using a concave hull to account for irregular acquisition geometries, and then extract the largest contiguous rectangular region with regular grids (GeeksforGeeks, 2025).

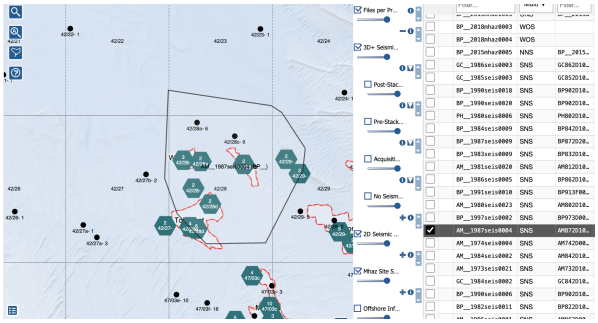


Figure 2: UKNDR GUI for seismic data filtering and downloading

Checkshot-Based Velocity Volume Construction for Time-Depth Seismic Conversion

Post-stack seismic volumes are migrated data and may be available in either the time or depth domain. In cases where the data are time migrated, a time-to-depth conversion is required for integration with well information. Hence, we propose to use checkshot data, which provide depth and corresponding two-way travel times, to compute average velocities for time-depth conversion of time migrated seismic volumes. These velocity estimates are then spatially interpolated using a Radial Basis function (RBF) (Skala, 2017) to construct a three dimensional average velocity volume as shown in figure 3.

The general RBF interpolant is defined as:

$$f(x) = \sum_{i=1}^{N} \lambda_i \, \phi(\|x - x_i\|) \tag{1}$$

- x represents a spatial location where velocity is estimated (e.g., grid point in the seismic volume: (x, y) or (x, y, z))
- x_i are the locations of known data points (checkshot positions)
- $f(x)$ represents interpolated value, i.e., velocity at location x
- λ_i weights determined from known velocities at the wells, and
- $\phi(\|x - x_i\|)$ radial basis function that depends on the distance between x and each data point x_i

We intentionally constructed an average velocity volume rather than an interval velocity, in order to obtain a smooth velocity field for depth conversion. Although interval velocity is an intrinsic rock property, our objective is not to construct highly accurate velocity models, but rather to obtain a consistent mapping between time and depth that enables alignment of seismic data with well information. In practice, estimating precise interval velocities is challenging (Kosloff et al., 2002), particularly in areas with sparse well control or limited data quality. Velocity information is typically derived from time–velocity measurements such as stacking velocities or checkshot data (Herron, 2011). Instead of focusing on exact velocity reconstruction, we aim to capture the overall statistical behavior of

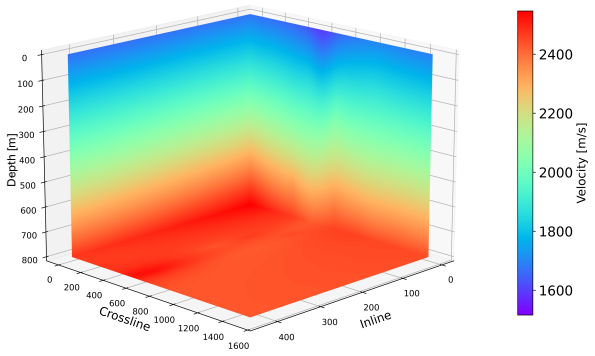
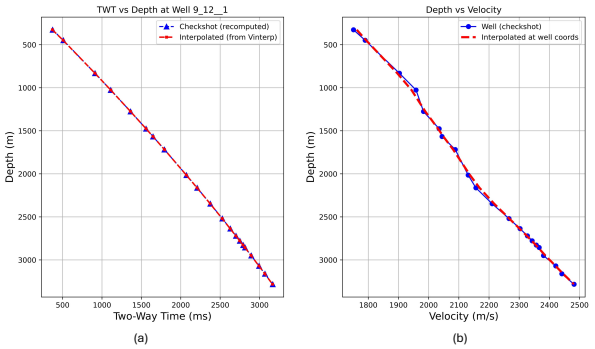


Figure 3: 3-D visualization of smooth velocity field constructed using checkshots



(a) Comparison of two-way travel time versus depth for checkshot data (blue curve) and the corresponding interpolated well velocity at the same location (red curve) (b) Comparison of depth vs velocity for checkshot measurements and the corresponding interpolated well velocity at the same location

Figure 4

the velocity field.

As part of quality control for the interpolated velocity model, we performed consistency checks between the original checkshot measurements and the interpolated velocity values at corresponding well locations shown in figure 4, comparing time–velocity and depth–velocity profiles derived from checkshots with those obtained from the interpolated average velocity volume at the same spatial positions. The depth versus velocity and two way travel time vs velocity trends show good agreement, indicating that the interpolated velocity model preserves the underlying well control concluding that the constructed average velocity volume remains consistent with measured checkshot data and is suitable for reliable time–depth conversion.

Time to Depth Conversion of Seismic Data

The time migrated seismic volumes are converted to depth in OpendTect (dGB Earth Sciences, 2026) and the depth converted seismic is illustrated in figure 6. The conversion from two way travel time to depth is governed by the fundamental

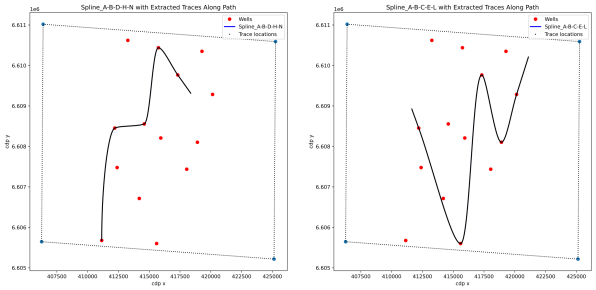


Figure 5: Quasi-2D lines passing through different well locations

relationship between velocity and travel time. For vertically propagating waves, depth is obtained by integrating interval velocity over time: $z(t) = \int_0^t \frac{v(\tau)}{2} d\tau$ where, $z(t)$ denotes depth, $v(\tau)$ is the interval velocity as a function of time, and t represents the two-way travel time. The factor of $\frac{1}{2}$ accounts for the fact that seismic data are recorded in two way travel time. The depth at time sample t_k is computed as $z_k = \sum_{j=1}^{k} \frac{v_j}{2} \times \Delta t$. Although, not accurate, but this limitation is acceptable as the objective is to learn the statistical distribution of velocity models that at a later stage can be used to train generative neural networks to carry out seismic inference.

Training Data

The final step in the data curation pipeline is to generate two-dimensional depth-domain seismic sections from curated three-dimensional volumes. The well locations are first mapped within the seismic boundary, after which quasi-2D lines are constructed along the trajectories that pass through the selected wells as depicted in figure 5.

The extracted seismic sections corresponding to the quasi-2D lines were resampled to 256×512 using a 2D FFT, where a smooth low-pass filter with a cosine taper suppresses high-frequency components, preserving low frequencies while gradually attenuating frequencies to avoid sharp cutoffs. Since well logs typically have finer vertical sampling than seismic traces, filtering and resampling are applied to wells to harmonize resolutions with seismic before training of a diffusion based Gen-AI model. For details on this training, we refer to another abstract by the authors submitted to the proceedings of this conference. The final data set consists of 2-D seismic sections in depth-domain of dimension 256x512 with a 12.5m sampling interval integrated with well control, as illustrated in figure 7. All curated datasets are stored in HDF5 format.

RESULTS

A diffusion-based generative model (Erdinc et al., 2024) was initially trained on Compass dataset having the same dimension and sampling interval as these curated datasets. For the initial experiments, the model was retrained using our curated data as represented in figure 7, from only 40 wells, and was able to generate realistic velocity model distributions that closely align with the ground truth as shown in figure 8. These re-

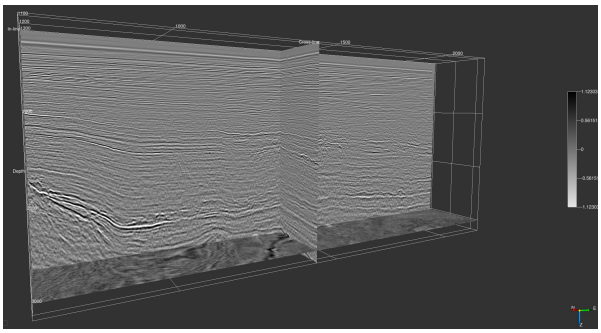


Figure 6: Seismic image after converting from Time to Depth domain visualised in OpendTect

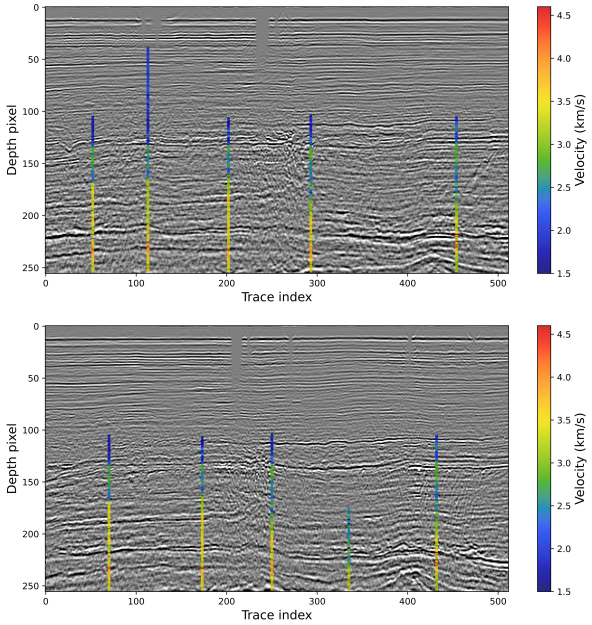
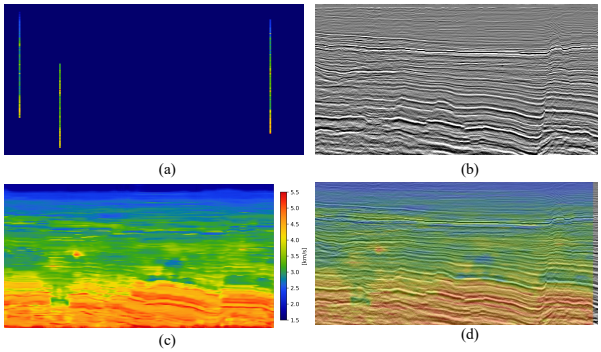


Figure 7: Both the figures show different sections of 2-D resampled seismic in depth tied to wells



(a) represents the ground truth velocity the network learns from the wells, (b) represents the corresponding seismic section, (c) is a velocity sample as reconstructed by the diffusion model, and (d) presents the overlaid velocity sample on the seismic image.

Figure 8

sults are preliminary and future work will involve scaling the training to at least 1000 real well logs to further improve the model's accuracy and generalization.

CONCLUSION

We introduce an automated data curation pipeline that streamlines seismic data preparation and by leveraging real field datasets, the proposed framework addresses key limitations of existing synthetic benchmark velocity models, including the lack of realistic geological complexity and variability. Unlike prior approaches that rely on deterministic models or simplified synthetic data, our methodology focuses on capturing the statistical characteristics from real seismic observations. These curated datasets include not only velocity but also additional properties such as density, enabling their use in evaluating multiparameter inversion algorithms in the future. Currently, we evaluate these datasets in a 2-D setting and are in the process of extending our framework to 3-D.

ACKNOWLEDGMENTS

This work was carried out in collaboration with the UK National Data Repository (NDR). *Contains information provided by the North Sea Transition Authority and/or other third parties.* During the preparation of this work, the authors used ChatGPT for language refinement and to improve readability. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

DATA AND MATERIALS AVAILABILITY

Data associated with this work are available here and the curated datasets with codes will be made available upon acceptance.

REFERENCES

Alaudah, Y., P. Michałowicz, M. Alfarraj, and G. AlRegib, 2019, A machine-learning benchmark for facies classification: Interpretation, **7**, SE175–SE187.

Chen, H., J. Chen, M. D. Sacchi, J. Gao, and P. Yang, 2025, Unsupervised seismic acoustic impedance inversion based on generative diffusion model: Geophysics, **90**, M109–M121.

Consolvo, B., 2023, Seismic data to subsurface models with openfwi: Training an ai model on the latest intel xeon cpu with pytorch 2.0: `https://medium.com/better-programming/seismic-data-to-subsurface-models-with-openfwi-bc8a02135428` (Medium article, accessed 2026).

Deng, C., S. Feng, H. Wang, X. Zhang, P. Jin, Y. Feng, Q. Zeng, Y. Chen, and Y. Lin, 2022, Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion: Presented at the Advances in Neural Information Processing Systems (NeurIPS), Curran Associates, Inc.

dGB Earth Sciences, 2026, Opendtect pro & dgb plugins documentation - 7.0: `https://doc.opendtect.org/7.0.0/doc/dgb_userdoc/Default.htm`. (Accessed: 2026-02-12).

Donoho, D., 2024, Data Science at the Singularity: Harvard Data Science Review, **6**. (https://hdsr.mitpress.mit.edu/pub/g9mau4m0).

Erdinc, H. T., R. Orozco, and F. J. Herrmann, 2024, Generative geostatistical modeling from incomplete well and imaged seismic observations with diffusion models: arXiv preprint arXiv:2406.05136.

Fehler, M., and A. Cheng, 2007, Seam: The seg advanced modeling project, phase i: AGU Fall Meeting Abstracts.

Gahlot, A. P., R. Orozco, and F. J. Herrmann, 2025, Advancing geological carbon storage monitoring with 3d digital shadow technology: arXiv preprint arXiv:2502.07169.

GeeksforGeeks, 2025, Largest rectangular area in a histogram using stack: `https://www.geeksforgeeks.org/dsa/largest-rectangular-area-in-a-histogram-using-stack/`. (Accessed: 2026).

Herron, D. A., 2011, First steps in seismic interpretation: Society of Exploration Geophysicists.

Janssen, V., 2009, Understanding coordinate reference systems, datums and transformations: International Journal of Geoinformatics, **5**, 41–53.

Jin, P., Y. Feng, S. Feng, H. Wang, Y. Chen, B. Consolvo, Z. Liu, and Y. Lin, 2024, An empirical study of large-scale data-driven full waveform inversion: Scientific Reports, **14**, 20034.

Jones, C. E., J. A. Edgar, J. I. Selvage, and H. Crook, 2012, Building complex synthetic models to evaluate acquisition geometries and velocity inversion technologies: 74th EAGE Conference and Exhibition Incorporating EUROPEC 2012, European Association of Geoscientists & Engineers, cp–293–00580.

Kosloff, D., Y. Sudman, and J. W. C. Sherwood, 2002, Velocity and interface depth determination by tomography of depth migrated gathers: Geophysics, **67**, 1388–1399.

Mekonnin, A., K. Wacławiak, M. Humayun, S. Zhang, and H. Ullah, 2025, Hydrogen storage technology, and its challenges: A review: Catalysts, **15**, 260.

North Sea Transition Authority, 2026, Uk national data repository: `https://www.nstauthority.co.uk/data-and-insights/data/uk-national-data-repository/`. (Contains information provided by the North Sea Transition Authority and/or other third parties).

Orozco, R., A. Siahkoohi, M. Louboutin, and F. J. Herrmann, 2025, Aspire: Iterative amortized posterior inference for bayesian inverse problems: Inverse Problems, **41**, 045001.

Pangman, P., 2014, Seam phase ii—land seismic challenges: The Leading Edge, **33**, 828–830.

Skala, V., 2017, Radial basis function interpolation and applications: An incremental approach: Latest Trends on Applied Mathematics, Simulation, Modelling, 1–8.

Yin, Z., R. Orozco, and F. J. Herrmann, 2025, Wiser: Multimodal variational inference for full-waveform inversion without dimensionality reduction: Geophysics, **90**, A1–A7.

Yin, Z., R. Orozco, M. Louboutin, and F. J. Herrmann, 2024, Wise: Full-waveform variational inference via subsurface extensions: Geophysics, **89**, A23–A28.