Deep learning for seismic velocity estimation and mapping from sparse data in underground environment

Apprentissage profond pour l'estimation et la cartographie des vitesses sismiques à partir de données éparses en environnement souterrain

Valentin Tschannen[1], *Kilyann* Richard[2], *Élodie* Morgan[2], *Tristan* Barbagelata[1], *Timothée* Lamare[3], *Béatrice* Yven[3], *Julien* Cotton[3*]

[1]Aquila Data Enabler, 92400 Courbevoie, France
[2]SpotLight, 91300 Massy, France
[3]Andra, 92298 Châtenay-Malabry Cedex, France

Abstract. Characterizing rocks around underground engineered structures without drilling requires weakly intrusive sensing and methods able to exploit sparse, non-conventional seismic data. This constraint is particularly strong in High Level Waste (HLW) cells, where sensor placement is limited by design requirements and by the need to preserve the surrounding rock, resulting in reduced illumination and data that are not well suited to standard data processing. To explore alternatives, a learning-based inversion framework was developed using stochastic velocity models, finite-difference simulations and spectral conditioning. A U-Net architecture trained on numerous synthetic datasets reconstructs coherent velocity maps despite sparse geometries, illustrating both the potential and the current limitations of deep learning approaches for such constrained underground acquisitions.

Résumé. Caractériser la roche autour d'ouvrages souterrains sans recourir à des forages impose des dispositifs faiblement intrusifs et des méthodes capables d'exploiter des données sismiques non conventionnelles. Cette contrainte est marquée dans les alvéoles HA (Haute Activité), où la pose des capteurs est fortement limitée par le concept de stockage et la préservation de la roche, rendant ces données peu adaptées à une tomographie classique. Pour explorer des alternatives, un schéma d'inversion par apprentissage a été développé à partir de nombreux modèles stochastiques de vitesse, de simulations en différences finies et d'un conditionnement spectral. Entraîné sur ces jeux synthétiques, un réseau U-Net a reconstruit des cartes de vitesse cohérentes malgré la géométrie contrainte des données d'entrée, illustrant le potentiel et les limites actuelles des approches d'apprentissage profond dans ce type d'acquisition.

* Corresponding author: julien.cotton@andra.fr

1 Introduction

Seismic velocities provide an integrated indicator of rock mechanical behavior and can be useful for mapping local variations around underground structures. However, seismic surveys in underground environments are often constrained by limited access, restricted sensor placement, and reduced illumination of the medium. As a result, the recorded datasets tend to be sparse, with low aperture and limited redundancy, which complicates their use for imaging purposes.

Conventional approaches for estimating seismic velocities include methods based on travel-time analysis or full-waveform inversion. While effective under favourable acquisition geometries, these techniques face limitations when data coverage is restricted. Travel-time tomography requires sufficient angular diversity and redundancy, which are rarely achievable around excavated structures.

Recent advances in deep learning have demonstrated the ability of architectures, such as U-Net models, to reconstruct spatial properties from seismic waveforms, using either real or synthetic datasets. Such methods are attractive for constrained underground configurations, as they can exploit large volumes of simulated data and provide rapid inference once trained. Despite this potential, relatively few studies have examined their performance in low-aperture settings with acquisition geometries typical of underground environments.

1.1 The Excavation-Damaged Zone (EDZ) characterization at Andra's Underground Research Laboratory (URL)

The excavation of underground structures (cells, drifts, boreholes) at Andra's URL induces a damaged zone (EDZ) in the surrounding claystone, characterised by stress redistribution, microcracking and local variations in elastic properties. This phenomenon has been documented by Andra for more than two decades through mechanical, hydraulic and geophysical investigations, providing a well-established understanding of its limited spatial extent around structures [1-2]. In this study, seismic measurements are not intended to identify the EDZ itself, but rather to examine whether velocity variations derived from sparse datasets can contribute to confirming or refining its estimated extent around the investigated structure.

2 Acquisition settings

The acquisition setup was designed to record seismic wavefields around a full-scale HLW cell demonstrator at Andra's URL. Seismic excitation was produced using a TIRA inertial vibration exciter operated at approximately 90 % of its nominal capacity, corresponding to a force of about 550 N. The source generated a linear sweep from 2000 Hz to 200 Hz over a duration of 9 s and was applied directly against the wall of the access drift. Wave propagation was recorded using Distributed Acoustic Sensing (DAS) with an optical fiber installed in the annular space between the steel liner and the claystone (Figure 1). The interrogator provided measurements every 0.5 m along the fiber, with a temporal sampling interval of 200 µs, yielding a dense but strongly constrained dataset aligned with the cell. For research purposes, the same fiber was also deployed in parallel boreholes located near the demonstrator, using an in-and-out geometry in each borehole to obtain additional reference traces (these boreholes would not be feasible in an operational setting, to preserve the rock surrounding the HLW cells).

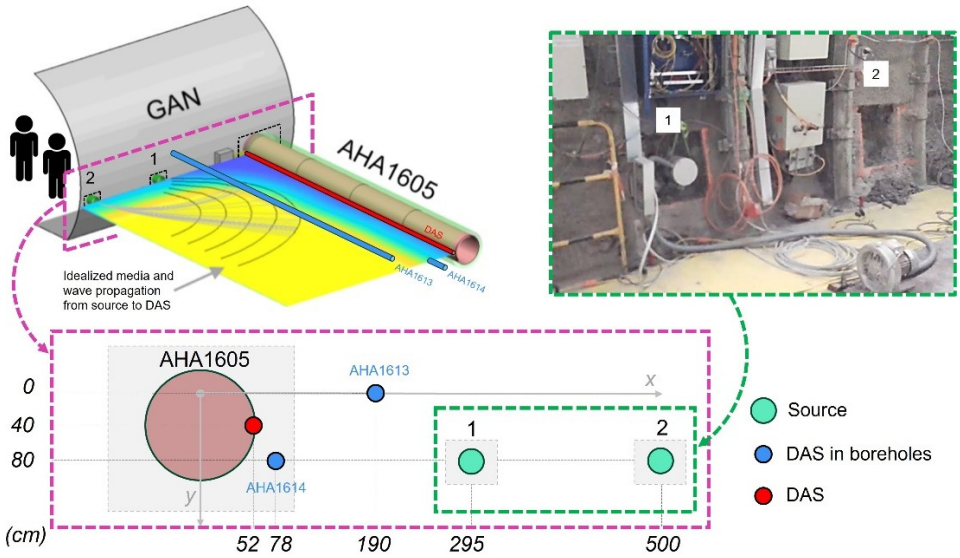

Fig. 1. Acquisition scheme and setup with a photograph showing the source positions

2.1 Data quality and geometrical control

The data quality can be assessed in Figure 2. The V-shaped pattern in each gather results from the in-and-out deployment of the optical fiber both in the vicinity of the HLW demonstrator and in the surrounding boreholes. Along the first pass, the distance to the source increases and first-arrival times lengthen; on the return pass, offsets decrease again, producing a symmetric trend.

Although the first arrival can be identified and picked, the associated expected uncertainty mainly reflects the width and shape of the source wavelet, its phase and, potentially, variations induced by the incidence angle of the wave on the optical fiber [3]. These factors affect the precision with which the onset of the wavelet can be located in time.

Interpreting these arrivals in terms of a two-zone structure composed of intact Callovo–Oxfordian (COx) claystone and an EDZ would require additional assumptions. In particular, the velocities in the intact COx and in the EDZ would need to be fixed or prescribed. Under this hypothesis, the first-arrival time could be expressed as the sum of travel times in each medium, allowing the source-to-interface distance to be determined geometrically. Such an interpretation depends directly on accurate knowledge of the source–receiver offset, and it also idealizes the transition, which may in reality be smooth rather than abrupt.

These assumptions define the conditions under which this simple geometric reasoning can be used for coarse consistency checks or to explore the sensitivity of inferred quantities to uncertainties in velocities, picking and geometry. By explicitly characterizing the uncertainties associated with each term, for instance within a probabilistic formulation treating them as distributions rather than fixed values, such approaches could in the future provide structured prior information within a broader inversion workflow and help guide or stabilize more advanced learning-based methods as discussed further.

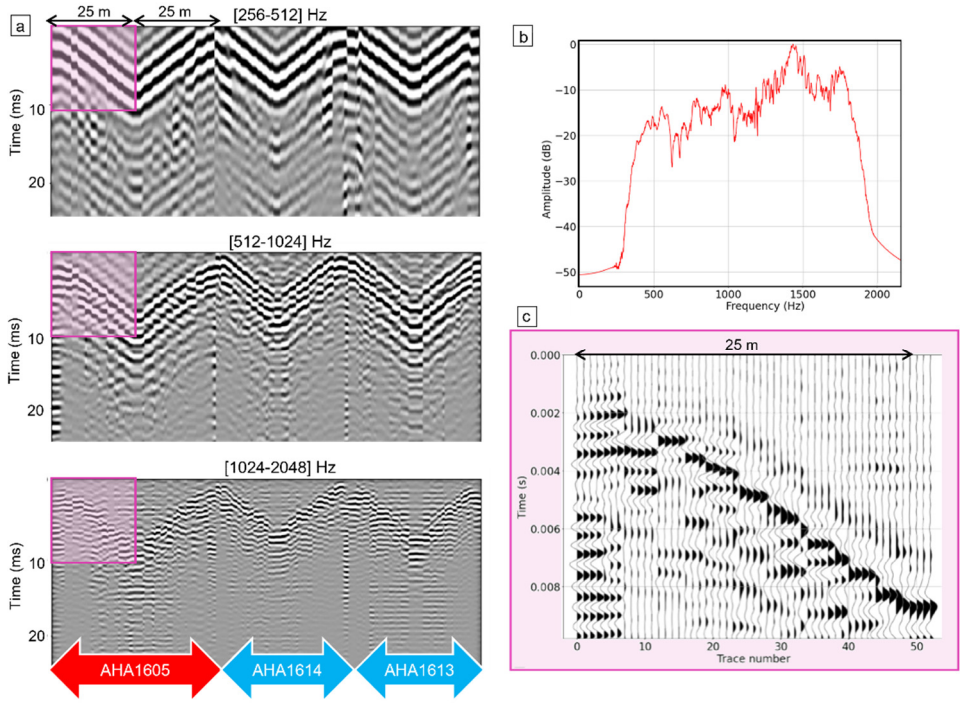

Fig. 2. Overall seismic data quality. a) DAS records decomposed into octave bands (from low to high frequencies, top to bottom); the pink inset marks the early-arrival window enlarged in c). b) Amplitude spectrum of the recorded signal. c) Time–distance zoom (first 10 ms over a 25 m fiber segment) highlighting the direct P-wave first break.

2.2 Data Generation for deep learning

The synthetic training dataset was constructed through a stochastic workflow designed to generate a broad range of spatial velocity variations representative of possible subsurface conditions around the demonstrator. Spatial velocity maps were produced using Gaussian random fields with prescribed covariance structures. Several covariance kernels were explored, including Gaussian, exponential and Matern families, to vary smoothness, correlation length and anisotropy. Sampling was performed using algorithms suited for large-scale random field generation, in particular the RMWSPy method [4] and the GPRFS strategy [5]. This stochastic approach was essential to avoid imposing any specific structural pattern on the learning process: by training in random fields rather than on predefined geological geometries, the network is exposed to a wide diversity of plausible configurations (Figure 3) and is less likely to memorize shapes or artefacts. This reduces inductive bias and encourages the model to learn general relationships between waveform characteristics and the underlying velocity structure.

For each stochastically generated velocity realization, seismic wave propagation was simulated using finite-difference time-domain solvers. Several implementations were assessed, including Deepwave [6] and CREWES [7], and the modelling was carried out with SimWave [8], a GPU-optimized scheme offering second-order accuracy in time and fourth-order accuracy in space, together with perfectly matched layers for boundary absorption. The acquisition geometry (source and receiver layout, offsets and sampling) was configured to reproduce the essential characteristics of the field experiment around the HLW cell. The resulting output consisted of multi-source synthetic seismograms paired with their exact

underlying velocity fields (Figure 4). The synthetic gathers display coherent first arrivals and well-formed diffractions consistent with velocity heterogeneities. These features confirm that the numerical modelling produces data of appropriate quality for testing the inversion workflow.

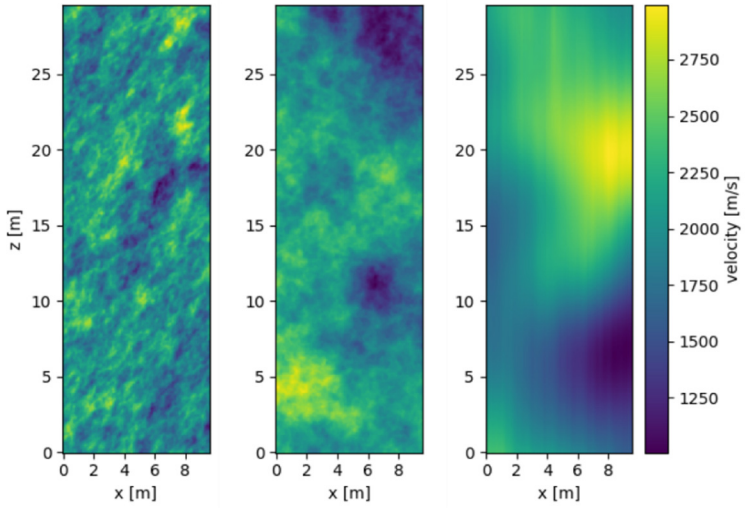

Fig. 3. Examples of random velocity models generated from two-dimensional Gaussian fields.

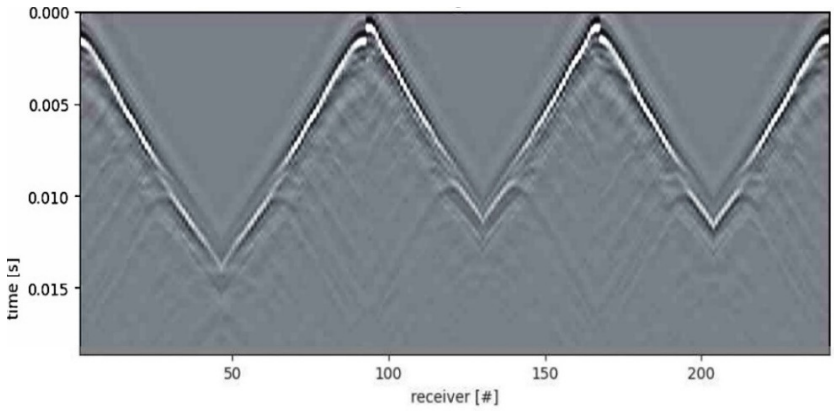

Fig. 4. A synthetic seismic simulation corresponding to one of the numerous generated velocity models, as it would be recorded by the acquisition setup described earlier.

Because the numerical simulations do not include the acquisition chain response or the exact source signature, a spectral adaptation stage was applied. A representative wavelet was estimated from the field dataset through three complementary approaches: comparison with the theoretical sweep used during acquisition, autocorrelation of uncorrelated traces, and statistical extraction from pre-processed records. These wavelets differ in bandwidth and sensitivity to noise, but each provides a realistic approximation of the operational source signature. Synthetic traces were convolved with the selected wavelet so that their amplitude and frequency characteristics match the field recordings. This ensures that the training waveforms reflect both the physical propagation effects and the spectral shaping introduced by the real acquisition system.

2.3 Inversion by deep learning

The inversion task was formulated as a supervised deep learning problem in which the objective is to estimate a spatial map of seismic wave propagation velocity from multi source seismic waveforms [9-10]. The training pairs consisted of synthetic seismograms generated for a given ground truth velocity field. Prior to training, each waveform was normalized and arranged in a fixed input structure that preserved the source dependent organization of the shots. The velocity fields were also normalized to stabilize the optimization and to ensure that the learning process focused on relative variations rather than on absolute amplitude differences. The network architecture (Figure 5) adopted for the inversion was based on the U-Net structure [11].

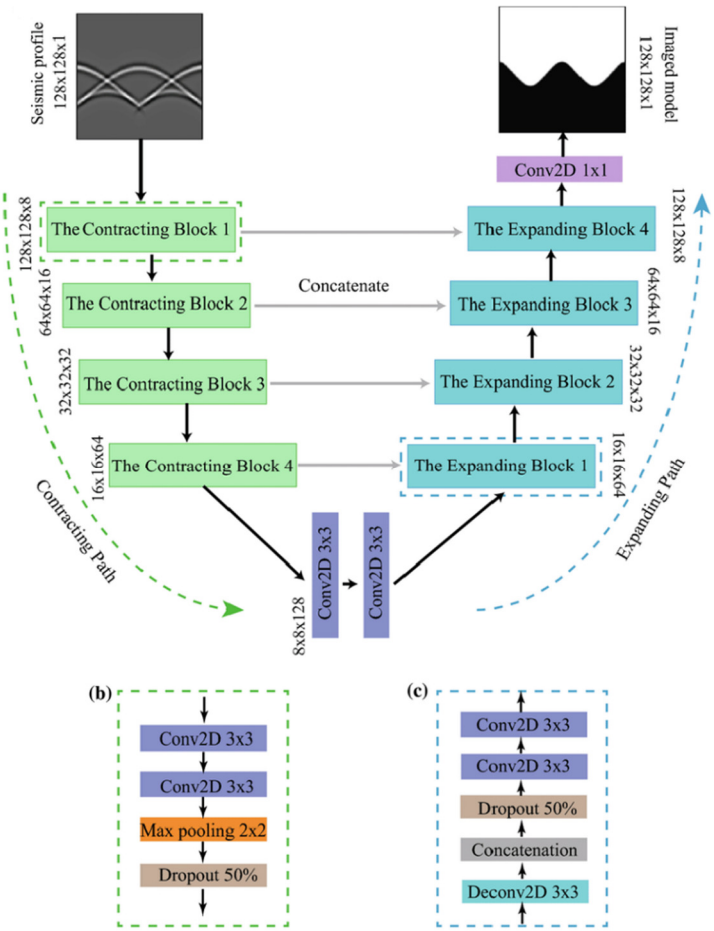

Fig. 5. U-Net architecture used for seismic inversion, showing tensor dimensions, convolution blocks, pooling and upsampling operations, and skip connections.

This fully convolutional encoder decoder formulation preserves spatial resolution while enabling multi scale feature extraction. Convolutional blocks were combined with instance normalization and rectified linear activation to maintain numerical stability. The encoder progressively extracted features from the waveform tensor, which incorporates spatial and temporal correlations, while the decoder reconstructed a two-dimensional velocity map with the same grid size as the synthetic models. The dataset was divided into training, validation

and test subsets to monitor convergence and assess generalization. Reconstruction quality was evaluated through quantitative metrics and visual inspection of the predicted velocity maps. Several configurations were benchmarked, and the selected architecture provided a satisfactory compromise between computational cost and reconstruction fidelity.

Training required high performance computing resources since the workflow combined large synthetic datasets, forward simulations for each stochastic model and the exploration of multiple variants of the U-Net architecture. The computational tasks were carried out on the Jean Zay supercomputer of GENCI and IDRIS, which provided both the GPU and CPU capacity needed to complete the data generation and learning phases under realistic time constraints.

2.4 Deep learning Inversion: Results on Synthetic Data

The trained network was then applied to synthetic data generated by forward simulation in stochastic velocity models. This step evaluates the behavior of the model under acquisition geometries and noise levels closer to the field experiment. The reconstructed maps exhibit consistent features and reproduce the main gradients present in the velocity models, although limitations remain, particularly those associated with the intrinsic two-dimensional formulation and the restricted illumination imposed by the acquisition geometry (Figure 6). As expected, reconstruction accuracy is highest in the well-illuminated region defined by the source–receiver geometry. The network recovers the main velocity trends where data coverage is sufficient and does not introduce spurious high-wavenumber content beyond what the acquisition can resolve.

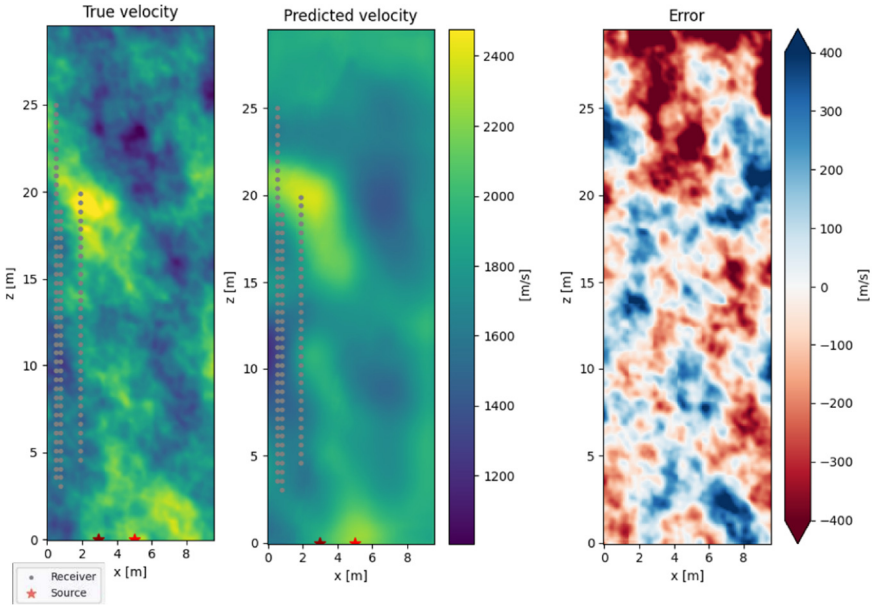

Fig. 6. Example of stochastic velocity-model reconstruction from synthetic data. The figure displays the true velocity model, the network prediction, and the corresponding reconstruction error. The areas with limited illumination show larger residuals.

2.5 Generalization Assessment on Pseudo-real Data

A second evaluation of the learning-based inversion framework was performed on a dataset derived from the CDZ (Compression of Damaged Zone) experiment conducted at Andra's

URL [2]. Several seismic acquisitions were performed comprising a baseline survey, followed by multiple monitor acquisitions conducted while increasing mechanical loading applied to the gallery wall. The aim of the experiment was to measure first arrivals under varying stress conditions to perform tomography and quantify structural changes within the damaged zone. In contrast with the sparse DAS acquisition deployed around the HLW demonstrator previously presented, the configuration relied here on small impulsive seismic sources placed directly inside boreholes, together with multiple lines of receivers also installed in boreholes. This arrangement provided significantly richer ray coverage and a more favorable illumination of the medium, enabling tomography. The available tomographic velocity inversion of the CDZ dataset was here used to generate synthetic seismic data through finite-difference simulations. The acquisition geometry, including the number of sources, their depth positions and the distribution of receivers, reproduced the CDZ configuration. Impulsive sources were modelled, and a spectral adaptation step was applied to match the frequency content of the CDZ data. These numerically generated seismograms constitute a "pseudo-real dataset", that can be seen as an intermediate test case between the fully synthetic examples generated from stochastic fields (previous part) and a real recorded dataset. The pseudo-real dataset provides an opportunity to evaluate the workflow in a different acquisition context involving repeated seismic measurements, while keeping noise-free inputs since the wavefields are produced by ideal finite-difference propagation in the tomographic velocity model. Applied to this dataset, the network reconstructs the main velocity structures present in the tomographic reference model. The geometry of the low-velocity zones and the overall velocity gradients are correctly recovered, providing a mathematical validation of the workflow under controlled conditions (Figure 7).

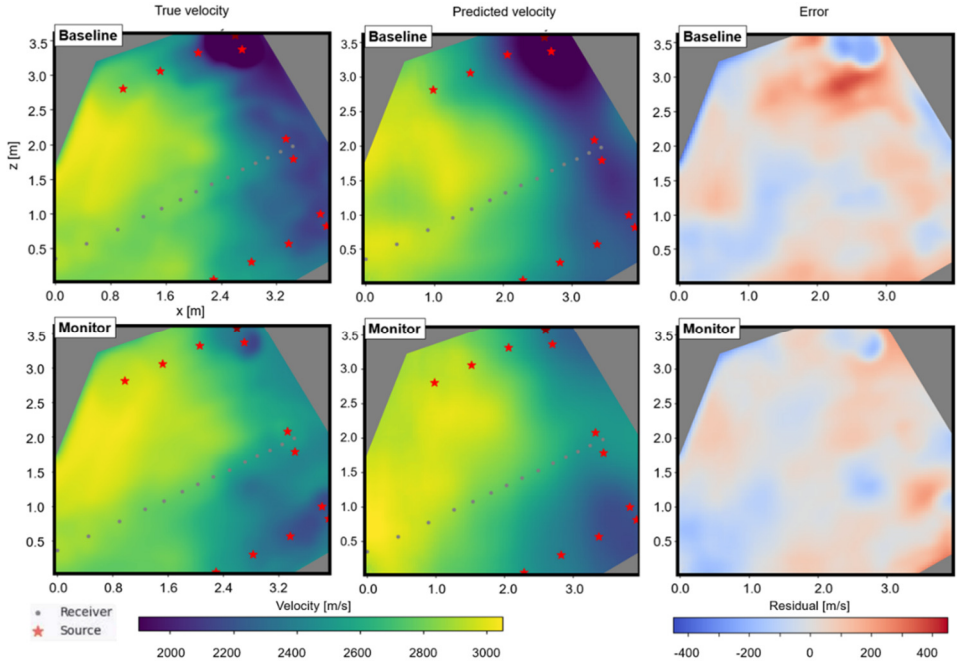

Fig. 7. Deep learning velocity inversion using pseudo-real CDZ data. Rows correspond to the baseline and to a later monitor (performed under mechanical loading). From left to right: 1) reference tomographic velocity model, 2) network velocity prediction, 3) spatial distribution of residuals.

3 Discussions

The application of the learning-based inversion framework to real DAS data did not yield conclusive results at this stage. Several factors contribute to this outcome. The field acquisition around the HLW demonstrator provides limited illumination, with few usable source–receiver paths and strong geometric constraints imposed by sensor placement. The resulting wavefields exhibit reduced redundancy, significant noise (aliasing) and partial coupling variations along the fiber, which differ markedly from the conditions represented in the synthetic training set. In addition, the available real dataset lacks the diversity required to support a reliable empirical fine-tuning stage, and the discrepancy between two-dimensional training and three-dimensional wave propagation further increases the mismatch between training and application domains.

These limitations suggest several directions for improvement. The training dataset could be enriched through targeted data augmentation strategies, including variations in acquisition geometry, noise levels, coupling conditions and source signatures, to expose the network to a broader set of realistic perturbations. Introducing three dimensional synthetic simulations for selected configurations could also reduce the gap between the training and application domains. Overall, the transition from synthetic or pseudo-real scenarios to real DAS data requires further adaptation of the training strategy and of the acquisition design. The results obtained so far indicate that these developments are necessary before robust reconstruction can be expected in the current configuration.

3.1 Perspectives on probabilistic priors for learning-based inversion

The simplified geometric formulation of travel times presented earlier, when expressed within a probabilistic or Bayesian framework, can provide structured prior information for learning-based inversion. In practice, a first stage would focus on estimating EDZ thickness and its uncertainty from travel-time data, using a lightweight model that ingests source–receiver geometry, picked arrivals and associated uncertainty distributions. The velocity distributions of the EDZ and of the intact claystone could be incorporated in the same manner, making all first-order quantities explicitly probabilistic and consistent with available constraints. This first-stage module could rely on a conditional variational auto-encoder [12] which appears well suited for integrating uncertain inputs, producing probabilistic outputs and capturing multi-modal solutions while remaining flexible under limited acquisition constraints. Such an approach would generate a distribution of admissible EDZ configurations rather than a single deterministic estimate. A dedicated training phase on synthetic datasets, for which true geometries and uncertainty distributions are fully controlled, would be used to validate this probabilistic mapping. In a second stage, the resulting distributions could be injected as soft priors into a more expressive network operating on full waveforms, such as the U-Net-based velocity inversion workflow. The inversion would then refine a physically plausible domain instead of exploring the full model space, which may stabilize training under sparse illumination, mitigate sensitivity to waveform variability and preserve uncertainty throughout the process. This hybrid strategy would help constrain the inversion without imposing a fixed structure, while at the same time providing uncertainty estimates that remain interpretable in the context of sparse underground seismic acquisitions.

4 Conclusion

This study assessed a learning-based approach for estimating seismic velocity fields under sparse and constrained acquisition conditions relevant to underground structures. Synthetic

tests showed that the method can recover the main features of heterogeneous velocity models, and pseudo-real experiments based on the CDZ configuration indicated that generalization to more realistic structures is achievable. Results obtained on real DAS data remain inconclusive, reflecting limited illumination, noise and discrepancies between two-dimensional training and three-dimensional propagation. These observations suggest that further work is needed on data augmentation, physics guided constraints and adaptation of the acquisition geometry. Simple geometric analyses may also support hybrid strategies by constraining the range of admissible models. Overall, while promising, the approach requires additional developments before robust application to field data in such settings can be expected.

Acknowledgements

Financial support from BPI France within the framework of the France Relance program is gratefully acknowledged.

This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011014459 made by GENCI.

References

1. G. Armand, F. Leveau, C. Nussbaum, R. de La Vaissière, A. Noiret, D. Jaeggi & C. Righini, *Geometry and properties of the excavation-induced fractures at the Meuse/Haute-Marne URL drifts*, Rock Mech. Rock Eng., **47**, 21-41 (2014).
2. R. de La Vaissière, J. Morel, A. Noiret, P. Côte, B. Helmlinger, R. Sohrabi & C. Nussbaum, *Excavation-induced fractures network surrounding tunnel: properties and evolution under loading*, Geol. Soc. Spec. Publ., **400**, 279–291 (2014).
3. A. H. Hartog, *An introduction to distributed optical fibre sensors*, CRC Press (2017)
4. S. Hörning & B. Haese, *RMWSPy (v 1.1): A Python code for spatial simulation and inversion for environmental applications*, Environ. Model. Softw., **138**, 104970 (2021)
5. L. Räss, D. Kolyukhin & A. Minakov, *Efficient parallel random field generator for large 3D geophysical problems*, Comput. Geosci., **131**, 158–169 (2019)
6. A. Richardson, *Deepwave (v0.0.20)*, Zenodo (2023)
7. G. F. Margrave & M. P. Lamoureux, *Numerical methods of exploration seismology: With algorithms in MATLAB®*, Cambridge University Press (2019)
8. J. Freire de Souza, J. B. D. Moreira, K. J. Roberts, R. d. R. A. Gaioso, E. S. Gomi, E. C. N. Silva & H. Senger, *simwave – A Finite Difference Simulator for Acoustic Waves Propagation*, arXiv, 2201.05278 (2022).
9. F. Yang & J. Ma, *Deep-learning inversion: A next-generation seismic velocity model building method*, Geophysics, **84**, R583–R599 (2019)
10. S. Li, B. Liu, Y. Ren, Y. Chen, S. Yang, Y. Wang & P. Jiang, *Deep-learning inversion of seismic data*, arXiv, 1901.07733 (2019)
11. O. Ronneberger, P. Fischer & T. Brox, *U-net: Convolutional networks for biomedical image segmentation*, Med. Image Comput. Comput.-Assist. Interv., **9351**, 234–241 (2015)
12. K. Sohn, H. Lee & X. Yan, *Learning structured output representation using deep conditional generative models*, Adv. Neural Inf. Process. Syst., **28** (2015).