https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

OpenSWI: A Massive-Scale Benchmark Dataset for Surface Wave Dispersion Curve Inversion

Feng Liu[1,2], Sijie Zhao[2,3], Xinyu Gu[2], Fenghua Ling[2], Peiqin Zhuang[2], Yaxing Li[4], Rui Su[2], Lihua Fang[5], Lianqing Zhou[5], Jianping Huang[4], and Lei Bai[2]

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[3]School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China
[4]Chengdu University of Technology, Chengdu 610059, China
[5]Institute of Earthquake Forecasting, China Earthquake Administration, Beijing 100036, China

Correspondence: Yaxing Li (yxli2024@cdut.edu.cn) and Rui Su (surui@pjlab.org.cn)

Abstract. Surface wave dispersion curve inversion plays a critical role in both shallow geophysical exploration and deep geological studies, yet it remains hindered by sensitivity to initial models, susceptibility to local minima, and low computational efficiency. Recently, data-driven deep learning methods, inspired by their success in computer vision and natural language processing, have shown promising potential to overcome these challenges. However, the lack of large-scale and diverse benchmark datasets remains a major obstacle to the development and evaluation of such methods. To address this gap, we introduce **OpenSWI**, a comprehensive benchmark dataset generated through the Surface Wave Inversion Dataset Preparation (SWIDP) pipeline. OpenSWI comprises two synthetic datasets tailored to different research scales and application scenarios, namely **OpenSWI-shallow** and **OpenSWI-deep**, as well as an AI-ready real-world dataset for generalization evaluation, **OpenSWI-real**. OpenSWI-shallow is derived from the 2-D geological model dataset OpenFWI, containing over 22 million 1-D velocity profiles paired with their fundamental-mode phase and group velocity dispersion curves, spanning a broad spectrum of shallow geological structures (e.g., flat layers, faults, folds, and realistic stratigraphy). OpenSWI-deep is built from 14 global and regional 3-D geological models, comprising approximately 1.26 million high-fidelity 1-D velocity-dispersion data pairs for deep earth studies. OpenSWI-real, compiled from open-source projects, contains two sets of observed dispersion curves and their corresponding 1-D reference models, serving as a benchmark for evaluating the generalization of deep learning models. To demonstrate the utility of OpenSWI, we trained deep learning models on OpenSWI-shallow and OpenSWI-deep, and evaluated them on OpenSWI-real. The results show strong agreement between the predicted and reference velocity models, confirming the diversity and representativeness of the OpenSWI dataset. To facilitate the advancement of intelligent surface wave dispersion curve inversion techniques, we release the OpenSWI dataset (https://doi.org/10.5281/zenodo.16874111) and the SWIDP toolbox along with associated resources (https://doi.org/10.5281/zenodo.16884901), providing open resources to support the research community.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

1 Introduction

Surface wave dispersion curve inversion is a fundamental geophysical technique for reconstructing subsurface shear wave velocity profiles by fitting theoretical dispersion curves to measured data (Xia et al., 1999; Shapiro and Campillo, 2004; Wathelet et al., 2004). It is widely applied in shallow engineering surveys, including site response and microzonation studies (Park et al., 1999; Socco and Strobbia, 2004; Foti et al., 2014), as well as in studies of lithospheric structure and evolution at greater depths (Shapiro and Ritzwoller, 2002; Shapiro and Campillo, 2004; Yang and Ritzwoller, 2008). In shallow subsurface investigations, this technique is valuable for identifying complex geological features such as weathering layers and overburden, while at greater depths, it provides critical insights into tectonic evolution (Reid et al., 2025). Despite its widespread applicability, traditional inversion methods are heavily dependent on initial models and nonlinear optimization, leading to high computational costs and susceptibility to getting trapped in local minima(Shapiro and Ritzwoller, 2002; Wathelet et al., 2004; Chen et al., 2025). These limitations hinder their applicability to large-scale, high-resolution imaging tasks.

In recent years, rapidly developing deep learning methods have revolutionized the process of surface wave dispersion curve inversion. These data-driven approaches leverage deep neural networks, such as fully connected networks (FNNs), convolutional neural networks (CNNs), and Transformer networks, to learn the mapping between dispersion curves and subsurface shear wave velocity profiles (Hu et al., 2020; Yablokov et al., 2021; Wang et al., 2022; Cai et al., 2022; Huang et al., 2024; Liu et al., 2025; Jiang et al., 2025). By effectively eliminating reliance on initial models and iterative optimization, these methods significantly improve inversion efficiency and performance (Chen et al., 2025). Once trained, the models can rapidly invert large-scale datasets in seconds, making them well-suited for real-time applications, such as field deployment and imaging. However, their performance and generalization ability are strongly influenced by both the quality and diversity of the training data (Luo et al., 2022). Previous research has demonstrated that large-scale, diverse datasets substantially enhance deep model performance, particularly in scenarios with no labeled data (zero-shot learning) or limited labeled data that requiring fine-tuning (few-shot learning) (Luo et al., 2022; Liu et al., 2025). Therefore, the development of dispersion curve datasets that encompass representative geological features, multi-scale structures, and sufficient sample sizes is crucial for advancing intelligent inversion methods.

Despite the importance of diverse datasets for deep learning methods, the construction of benchmark datasets specifically for surface wave dispersion curve inversion remains limited. In contrast, other areas of seismic research have seen the successful creation of large-scale datasets. For instance, in seismic monitoring, datasets like STEAD (Mousavi et al., 2019) and INSTANCE (Michelini et al., 2021) contain millions of waveform data traces. Similarly, full-waveform inversion efforts have led to the creation of model collections such as OpenFWI (Deng et al., 2021) and EFWI (Feng et al., 2023), each comprising hundreds of thousands of geological velocity models. Seismic exploration has also benefited from the development of standardized workflows, leading to benchmark datasets like cigFacies (Gao et al., 2025) and cigChannels (Wang et al., 2025). However, in the specific domain of surface wave dispersion curve inversion, there is still a significant lack of representative, well-structured, and publicly accessible datasets. One of the main challenges lies in the necessity of paired dispersion curves and velocity profiles to generate high-quality training samples. Actual observational data are often proprietary and not available

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

to most of the researchers Merrifield et al. (2022). Moreover, observed dispersion curves are often compromised by limitations in observation conditions and subjective picking, resulting in issues such as noise contamination and data gaps (Socco and Strobbia, 2004; Bensen et al., 2007). Additionally, the non-uniqueness of the corresponding velocity profiles further complicates the development of supervised models (Foti et al., 2009), making it more difficult to train deep learning algorithms effectively.

To address these challenges, synthetic surface wave dispersion curve data have emerged as a feasible alternative. Synthetic data, generated through a series of forward modeling processes, can effectively simulate field-observed dispersion curves. Since the corresponding velocity profiles are known in the simulation, this method naturally avoids pairing errors. Deep neural networks trained on synthetic data have demonstrated good applicability and inversion performance in shallow subsurface geological exploration (Cao et al., 2020; Aleardi and Stucchi, 2021; Yablokov et al., 2021, 2023; Gan et al., 2024) and deep structural imaging (Hu et al., 2020; Wang et al., 2022; Huang et al., 2024; Jiang et al., 2025; Liu et al., 2025). However, existing publicly available datasets are still largely limited to specific geological features or particular regions, lacking sufficient geological diversity and regional coverage. Given the complexity of shallow geology and the regional variability of deep structures, constructing a synthetic dataset with greater geological complexity, broader coverage, and larger sample sizes is essential for improving the generalization ability and practical applicability of models.

In this paper, we introduce OpenSWI, a comprehensive benchmark dataset designed for surface wave dispersion curve inversion, developed through the dataset construction workflow SWIDP (Figure 1). OpenSWI includes two synthetic benchmark datasets, OpenSWI-shallow and OpenSWI-deep, each tailored to different research scales and application scenarios, as well as an AI-ready real-world dataset, OpenSWI-real, specifically for evaluating model generalization. The OpenSWI-shallow dataset, built upon the publicly available 2-D geological model dataset OpenFWI, incorporates a broad range of geological features, such as flat layers, faults, folds, and actual geological structures, containing approximately 22 million 1-D velocity profiles paired with their corresponding fundamental-mode surface wave dispersion curves. This makes it the largest and most geologically diverse dataset available for shallow subsurface studies. To further enhance structural diversity and sample variability, SWIDP integrates a Diffusion Probabilistic Model (DDPM), which learns the distribution of 2-D geological models and allows the continuous generation of more varied shallow subsurface data. The OpenSWI-deep dataset, generated by collecting, curating, and integrating 14 global and regional 3-D geological models, consists of approximately 1.26 million high-fidelity 1-D dispersion data samples, providing a large-scale benchmark for deep subsurface imaging tasks. OpenSWI-real, derived from two publicly available observational datasets and their reference velocity models, is directly applicable for performance testing and generalization validation of deep learning models in real-world applications. To evaluate the practical utility of these datasets, we trained two Transformer-based models using OpenSWI-shallow and OpenSWI-deep, then validated them on OpenSWI-real. Experimental results show that the inversion results of the trained models on real-world data are highly consistent with reference models, confirming the effectiveness and representativeness of the OpenSWI datasets for real-world applications. All datasets, along with the associated toolchain (including profile extraction, forward modeling and training examples), have been fully open-sourced, offering a reusable, high-quality benchmark platform for advancing future research in intelligent surface wave dispersion curve inversion.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

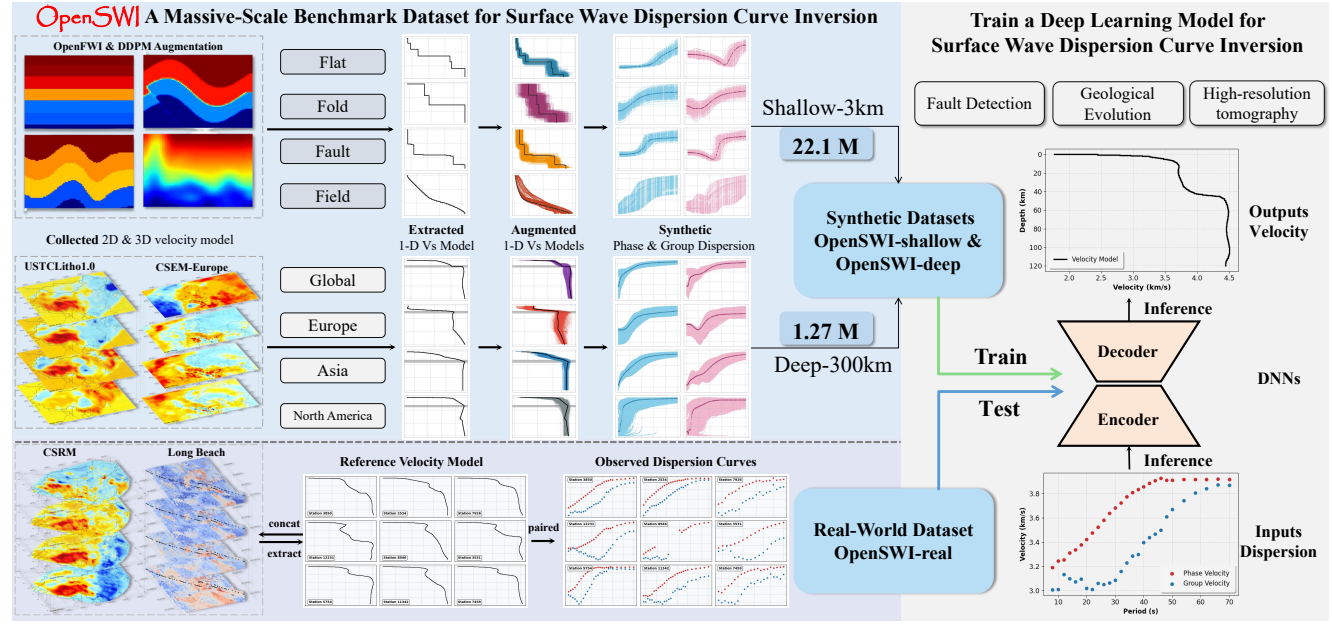


Figure 1. Overview of the workflow for constructing the OpenSWI benchmark datasets and their application in deep learning-based surface wave dispersion curve inversion. The workflow starts with the collection and quality control of raw data, followed by the extraction and augmentation of 1-D velocity profiles, and the simulation of dispersion curves to generate two synthetic datasets, OpenSWI-shallow and OpenSWI-deep, tailored for different research scales and application scenarios (blue box). To evaluate the generalization capability of deep learning models, a real-world dataset, OpenSWI-real, is also curated (purple box). Finally, a simple deep learning model, trained on the benchmark datasets, is applied to real observational data, as depicted in the white box on the right.

2 Construction of the Large-scale OpenSWI Benchmark Datasets

2.1 Integrated Workflow for Dataset Construction

We present an integrated workflow for constructing large-scale benchmark datasets for surface wave dispersion curve inversion. The workflow is designed to ensure geological diversity and realism of the data sources, employ modular and fully automated processing, and ensure high accuracy and computational efficiency in forward modeling. It encompasses all major stages, from the collection and standardization of raw geological models, through quality control and parameterization, to the simulation of fundamental-mode dispersion curves, providing a reproducible pipeline for large-scale dataset generation.

2.1.1 Collection and Quality Control of Geological Models

The first step in constructing a high-quality dataset for dispersion curve inversion is the collection of representative velocity models from diverse geological settings. These velocity models were primarily obtained from open-access geological databases and previously published studies, such as OpenFWI datasets (Deng et al., 2021)—which contain 2-D geological

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

models covering various sedimentary and tectonic settings—and LITHO1.0 geological models (Pasyanos et al., 2014), providing lithospheric-scale structural information. Table 1 summarizes the original data sources employed in this study. These rigorously curated and geologically validated models form a reliable foundation for constructing the OpenSWI datasets.

However, because the raw velocity models originated from different research groups and projects, they exhibited considerable variability in several aspects, such as data characteristics (e.g., depth range, spatial resolution), parameter types (e.g., S-wave velocity (v_s), P-wave velocity (v_p), or combined shear-wave velocities in both horizontal and vertical directions (v_{sv} and v_{sh})), and storage formats (e.g., `.npz`, `.txt`, or `.nc`). As a result, a unified quality control and standardization process was applied to ensure data consistency and representativeness before their use. The quality control procedures included the following steps:

1. **Data correction and cleaning:** Models containing missing or abnormal values (e.g., `zero` or `NaN` values in some geological models) were corrected through interpolation or single-point removal to ensure data completeness.

2. **Parameter conversion:** For models that provided only v_p, the corresponding v_s were estimated using the empirical relationships proposed by Brocher (2005). In cases where models included v_{sv} and v_{sh}, an equivalent v_s was derived using the geometric mean.

3. **Plausibility verification:** Geological structures within the models were systematically examined to remove anomalies inconsistent with geological principles or unsuitable for forward modeling.

These quality control measures substantially improved the accuracy and applicability of the geological models, thereby providing a robust and standardized data foundation for dispersion curve forward modeling and subsequent machine learning model training.

2.1.2 Extraction and Parameterization of 1-D Velocity Profiles

After completing the quality control and standardization of the geological models, the next step was to construct 1-D velocity profiles suitable for forward modeling. As illustrated in Figure 2, this process involved multiple stages, including profile extraction from 2-D or 3-D geological models, removal of redundant samples, structural rationalization, and parameter completion.

Each 1-D profile contains key physical parameters extending from the surface to the target depth range, including depth, S-wave velocity (v_s), P-wave velocity (v_p), and density (ρ). The procedure is described as follows:

1. **Extraction and de-duplication of 1-D profiles**: Vertical 1-D v_s profiles were extracted from 2-D geological cross-sections and 3-D geological models at each surface grid point (or a selected subset of grid points). In some geological models, particularly those with horizontally layered structures, duplicate or nearly identical profiles were identified. These duplicates were removed to ensure each sample was both unique and representative.

2. **Structure refinement of 1-D profiles**: To enhance numerical stability during forward modeling, abnormally thin layers or those with extreme velocity values were merged or smoothed. This step minimized non-physical artifacts in the simulations and improved the physical plausibility of the 1-D models.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Table 1. Original data sources used in constructing the OpenSWI datasets, summarizing dataset categories (e.g., OpenSWI-shallow, OpenSWI-deep, and OpenSWI-real), references, primary geological settings (e.g., Flat, Flat-Fault, Fold, Fold-Fault, and Field) or geographic coverage (e.g., global, China, Europe, the United States), recorded velocity parameters (e.g., P-wave velocity v_p, S-wave velocity v_s, combined shear-horizontal velocity v_{sh}, and shear-vertical velocity v_{sv}), as well as the size of the raw data, expressed as N velocity profiles $\times$ M model variables $\times$ 2-D velocity model shape (for OpenSWI-shallow) or L layers (for OpenSWI-deep and OpenSWI-real).

Group	Reference	Datasets	Geological Feature /Cover Region	Model Variable	Model Size
OpenSWI shallow	Deng et al. (2021)	OpenFWI-FlatVelA	Flat	v_p	30,000 × 1 × 70 × 70
		OpenFWI-Flat-FaultA	Flat + Fault	v_p	54,000 × 1 × 70 × 70
		OpenFWI-CurveVel	Fold	v_p	30,000 × 1 × 70 × 70
		OpenFWI-Fold-Fault	Fold + Fault	v_p	54,000 × 1 × 70 × 70
		OpenFWI-StyleA	Field	v_p	67,000 × 1 × 70 × 70
OpenSWI deep	Pasyanos et al. (2014)	LITHO1.0	Global	$depth$, v_s	40,962 × 2 × 96
	Xin et al. (2019)	USTClitho1.0	China	$depth$, v_s	9,125 × 2 × 12
	Shen et al. (2013)	Central-and-Western US	USA	$depth$, v_s	6,803 × 2 × 72
	Shen et al. (2016)	Continental China	China	$depth$, v_s	4,516 × 2 × 400
	Xie et al. (2018)	US Upper-Mantle	USA	$depth$, v_s	3,678 × 2 × 600
	Lu et al. (2018)	EUcrust	European	$depth$, v_s	43,520 × 2 × 80
	Berg et al. (2020)	Alaska	Alaska	$depth$, v_s	19,408 × 2 × 156
	Çubuk-Sabuncu et al. (2017) Blom et al. (2020)	CSEM-Europe	European	$depth$, v_{sh}, v_{sv}	21,931 × 3 × 61
	Blom et al. (2020)	CSEM-Eastmed	Eastern Mediterranean	$depth$, v_{sh}, v_{sv}	12,782 × 3 × 81
	Fichtner and Villaseñor (2015)	CSEM-Iberian	Western Mediterranean	$depth$, v_{sh}, v_{sv}	9,102 × 3 × 81
	Colli et al. (2013)	CSEM-South Atlantic	South Atlantic	$depth$, v_{sh}, v_{sv}	7,371 × 3 × 51
	Rickers et al. (2013) Krischer et al. (2018)	CSEM-North Atlantic	North Atlantic	$depth$, v_{sh}, v_{sv}	14,541 × 3 × 51
	Simutė et al. (2016)	CSEM-Japan	Japanese Island	$depth$, v_{sh}, v_{sv}	14,641 × 3 × 61
	Fichtner et al. (2009) Fichtner et al. (2010)	CSEM-Astralasia	Australasian	$depth$, v_{sh}, v_{sv}	4,131 × 3 × 51
OpenSWI real	Fu et al. (2022)	LongBeach	USA	$depth$, v_s	5,297 × 2 × 241
	Xiao et al. (2024)	CSRM	Continental China	$depth$, v_s	12,901 × 2 × 145

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

3. **Interpolation and standardization**: Uniform layer-thickness interpolation was applied to ensure model consistency across different application scenarios. For shallow subsurface models, layers were resampled at 40 m intervals, whereas for deep-Earth models, a coarser 1 km interval was adopted. This standardization facilitated large-scale batch processing and streamlined integration with deep learning frameworks. We note, however, that some studies may prefer non-uniform layer-thickness schemes (e.g., finer resolution in the shallow part and coarser resolution at greater depths). To support such flexibility, users can easily regenerate alternative dataset versions using the original construction scripts we provide.

4. **Completion of Other Physical Parameters**: To derive complete physical models, v_p and ρ were computed as follows. For depths < 120 km, empirical relationships from Brocher (2005) were used to estimate v_p and ρ based on known v_s values, ensuring physical consistency. For depths ≥ 120 km, where empirical formulas are less applicable, a constant Poisson's ratio of 1.79 was assumed to calculate v_p from v_s, with ρ subsequently derived from v_p using Brocher's empirical relationship (Brocher, 2005).

Through these steps, we generated a comprehensive collection of 1-D velocity profiles characterized by geological diversity, physical consistency, and numerical stability.

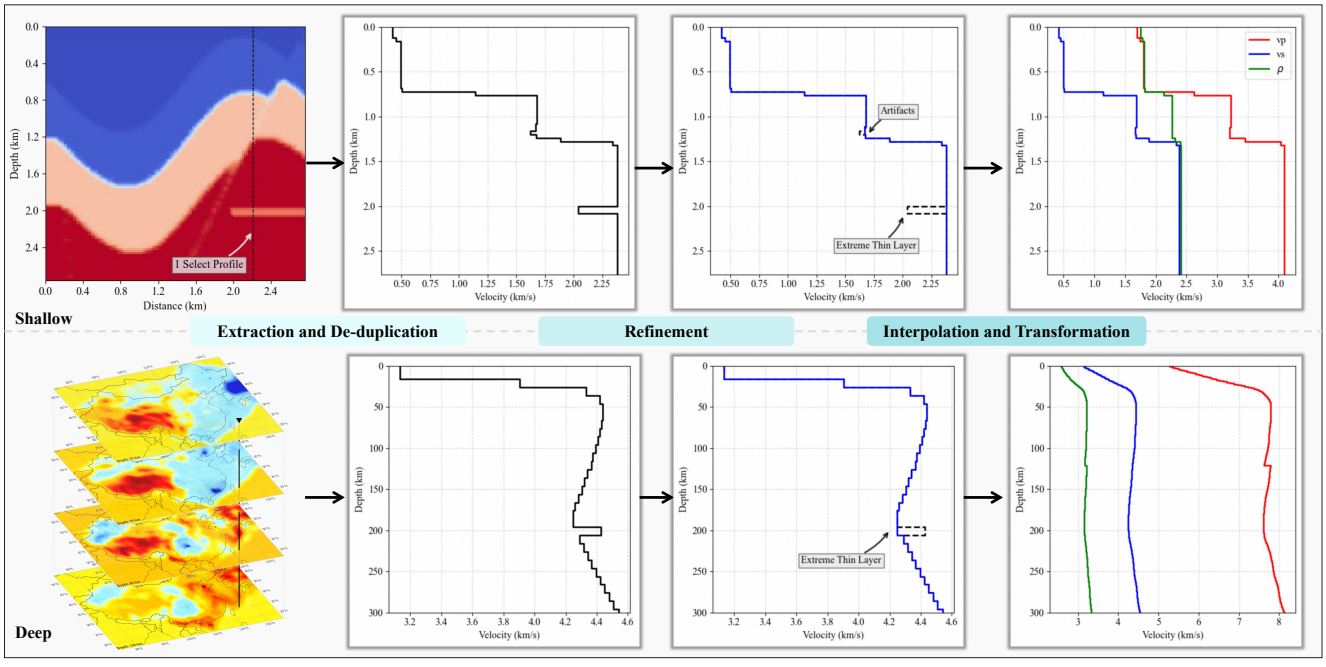


Figure 2. Workflow for extracting and parameterizing 1-D velocity profiles. The upper row shows the process for OpenSWI-shallow, derived from multiple 2-D geological cross-sections, while the lower row illustrates the process for OpenSWI-deep, based on curated 3-D geological models. The workflow includes profile extraction, de-duplication, structure refinement, interpolation, standardization, and parameter conversion to generate depth, v_s (blue), v_p (red), and ρ (green) for forward modeling of surface wave dispersion curves.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

2.1.3 Augmentation of Velocity Models for Geological Diversity

Although the 1-D velocity profiles extracted and transformed from the aforementioned 2-D and 3-D geological models already surpass those used in previous studies in both quantity and diversity, they still cannot fully capture the complete range of geological types and their characteristic variations. To further broaden the dataset's representativeness and establish a scalable data construction workflow, we designed and implemented multiple data augmentation strategies based on the original 2-D geological profiles and the processed 1-D velocity models, as outlined below:

1. **Perturbation-based augmentation of shallow 1-D velocity profiles:** For near-surface geological models, controlled perturbations were applied to both velocities and layer thicknesses while preserving the overall layer structure, thereby enhancing variability across different geological scenarios. The procedure includes: (1) extracting the primary layers from the 1-D profiles; (2) applying random perturbations with constrained amplitudes to the velocity and thickness of each layer; and (3) performing structural plausibility checks on the perturbed profiles, followed by interpolation and parameter conversion as detailed in Section 2.1.2 to ensure physical and numerical validity. The top row of Figure 3 illustrates the augmentation workflow and the resulting variations for a representative 1-D profile.

2. **Feature-aware augmentation of deep 1-D velocity profiles:** For deep geological structures characterized by distinct geophysical interfaces (e.g., the Moho discontinuity), we implemented a feature-aware perturbation strategy to improve model sensitivity to key geological boundaries. The procedure involves: (1) identifying the Moho interface in each 1-D profile; (2) fitting the crustal and mantle layers above and below the interface with cubic spline functions, where the number of spline nodes is randomly selected between 3–6 and 8–12, respectively; and (3) applying random perturbations to the velocity values at the spline nodes, followed by curve smoothing and re-interpolation to generate new deep velocity profiles. The bottom row of Figure 3 illustrates the complete workflow and resulting variations for a representative 1-D profile.

3. **Generative-model-based augmentation of 2-D geological models:** To further enrich geological feature diversity and enable scalable dataset expansion tailored to user needs, we employed deep generative techniques, such as diffusion probabilistic models (DDPMs, Ho et al. (2020)), using the 2-D geological cross-section data collected in Section 2.1.1. These models learn spatial feature distributions and synthesize additional 2-D geological models with improved geological consistency and structural diversity. This component of the workflow is described in greater detail in the subsequent section on shallow-subsurface dataset construction.

These augmentation strategies substantially enriched the dataset in terms of geological types, structural complexity, and the representation of key features. As a result, they provide deep learning models with more diverse and comprehensive training samples, thereby improving generalization and robustness when applied to complex geological settings.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

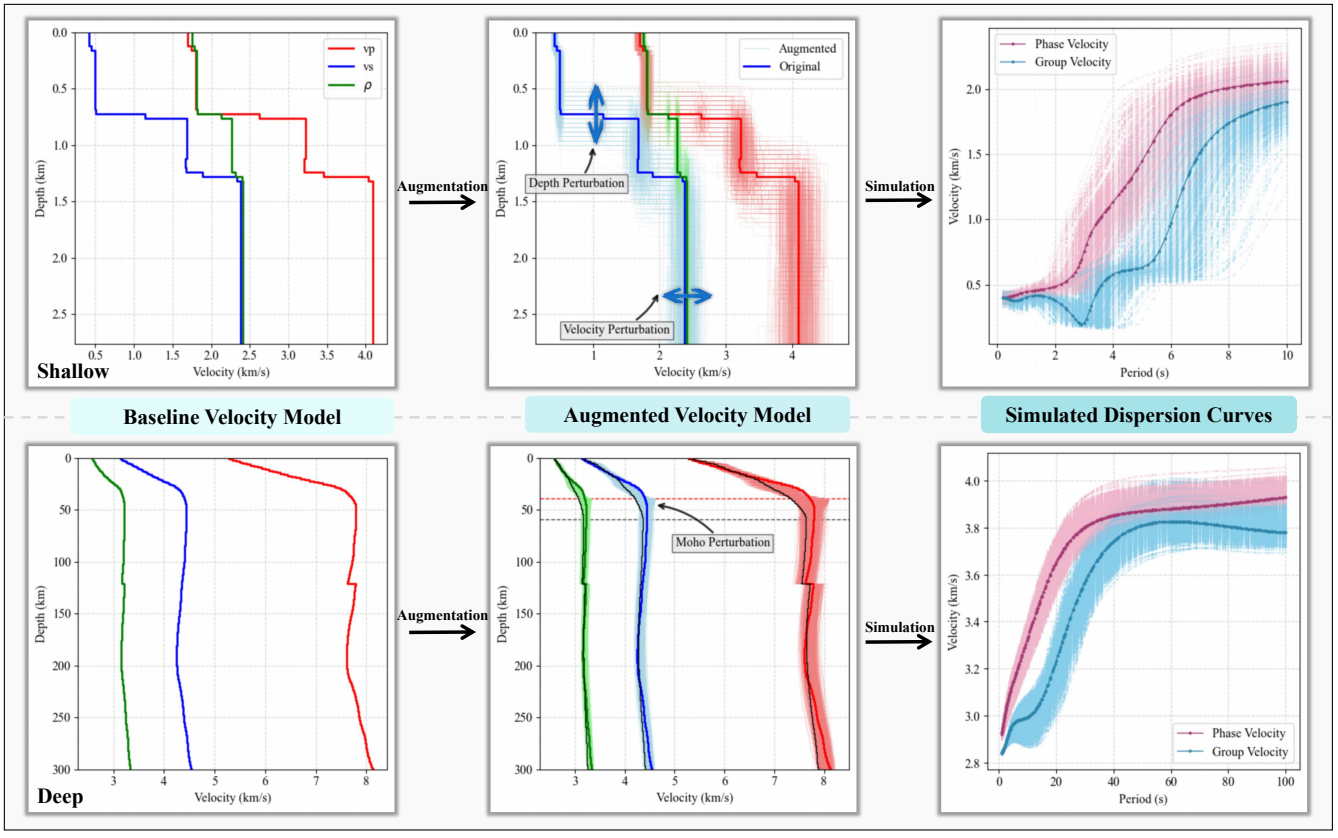


Figure 3. Illustration of data augmentation and forward simulation examples. The top row shows perturbation-based augmentation applied to OpenSWI-shallow data, which increases variability in shallow 1-D velocity profiles. The bottom row shows feature-aware perturbation applied to OpenSWI-deep data, focusing on key structural features such as the Moho discontinuity. Thick lines represent the original 1-D profiles and their corresponding dispersion curves, while thin lines represent the augmented profiles and dispersion curves.

2.1.4 Forward Modeling of Surface-wave Dispersion Curves

Based on the constructed 1-D velocity profiles, we employed efficient geophysical forward modeling tools to generate the corresponding surface-wave dispersion curves. Forward modeling is a critical step in dataset construction, ensuring that the simulated dispersion curves faithfully capture the propagation characteristics of surface waves in different subsurface media. The workflow comprises three main components:

- **Defining the period range of dispersion curves:** In practice, the period range and sampling points of observed dispersion curves vary considerably. To enhance the diversity and applicability of the dataset, we designed a hybrid sampling strategy for constructing the period axis. This strategy integrates uniform, random, and logarithmic sampling, with increased sampling density in the high-frequency range (Wang et al., 2023b; Liu et al., 2025). Such design ensures broad

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

coverage of surface-wave responses across different period bands, improving both the representativeness and utility of the simulated data.

– **Forward computation of dispersion curves:** The forward modeling of surface wave dispersion curves fundamentally involves numerically solving the dispersion equation across a range of frequencies (f), where frequency is defined as the reciprocal of the period T (i.e., $f = 1/T$), to determine the corresponding phase velocity c for each mode (Thomson, 1950; Haskell, 1953; Liu et al., 2024). This process can be formulated as a root-finding problem for the dispersion function D:

$$D(c, f, \mathbf{m}) = 0, \tag{1}$$

where $\mathbf{m}$ denotes the elastic parameters of the layered medium, and D encapsulates the frequency-dependent behavior of wave propagation in this structure. Solving Eq. (1) for each frequency yields the phase velocity dispersion curve $c(f)$ (also denoted as v_{phase}), which characterizes the propagation speed of each harmonic component of the wavefield.

In addition to phase velocity, the group velocity v_{group}, which describes the propagation speed of wave packets, is a critical quantity for surface wave analysis. It is obtained as the derivative of angular frequency $\omega = 2\pi f$ with respect to wavenumber k, and can be expressed in terms of the phase velocity as:

$$v_{\text{group}} = \frac{d\omega}{dk} = c(f) - f\frac{dc(f)}{df}. \tag{2}$$

The group velocity curve complements the phase velocity curve by offering additional sensitivity to subsurface structure and is especially useful in tomographic and inversion applications where energy transport characteristics are of interest.

For each 1-D velocity model, we used the Python library `Disba` (https://keurfonluu.github.io/disba), adapted from the classical seismological software package `Computer Programs in Seismology (CPS)` (Herrmann, 2013), to compute the dispersion curves. This tool efficiently calculates the fundamental-mode phase-velocity and group-velocity characteristics of Rayleigh waves and outputs complete period–velocity pairs (period, phase velocity, and group velocity) for each velocity model, ensuring comprehensive information for inversion tasks.

– **Parallelization and computational acceleration:** Given the large scale of the dataset, we implemented multi-process parallelization and matrix-based batch processing to significantly improve computational efficiency. These optimizations enabled the simulation of hundreds of thousands to millions of dispersion curves within a practical timeframe, meeting the data requirements of deep learning applications.

This workflow produced a large-scale, quality-controlled dataset of surface-wave dispersion curves. Figure 3 showcases examples of dispersion curves from the OpenSWI-shallow and OpenSWI-deep datasets. These simulated data provide a solid foundation for training deep learning–based inversion models, facilitating applications in resource exploration and imaging of Earth's internal structure.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Table 2. Comprehensive summary of the OpenSWI dataset, describing its categories (OpenSWI-shallow, OpenSWI-deep, and OpenSWI-real), associated period ranges (seconds, s), depth ranges (kilometers, km), and sampling intervals (kilometers, km), as well as the extracted and augmented 1-D velocity profiles ($depth$, v_p, v_s, ρ), expressed as N profiles $\times$ M model variables $\times$ L layers.

Group	Datasets	Period Range (s)	Depth Range (km)/ Depth Interval (km)	Extracted 1-D Velocity Profiles	Augmented 1-D Velocity Profiles
OpenSWI shallow	Flat	0.1-10	0-2.8 / 0.04	29,379 × 4 × 70	1,490,415 × 4 × 70
	Flat+Fault	0.1-10	0-2.8 / 0.04	292,933 × 4 × 70	2,925,151 × 4 × 70
	Fold	0.1-10	0-2.8 / 0.04	295,751 × 4 × 70	2,952,975 × 4 × 70
	Fold+Fault	0.1-10	0-2.8 / 0.04	537,751 × 4 × 70	5,369,692 × 4 × 70
	Field	0.1-10	0-2.8 / 0.04	2,338,248 × 4 × 70	9,345,103 × 4 × 70
	All	0.1-10	0-2.8 / 0.04	3,494,062 × 4 × 70	22,083,336x 4 × 70
OpenSWI deep	LITHO1.0	1-100	0-300 / 1.0	40,959 × 4 × 300	24,5771 × 4 × 70
	USTClitho1.0	1-100	0-300 / 1.0	9,125 × 4 × 300	54,750 × 4 × 70
	Central-and-Western US	1-100	0-300 / 1.0	6,803 × 4 × 300	40,818 × 4 × 70
	Continental China	1-100	0-300 / 1.0	4,516 × 4 × 300	27,096 × 4 × 70
	US Upper-Mantle	1-100	0-300 / 1.0	3,678 × 4 × 300	22,061 × 4 × 70
	EUcrust	1-100	0-300 / 1.0	43,520 × 4 × 300	261,155 × 4 × 70
	Alaska	1-100	0-300 / 1.0	19,408 × 4 × 300	116,448 × 4 × 70
	CSEM-Europe	1-100	0-300 / 1.0	21,931 × 4 × 300	131,586 × 4 × 70
	CSEM-Eastmed	1-100	0-300 / 1.0	12,782 × 4 × 300	76,692 × 4 × 70
	CSEM-Iberian	1-100	0-300 / 1.0	9,102 × 4 × 300	54,612 × 4 × 70
	CSEM-South Atlantic	1-100	0-300 / 1.0	7,371 × 4 × 300	44,226 × 4 × 70
	CSEM-North Atlantic	1-100	0-300 / 1.0	14,541 × 4 × 300	87,246 × 4 × 70
	CSEM-Japan	1-100	0-300 / 1.0	14,641 × 4 × 300	87,846 × 4 × 70
	CSEM-Astralasia	1-100	0-300 / 1.0	4,131 × 4 × 300	24,786 × 4 × 70
	All	1-100	0-300 / 1.0	212,508 × 4 × 300	1,275,093 × 4 × 70
OpenSWI real	LongBeach	0.263 - 1.666	0-1.4 / 0.04	5,297 × 4 × 35	-
	CSRM	8 - 70	0-120 / 1.0	12,901 × 4 × 120	-

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

2.1.5 Open-source Implementation

To promote reproducibility, scalability, and community engagement, we developed a standardized Python toolkit named `SWIDP` (Surface Wave Inversion Dataset Preparation pipeline). Built upon the key procedures described in Sections 2.1.1–2.1.4, `SWIDP` encapsulates core functionalities such as the extraction and parameterization of 1-D velocity profiles, data augmentation, and large-scale dispersion curve simulation.

By automating these processes, it enhances the efficiency, transparency, and consistency of dataset preparation. Designed with a modular architecture, `SWIDP` allows users to flexibly reuse or extend specific components, facilitating seamless adaptation to diverse research needs. Example codes are provided in Appendix A and B. The full source code is openly available at https://doi.org/10.5281/zenodo.16884901 (Liu, 2025b) and https://github.com/liufeng2317/OpenSWI, enabling both academic and industrial users to adopt and further develop the toolkit.

2.2 OpenSWI-shallow: Large-scale Benchmark for Complex Shallow Geology

2.2.1 Building Geological Model Foundations from OpenFWI

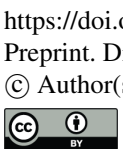
To establish a representative benchmark dataset for shallow-subsurface surface-wave dispersion curve inversion, we constructed a comprehensive collection of 2-D velocity models with diverse geological structures derived from the OpenFWI dataset (Deng et al., 2021). These models encompass five primary geological categories: flat layers (Flat), flat layers with faults (Flat–Fault), folded layers (Fold), folded layers with faults (Fold–Fault), and field-style models (Field) inspired by realistic observations. Each category contains approximately 30,000, 54,000, 30,000, 54,000, and 67,000 samples, respectively. All models share a grid resolution of 70×70 with a spatial sampling interval of 40 m, ensuring sufficient detail to capture the complexity and variability of shallow-subsurface geological features.

Based on these 2-D models, we systematically extracted a large number of 1-D velocity profiles according to the geological characteristics of each geological categories. To enhance the dataset's diversity and coverage, each original 1-D profile was augmented 4 to 10 times by independently applying perturbations of up to 10% in layer thickness and 5% in velocity. Following these perturbations, plausibility checks and interpolation adjustments were performed to ensure physical consistency and numerical stability. The final dataset comprises over 22 million 1-D velocity models spanning all geological categories. Detailed statistics of both the extracted and augmented profile counts for each category are summarized in Table 2.

Forward modeling of fundamental-mode Rayleigh-wave dispersion curves was then conducted for all 1-D models. Given that the maximum depth of these profiles is approximately 2.8 km, the simulated period range was defined from 0.2 s to 10 s, with 100 period points sampled per curve. To improve period coverage and model generalization capability, the sampling points were selected using a hybrid strategy combining uniform, random, and logarithmic sampling, contributing 50, 30, and 20 points, respectively. Each dispersion curve includes period, phase-velocity, and group-velocity information, serving as training and validation data for subsequent deep learning applications.

Figures 4 and 5 showcase the representativeness and statistical properties of the OpenSWI-shallow dataset. Figure 4 illustrates the diverse geological scenarios covered by the dataset through representative 2-D velocity models, systematically

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025



extracted 1-D profiles, and their augmented variants, together with the corresponding phase and group velocity dispersion curves. Figure 5 further summarizes the large-scale statistical distributions of profiles and dispersion characteristics across all geological types, highlighting the dataset's substantial improvements in structural diversity, distributional coverage, and suitability for data-driven surface wave inversion studies.

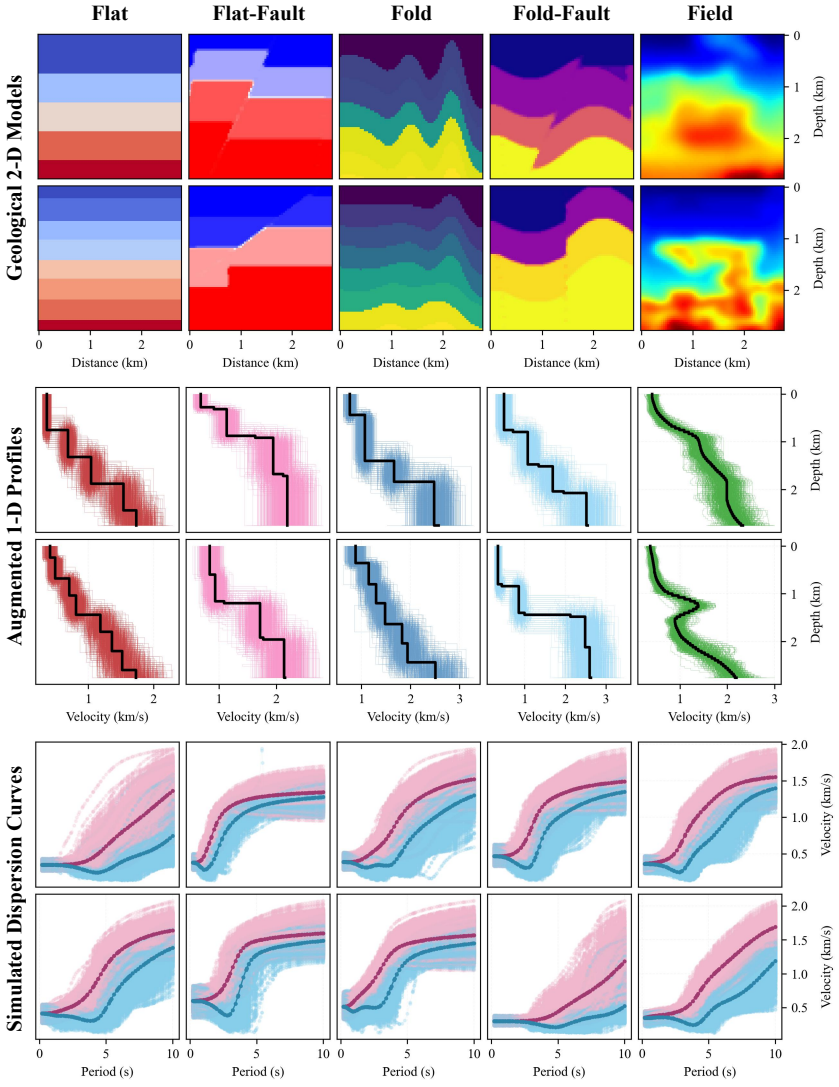


Figure 4. Representative samples from the OpenSWI-shallow dataset. The top two rows present original 2-D velocity models for five geological types: Flat, Flat–Fault, Fold, Fold–Fault, and Field. The middle two rows show the corresponding extracted 1-D velocity profiles (bold black lines) and their augmented variants (thin colored lines). The bottom two rows display the simulated Rayleigh-wave dispersion curves, with phase velocities shown in pink and group velocities in blue.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

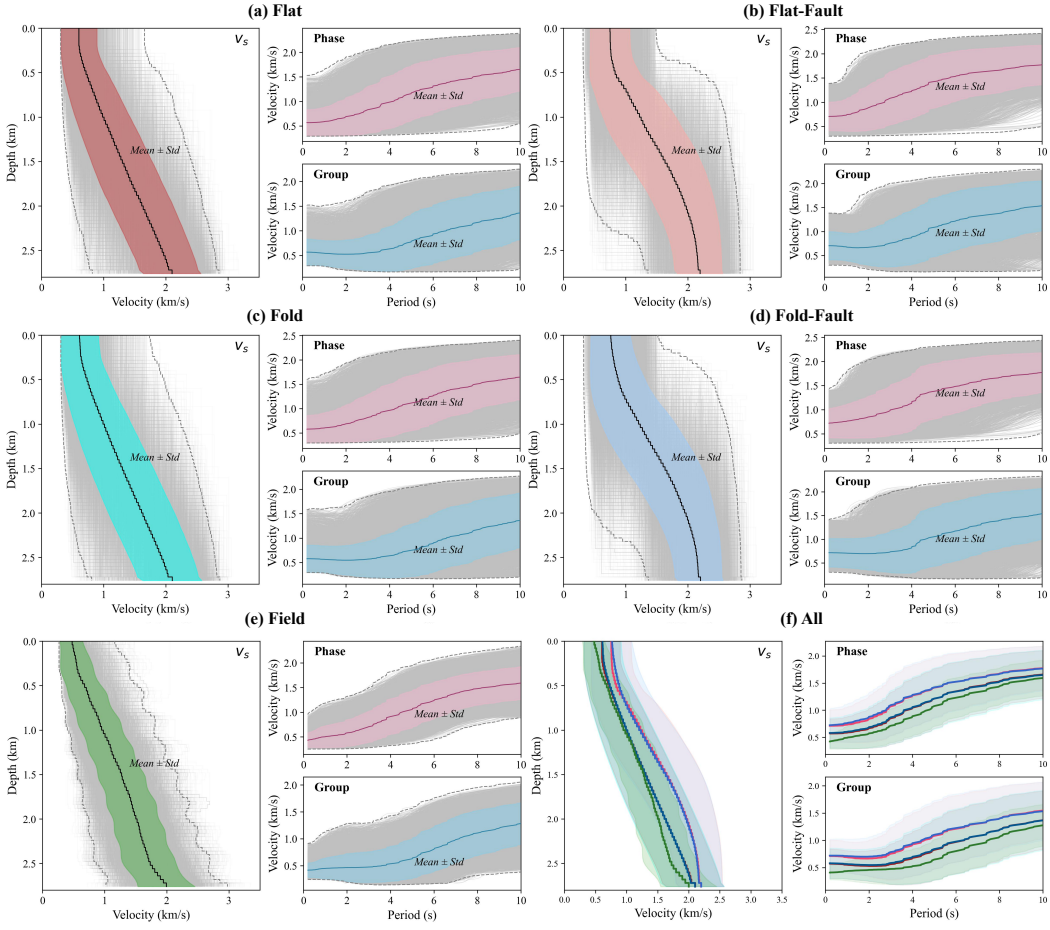


Figure 5. Statistical characteristics of the OpenSWI-shallow dataset. Distribution of 1-D velocity profiles and corresponding dispersion curves for each geological style: (a) Flat, (b) Flat-Fault, (c) Fold, (d) Fold-Fault, (e) Field. The black lines represent the mean, and the shaded regions indicate the ± 1 standard deviation range. Panel (f) summarizes the mean and variance across the five geological subsets.

2.2.2 Continual Expansion of Geological Structures with DDPM

Although the proposed OpenSWI-shallow dataset constructed from OpenFWI substantially improves geological structural diversity compared with existing dispersion curve datasets, it still cannot fully capture the complete range of velocity structure types observed in real subsurface settings. To further enhance the dataset's scalability in terms of structural complexity and geological representativeness, we incorporated a deep generative module based on Diffusion Probabilistic Model (DDPM), specifically designed for the shallow subsurface within the 0–3 km depth range.

This module uses 2-D velocity models from OpenFWI as training data to develop multiple DDPMs, which learn the distributional characteristics of different geological structures. Starting from Gaussian noise, the DDPMs iteratively generate velocity

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

models with realistic structural features, consistently reproducing faults, folds, and complex sedimentary units. Compared with traditional manual or perturbed augmentation, the DDPM-generated data provide clear advantages in structural continuity, geological realism, and controllable scalability, significantly expanding the foundational velocity model library (Ho et al., 2020; Wang et al., 2023a; Taufik et al., 2024). Details of the DDPM design and training are provided in Appendix C, and the code has been publicly released with the SWIDP pipeline for reproducibility.

Figure 6 illustrates the continual expansion of the OpenSWI-shallow dataset using the DDPM module. The diffusion model progressively transforms Gaussian noise into geologically realistic 2-D velocity models through a 1000-step denoising process, from which representative 1-D profiles are extracted and used to simulate Rayleigh-wave dispersion curves. This diffusion-based augmentation strategy substantially enriches the structural diversity and spatial coverage of the dataset, thereby improving the generalization capability of deep learning models.

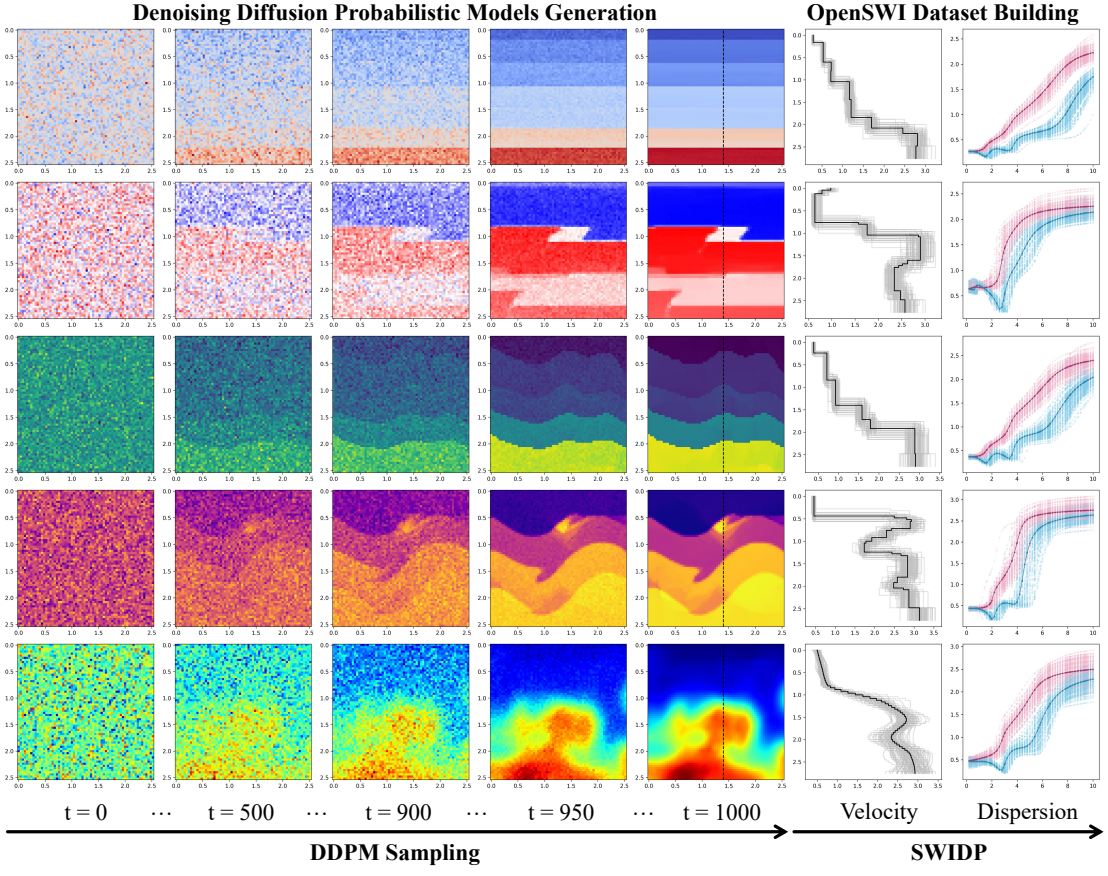


Figure 6. Continual expansion of the OpenSWI-shallow dataset using a diffusion-based generative module. The left panel illustrates a 1000-step denoising trajectory, where Gaussian noise is progressively transformed into 2-D velocity models with realistic geological structures. The right panel presents representative 1-D velocity profiles extracted from the generated models, along with their corresponding Rayleigh-wave dispersion curves simulated using the SWIDP pipeline.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

2.3 OpenSWI-deep: Global Coverage Benchmark for Deep Earth Imaging

Building upon the shallow-subsurface benchmark dataset introduced in Section 2.2, we further extended the OpenSWI framework to deeper Earth structures. However, for the deeper Earth structure, systematic datasets of regular velocity models remain largely unavailable. To address this gap, we compiled a collection of representative 3-D velocity models from published literature and geophysical studies. This collection includes one global-scale model and 13 high-resolution regional models, each constructed using different methodologies and data sources to maximize geological representativeness and geophysical applicability. Figure 7 shows the spatial distribution of these 14 models with horizontal slices at a depth of 60 km.

Among them, LITHO1.0 provides global information on the crust and upper mantle, encompassing sedimentary layers, crust, lithosphere, and asthenosphere, at a spatial resolution of 1°(Pasyanos et al., 2014). This model is widely used in seismic tomography and as a reference Earth model. USTClitho1.0, derived from double-difference tomography using seismic data from the Chinese National Seismic Network, resolves crustal and upper mantle structures down to 150 km depth at a horizontal resolution of 0.5°, supporting studies of regional deep structures (Xin et al., 2019). The Central and Western US (Shen et al., 2013) and Continental China (Shen et al., 2016) models integrate ambient noise and teleseismic surface waves with receiver function data and apply a Bayesian Monte Carlo inversion to image crust and upper mantle structures to 150 km depth at 0.5°resolution. The US Upper Mantle model uses long-period Rayleigh wave ambient noise and Markov chain Monte Carlo inversion to map shear-wave velocities down to 300 km across the continental United States (Xie et al., 2018). Similarly, the EUCrust model, based on four years of ambient noise data from 1,293 broadband stations, resolves the European crust and uppermost mantle with high resolution using Bayesian nonlinear methods (Lu et al., 2018). The Alaska model combines data from over 200 Transportable Array stations and integrates Rayleigh wave ellipticity, phase velocity, and receiver functions, using Markov chain inversion to image structures from the upper mantle to near-surface depths (140 km) (Berg et al., 2020).

We also included several regional models from The Collaborative Seismic Earth Model Project (CSEM), constructed through full-waveform inversion (Fichtner et al., 2006, 2013, 2018). These cover Europe (Çubuk-Sabuncu et al., 2017; Blom et al., 2020), the Eastern and Western Mediterranean (Blom et al., 2020; Fichtner and Villaseñor, 2015), the South and North Atlantic (Colli et al., 2013; Rickers et al., 2013; Krischer et al., 2018), the Japanese Islands (Simutė et al., 2016), and Australasia (Fichtner et al., 2009, 2010). These models are characterized by high resolution and structural consistency and are widely used for deep Earth imaging and geodynamic research.

All collected 3-D velocity models underwent quality control, including duplicate removal, anomaly detection, gap interpolation, and format homogenization to ensure consistency for dataset construction. From the processed models, we extracted approximately 212,508 1-D velocity profiles, with detailed statistics for each data source provided in Table 2. To further increase diversity, each profile was augmented five times using a hierarchical strategy: the depth of the Moho discontinuity was first identified, and profiles were divided into crust and upper mantle sections. Each section was parameterized using cubic spline curves, with 3–6 control nodes for the crust and 6–12 nodes for the upper mantle, followed by random perturbations of the nodes to introduce structural variations. This augmentation preserved key geological features (e.g., the Moho interface)

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

while significantly expanding the coverage and variability of the model library, ultimately yielding approximately 1.26 million augmented 1-D velocity models.

For each 1-D profile, we simulated fundamental-mode Rayleigh-wave dispersion curves over a 1–100 s period range, sampling 300 points using a combination of uniform, random, and logarithmic strategies (50, 30, and 20 points, respectively). Each dispersion curve, together with its associated velocity profile, constitutes a complete input–output pair for subsequent deep learning model training and validation. Figure 8 illustrates representative 1-D velocity profiles and their corresponding dispersion curves from the regional models, highlighting the relationships between velocity structures and surface-wave propagation under diverse geological conditions. These results provide high-quality initial data support for global geophysical imaging across different regions and scales.

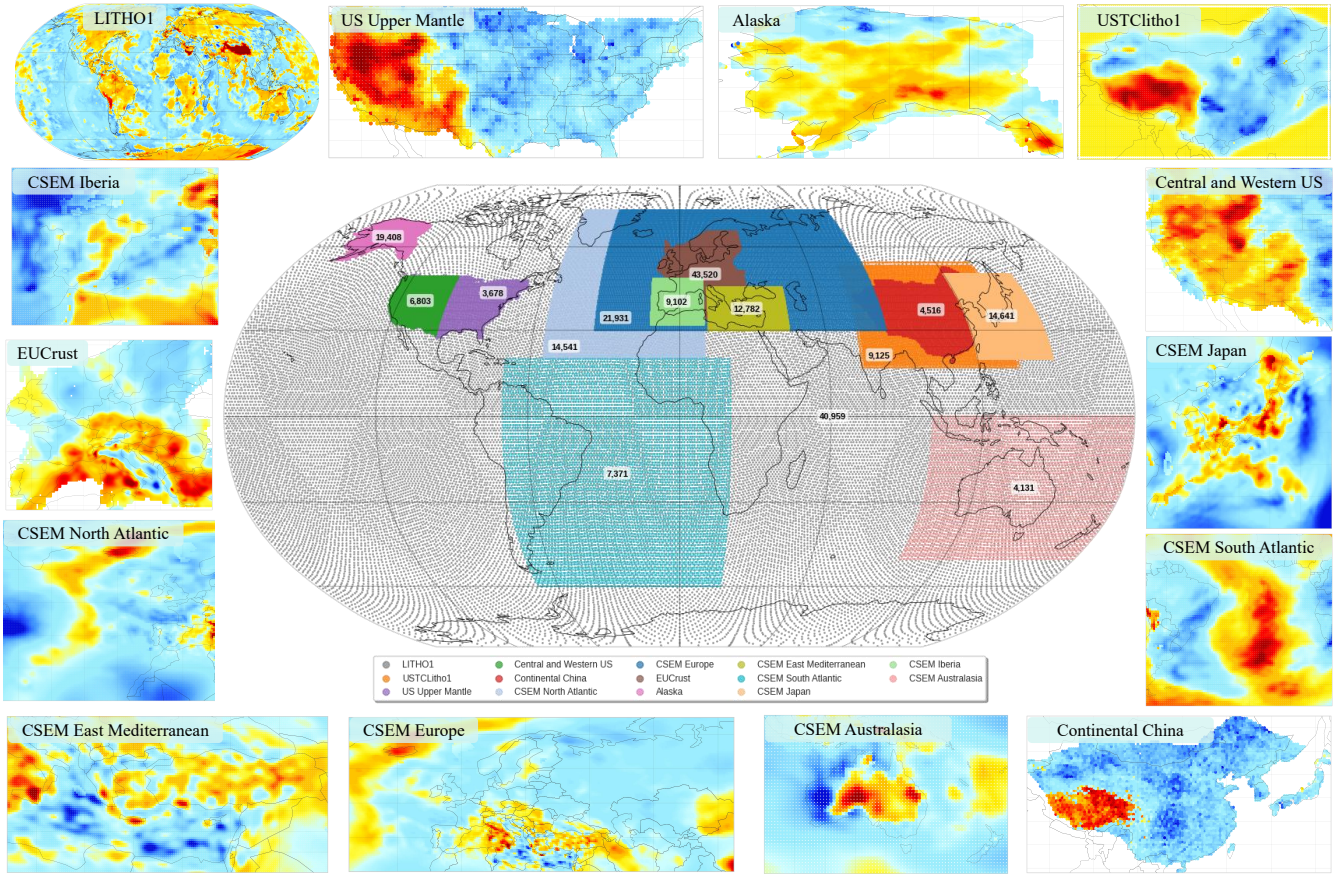


Figure 7. Spatial distribution of the 14 velocity models compiled for the OpenSWI-deep dataset. The collection includes one global-scale model and 13 high-resolution regional models obtained from published literature and geophysical studies. Horizontal slices at a depth of 60 km are shown to illustrate their geographic coverage and tectonic diversity.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

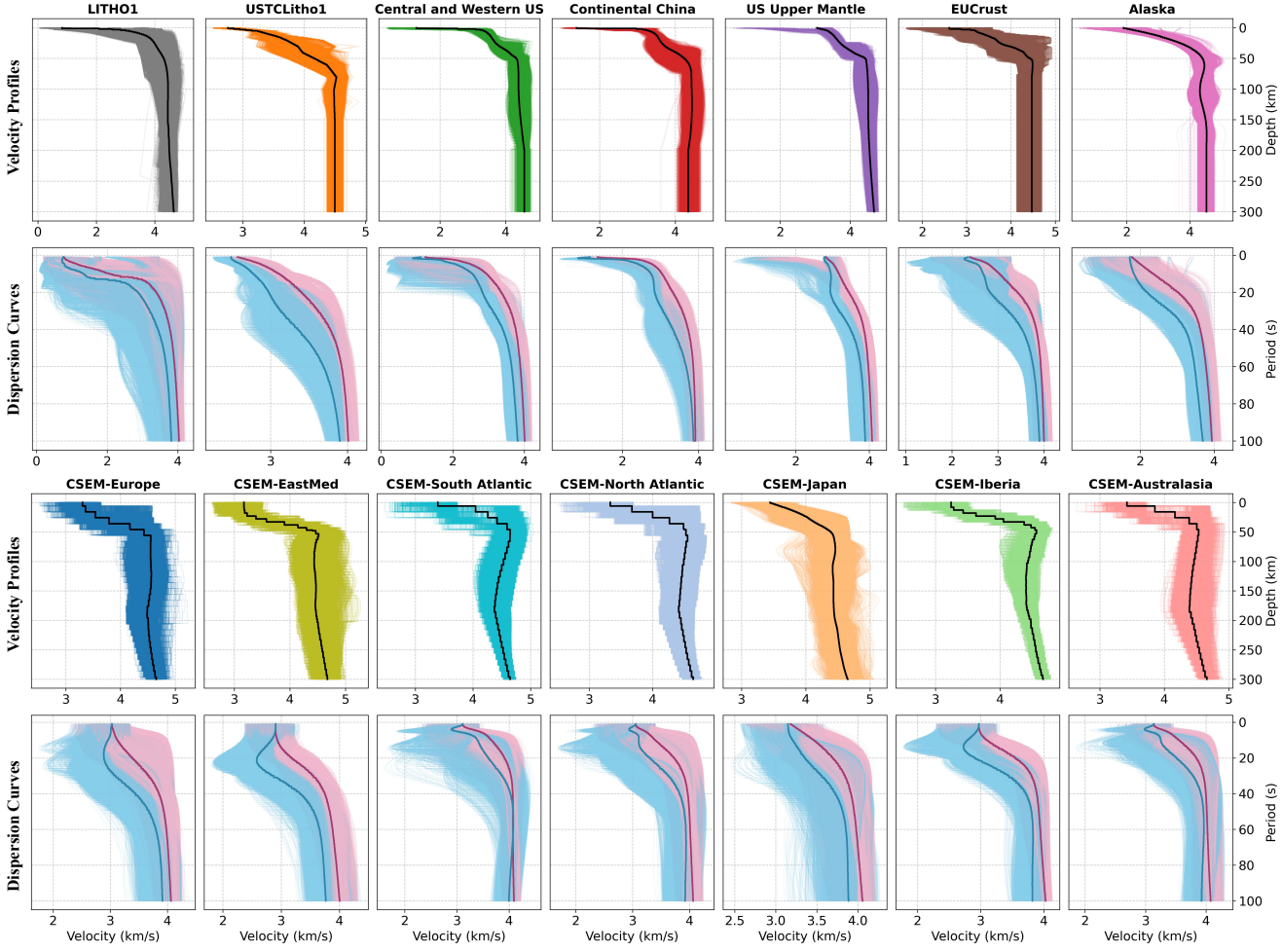


Figure 8. Representative samples from the OpenSWI-deep dataset. The first and third rows show 1-D velocity profiles extracted from the 14 sub-datasets, where the mean velocity model is indicated by a solid black line. The second and fourth rows display the corresponding fundamental-mode Rayleigh-wave dispersion curves over a 1–100 s period range, with the mean phase and group velocities shown in pink and blue, respectively.

2.4 OpenSWI-real: AI-ready Real-world Dataset for Generalization Testing

In addition to the large-scale synthetic velocity profile–dispersion curve datasets designed for model training, we curated multiple AI-ready real-world dispersion curve datasets to assess the adaptability and generalization capability of deep learning methods under practical geophysical conditions.

The first dataset is derived from the dispersion curve data processed by Fu et al. (2022) in the Long Beach region of the United States. As shown in Figure 9(a), over 5,200 short-period nodal stations were deployed between January and June 2011,

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

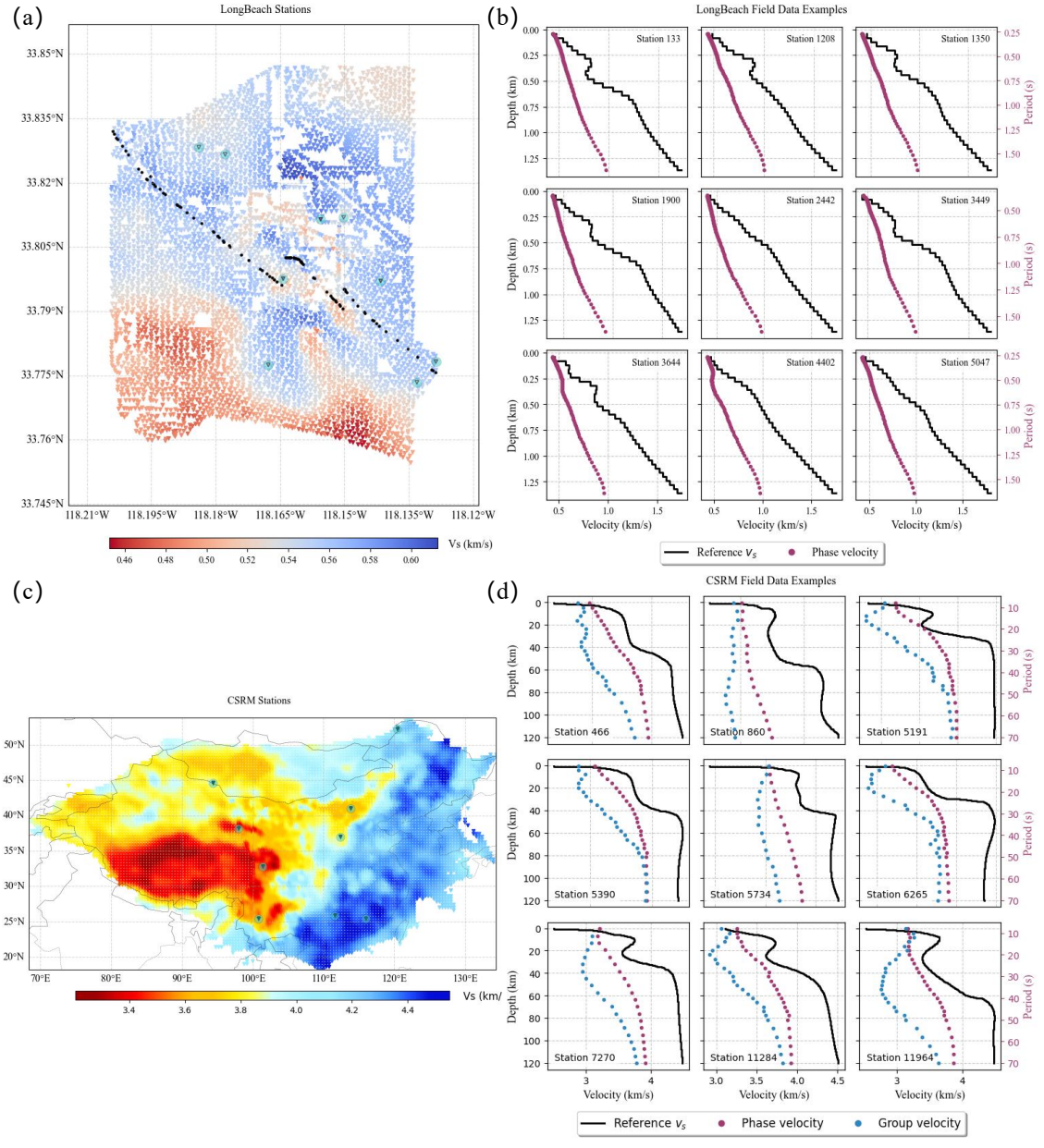


Figure 9. Overview of the OpenSWI-real dataset. (a) Station deployment for the Long Beach dataset in Southern California. (b) Representative observed phase velocity dispersion curves (purple dashed lines) and reference velocity models (black lines) from traditional inversion. (c) Distribution of selected grid points in the CSRM dataset across continental China, with background color denoting velocity at 70 km depth. (d) Representative examples from the CSRM dataset showing observed group (blue) and phase (purple) velocity curves and corresponding reference 1-D velocity profiles (black).

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

primarily for oilfield surveys (Lin et al., 2013), with an average station spacing of approximately 0.1 km. To achieve adequate spatial resolution, the dense array was divided into multiple subarrays, each with a 2 km radius. Dispersion curves were extracted automatically using a deep neural network after the frequency–Bessel (F–J) transform was applied to compute the frequency–phase velocity spectrum for each subarray. Figure 9(b) shows representative observed dispersion curves from 9 stations (purple dashed lines), together with 1-D reference shear-wave velocity profiles (black solid lines) obtained via traditional inversion methods. This dataset contains only phase velocity data, without group velocity information. After standardized processing, it comprises observed dispersion data from 5,297 stations (period range: 0.263–1.666 s) and corresponding reference velocity models (depth range: 0–1.4km, interpolated at 40 m intervals).

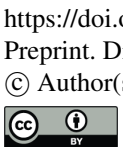
The second dataset originates from the China Seismological Reference Model Project (Wen et al., 2023; Xiao et al., 2024). Xiao et al. (2024) collected continuous seismic records from multiple networks, including the China National Seismic Network (CNSN), the China Seismic Array (ChinArray), and the Public Data Management Center (PDMC), spanning 4,196 seismic stations in total. Ambient noise cross-correlations between station pairs produced 639,171 empirical Green's functions, from which dispersion curves were extracted using frequency–time analysis. Additionally, 54,792 event–station dispersion curves were retrieved from 226 regional seismic events recorded by 1,463 stations. After gridding and quality control, the data were consolidated into 20,514 grid points and standardized to a period range of 8–70 s. To ensure reliability, we retained 12,901 grid points with at least 20 sampled period points. The resulting AI-ready dataset contains observed dispersion curves at these grid points (period range: 8–70 s) and their corresponding reference velocity models (depth range: 0–120 km, interpolated at 1 km intervals). Figure 9(c) shows the spatial distribution of the selected grid points across continental China, with background colors indicating the reference model velocity at 70 km depth. Figure 9(d) presents nine representative grid points, displaying observed dispersion curves (blue: group velocity; purple: phase velocity) and corresponding reference velocity profiles (black solid lines) derived from the traditional inversion results of Xiao et al. (2024).

3 Deep-learning-based Framework for Surface-wave Inversion

3.1 Transformer-based Architecture for Dispersion Curve Inversion

Deep learning-based surface-wave dispersion curve inversion seeks to learn a nonlinear mapping from input dispersion curves (including period, phase velocity, and group velocity) to corresponding 1-D subsurface shear-wave velocity profiles. In this study, we adopt a widely used Transformer-based architecture (Figure 10a) to enable end-to-end inversion (Liu et al., 2025; Jiang et al., 2025). The input to the model is a $3 \times N$ dispersion curve matrix, where the 3 rows represent period, phase velocity, and group velocity, and N denotes the number of sampled points. The model initially embeds the input via three separate 1×1 convolutional neural network (CNN) layers, yielding a feature representation of size $3 \times N \times E$, where E is the feature dimension. These embedded features are then processed through multiple Transformer blocks, which employ self-attention mechanisms to capture long-range dependencies across the dispersion curves. This global context modeling enhances the stability and accuracy of inversion results. Finally, a feature projection layer maps the global features extracted by the

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025



Open Access Earth System Science Data Discussions

Transformer to a velocity profile of length M, where M corresponds to the number of target depth layers, producing the final inversion output.

Given that the period range and target depth in real observational data vary, and that the maximum inversion depth strongly correlates with the observed period range, we incorporate the depth-aware strategy proposed by Liu et al. (2025) during training. This approach dynamically computes the maximum wavelength (period multiplied by velocity) for each input and adaptively determines the effective output depth range, thereby suppressing predictions at irrelevant depths and improving inversion accuracy. For the loss function, we adopt the Mean Squared Error (MSE), calculated exclusively over the effective depth range between predicted and ground-truth velocity profiles. To enhance robustness against noise and missing data commonly encountered in practice, we simulate these effects during data loading by adding 3% random Gaussian noise and randomly masking 10% of the dispersion data points.

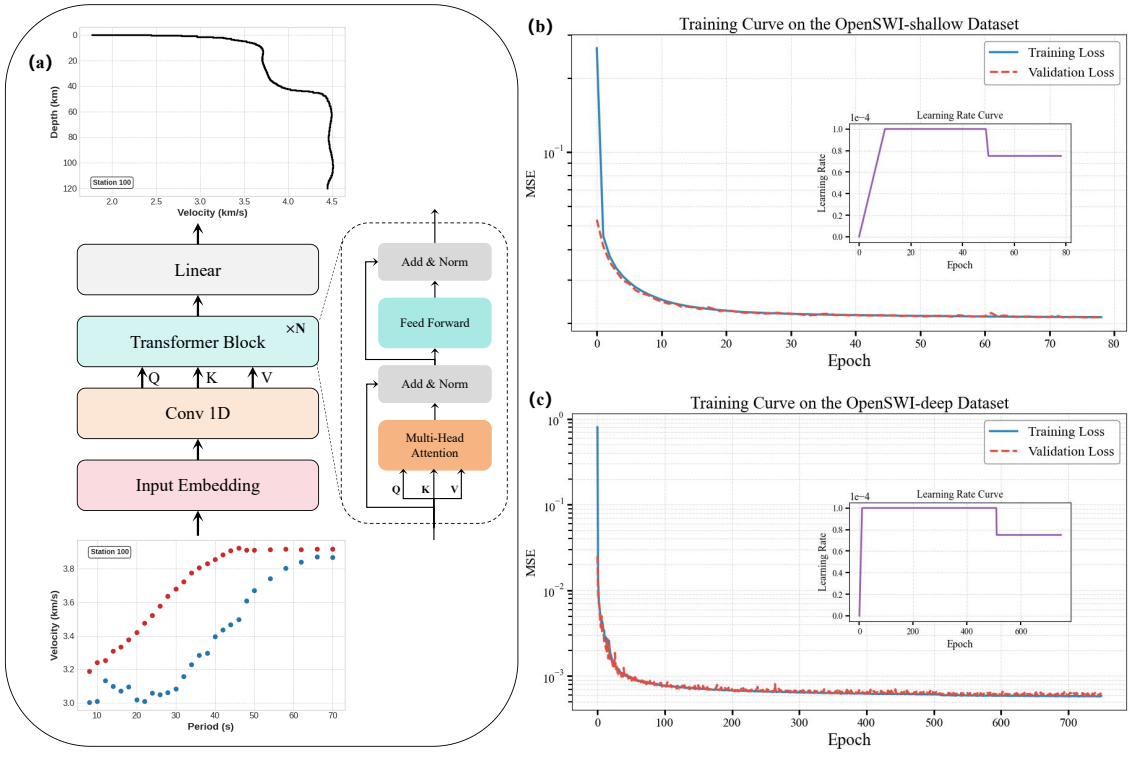


Figure 10. (a) The architecture of the deep neural network (Transformer) used in this work for surface wave dispersion curve inversion. The Training (blue) and validation (red) loss curve on the (b) OpenSWI-shallow, and (c) OpenSWI-deep datasets. The learning rate curve are presents in the inner figure with purple line.

Regarding training configuration, we use a larger batch size of 2048 and limit training to 100 epochs for the shallow dispersion dataset (OpenSWI-shallow) to optimize large-scale training efficiency. For the deeper dataset (OpenSWI-deep), a smaller batch size of 512 and up to 1000 epochs are employed. To avoid overfitting and reduce unnecessary computation, we

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

adopt an early stopping strategy, terminating training when the validation loss does not improve for 30 consecutive epochs for OpenSWI-shallow and 50 epochs for OpenSWI-deep. Both datasets are trained using the Adam optimizer, combined with a learning rate scheduler that integrates warm-up and step decay strategies to enhance training stability. During warm-up, the learning rate increases linearly from 1×10^{-9} to 1×10^{-4} over approximately 2 epochs for OpenSWI-shallow and 10 epochs for OpenSWI-deep. In the subsequent step decay phase, the learning rate is reduced to 75% of its value every 20 epochs for OpenSWI-shallow and every 200 epochs for OpenSWI-deep. Figures 10b and 10c present the training and validation error curves for both datasets alongside their corresponding learning rate schedules.

Model performance is first evaluated on the test sets by comparing predicted and ground-truth velocity profiles using Root Mean Squared Error (RMSE). Beyond this quantitative validation, the trained models are applied to real observational data to assess their generalization capabilities. Instead, we compare the synthetic and observed dispersion curves to compute the misfit errors, and assess the inversion quality by analyzing the distribution of these errors, including their mean and variance.

3.2 Large-scale Training with the OpenSWI-shallow and OpenSWI-deep Datasets

To comprehensively assess the effectiveness of the proposed deep neural network model for surface wave dispersion curve inversion, we conducted systematic training on both the OpenSWI-shallow and OpenSWI-deep datasets. Detailed architectural hyperparameters are provided in Appendix D. To ensure balanced representation across the training, validation, and test subsets, we employed stratified sampling strategies. Specifically, for the OpenSWI-shallow dataset, stratification was based on geological structure types (Flat, Flat-Fault, Fold, Fold-Fault, and Field), using a 90%/5%/5% split. For the OpenSWI-deep dataset, stratification was performed by geographic regions of the source models, following the same partitioning ratio.

During training, both training and validation errors were continuously monitored, as illustrated in Figure 10b and c. For both datasets, the error curves demonstrate stable convergence, suggesting that the model effectively captures the nonlinear relationship between surface wave dispersion curves and subsurface shear-wave velocity profiles. After training, evaluation on the held-out test sets yielded RMSE values of 0.1467 km/s for OpenSWI-shallow and 0.048 km/s for OpenSWI-deep, indicating that the predicted velocity models closely match the ground-truth profiles and confirming the model's high inversion accuracy under varying geological conditions.

Representative inversion results from both datasets are shown in Figures 11a and 11b, further demonstrating the model's capability to reconstruct subsurface velocity structures with high fidelity, including at greater depths. These results validate the generalization ability and practical applicability of the proposed method across diverse geological settings. It is worth noting that OpenSWI-shallow includes significantly more samples than OpenSWI-deep and encompasses a wider variety of geologically diverse and structurally complex velocity models. Consequently, achieving optimal performance on this dataset demands more specialized architectural designs and training strategies. While the model maintains strong overall inversion quality, it tends to oversmooth regions characterized by strong heterogeneity or abrupt structural changes, resulting in slightly muted responses in complex geological zones. This smoothing effect highlights current limitations in resolving fine-scale structural features and underscores the need for future enhancements, such as structure-aware regularization or multi-scale modeling techniques, to improve the representation of intricate subsurface variations.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

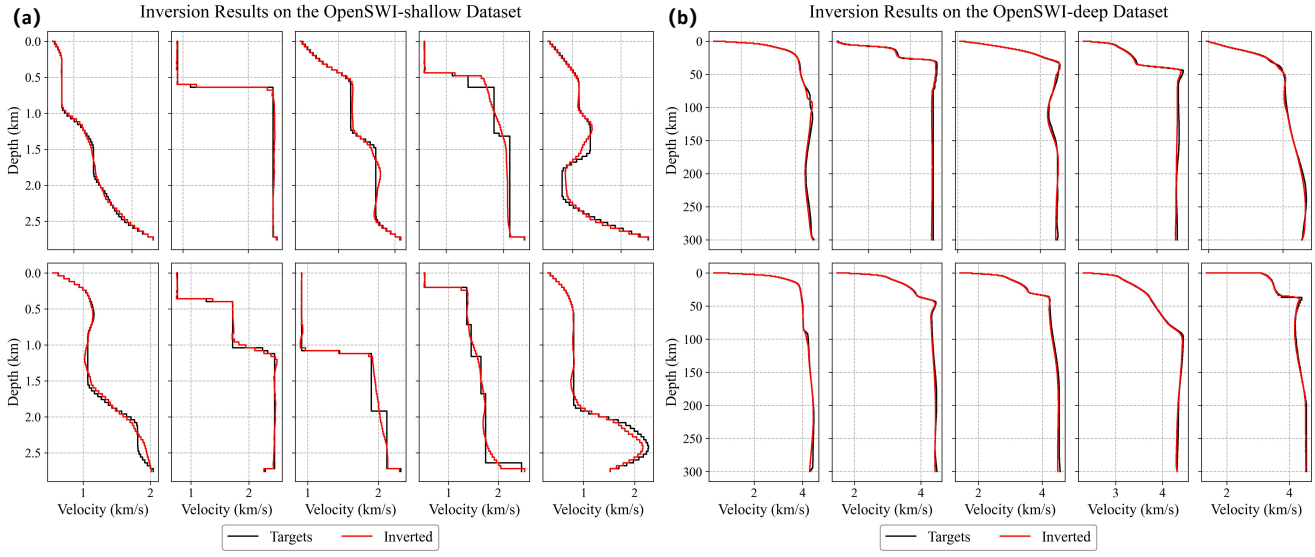


Figure 11. Representative inversion results on the test subsets of (a) OpenSWI-shallow and (b) OpenSWI-deep. The black lines represent the ground-truth velocity profiles, while the red lines denote the predicted results obtained by the trained neural network.

3.3 Generalization Testing on Real-world Observations Using OpenSWI-real

To evaluate the generalization capability of deep neural networks trained entirely on synthetic data, we directly applied the pretrained models to the OpenSWI-real dataset, which includes two representative real-world regions: Long Beach (shallow) and CSRM (deep).

In the shallow case, we used phase velocity dispersion curves from 5,297 stations in the Long Beach area as input to the shallow inversion network. The model generated a 1-D S-wave velocity profile for each station, which were then assembled into a 3-D velocity model of the region. Figure 12a presents horizontal slices of the predicted model at depths of 100 m, 200 m, 400 m, and 600 m, alongside the corresponding reference model. Figure 12b compares selected 1-D profiles from both models. Notably, despite the complete absence of Long Beach data during training, the model successfully reconstructs key subsurface velocity structures. In particular, the predicted profiles at 100 m and 200 m show excellent agreement with the reference model. Figure 12c shows the observed dispersion curves (black), as well as synthetic curves generated from the reference model (blue) and the neural network predictions (red). To quantitatively evaluate inversion performance, we computed the misfit between observed dispersion curves and those derived from the predicted velocity profiles. Figure 12d summarizes the error distributions. For the reference model, the mean and variance of the misfit are –33.9m/s and 14.7$(m/s)^2$, respectively, whereas the neural network predictions yield a mean misfit of 1.8m/s and a variance of 18.1$(m/s)^2$. These results demonstrate that the pretrained model generalizes effectively to real observational data and, in many cases, even outperforms the reference model, particularly in shallow geological settings.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

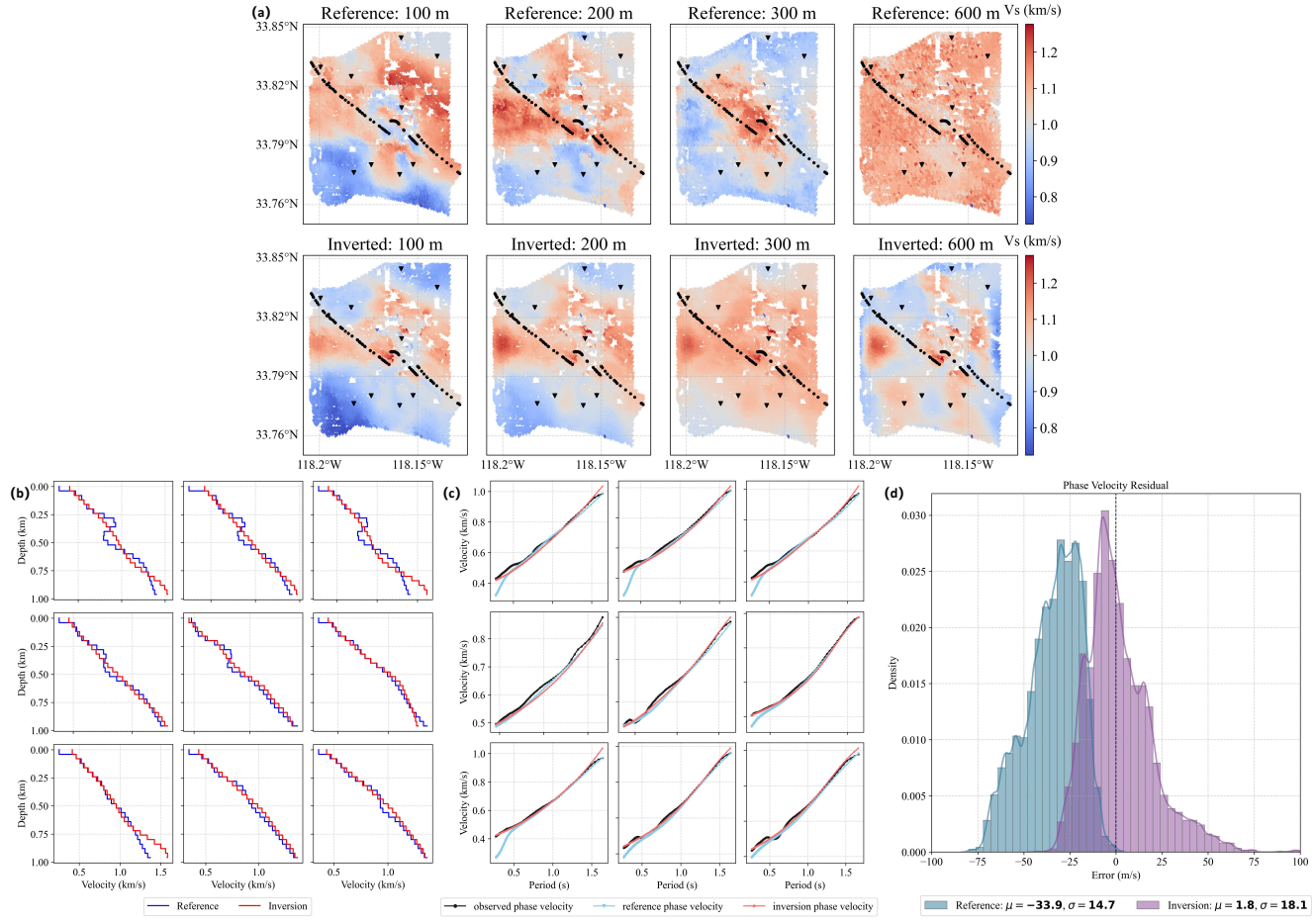


Figure 12. Generalization performance on real-world Long Beach data from the OpenSWI-real dataset. (a) Reference (Fu et al., 2022) and predicted v_s slices at depths of 100, 200, 300, and 600 m. (b) 1-D v_s profiles at nine representative locations, with reference and predicted models shown in blue and red, respectively. (c) Comparison of phase velocity dispersion curves, including observed curves (black), synthetic curves from the reference model (blue) and the predicted model (red). (d) Error distributions of phase velocity with respect to observed curves, based on synthetic dispersion curves from the reference (blue) and predicted (purple) models.

For the deep case, we applied the pretrained deep inversion network to both phase and group velocity dispersion curves at 12,901 grid points provided by the CSRM project (Wen et al., 2023; Xiao et al., 2024). Figure 13a compares the predicted and reference velocity structures at depths of 20 km, 40 km, 60 km, and 80 km. Figure 13b shows 1-D profile comparisons at nine representative grid points, where black lines denote the reference models and red lines indicate the neural network predictions. Figure 13c presents the observed dispersion curves (black), along with synthetic curves generated from the reference model (blue) and the predicted models (red). Figure 13d displays the distribution of misfits between synthetic and observed dispersion curves across all grid points. The reference model achieves a mean misfit of –72.9m/s with a variance of 65.5(m/s)^2, while the

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

neural network results exhibit a mean misfit of 24.8m/s and a lower variance of 49.6(m/s)^2. These findings suggest that the trained network can recover deep crustal velocity structures with accuracy comparable to, or better than, that of the reference model—even without any fine-tuning on real data.

In summary, these experiments confirm the strong generalization ability of the proposed method across a broad range of geological settings and depth regimes. More importantly, they highlight the effectiveness of the OpenSWI dataset series in enabling the training and evaluation of deep learning-based inversion techniques. With its extensive geological diversity, structural complexity, and broad spatial coverage, the OpenSWI dataset provides a solid foundation for learning transferable representations. As demonstrated, the resulting models can produce high-quality inversion results on real-world observations without retraining or domain adaptation, positioning OpenSWI as a valuable benchmark for advancing deep learning in realistic geophysical applications.

4 Discussion

The OpenSWI dataset marks a substantial advancement in the development of AI-ready benchmark datasets for surface wave dispersion curve inversion. Compared to existing public datasets, OpenSWI offers significantly larger scale, broader spatial coverage, and enhanced geological diversity. Specifically, the OpenSWI-shallow subset contains over 22 million 1-D velocity profiles and their associated dispersion curves representing shallow subsurface structures (depths < 3 km), while the OpenSWI-deep subset comprises approximately 1.28 million samples covering deeper Earth structures down to 300 km. In addition, the OpenSWI-real dataset provides real-world observational data for validating inversion methods under practical conditions. This comprehensive suite enables robust evaluation of machine learning–based approaches across synthetic and real data scenarios. Furthermore, a complete dataset construction toolkit, SWIDP, is released alongside the dataset, allowing users to flexibly generate customized datasets tailored to specific research needs.

Experimental results show that deep learning models trained exclusively on synthetic data from OpenSWI exhibit strong generalization to real-world observations, even without fine-tuning. This underscores the importance of large-scale, high-fidelity synthetic datasets in overcoming the challenges posed by the limited availability and annotation complexity of real seismic data. Practically, this indicates that reliable inversion results can be obtained even in regions with sparse or low-quality observations, thereby lowering the threshold for deploying machine learning models in real-world geophysical applications.

Despite these strengths, several limitations remain. First, the derivation of v_p and ρ from v_s through empirical relationships may introduce systematic biases, especially in regions with complex or atypical geological structures (Brocher, 2005). Second, although OpenSWI spans a wide array of tectonic and geological environments, it still underrepresents certain extreme (e.g. anisotropic media, fluid-saturated layers) or geodynamically active (e.g. mid-ocean ridges, highly deformed orogenic belts) settings, limiting its applicability in those areas. Third, the current dataset focuses primarily on fundamental-mode Rayleigh wave dispersion curves and does not incorporate higher modes or additional geophysical observables (e.g., ellipticity, receiver functions), which constrains its utility for joint inversion frameworks (Liu et al., 2024; Jiang et al., 2025). Lastly, although

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

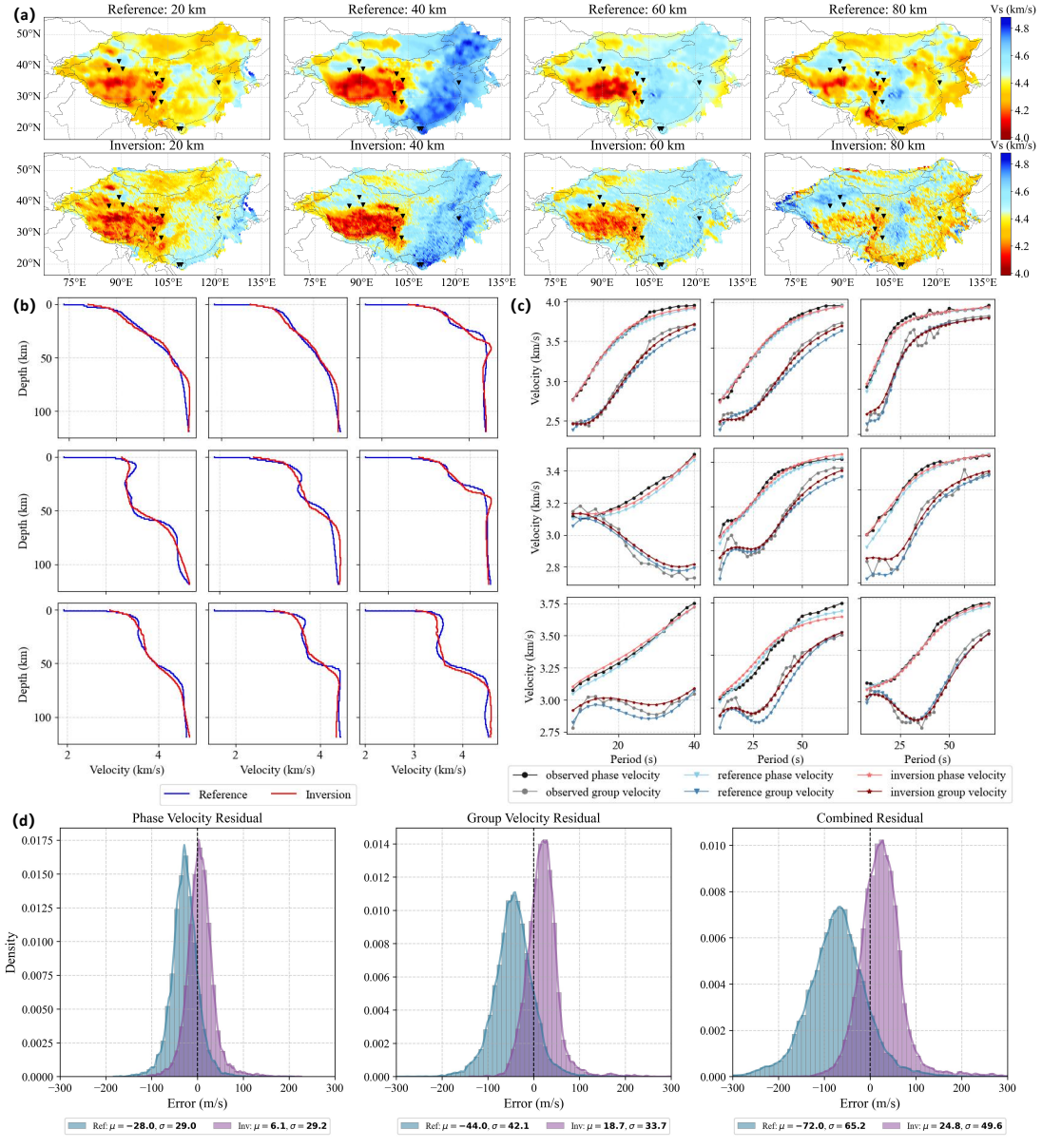


Figure 13. Generalization performance on real-world CSRM data from the OpenSWI-real dataset. (a) Reference (Xiao et al., 2024) and predicted v_s slices at depths of 20, 40, 60, and 80 km. (b) 1-D v_s profiles at nine representative locations, with reference and predicted models shown in blue and red, respectively. (c) Comparison of phase and group velocity dispersion curves, including observed curves (black), synthetic curves from the reference model (blue for group velocity, light blue for phase velocity), and from the predicted model (red for group velocity, pink for phase velocity). (d) Error distributions of phase velocity (left), group velocity (middle), and their sum (right) with respect to observed curves, based on synthetic dispersion curves from the reference (blue) and predicted (purple) models.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

OpenSWI incorporates a degree of noise and data incompleteness, it does not fully capture the complexities of real-world measurements, including uncertainties in source characteristics, instrument responses, and acquisition-related biases.

Future developments can be pursued along several interrelated directions. First, expanding the dataset's geographic coverage and geological diversity, particularly in tectonically extreme regions, would broaden its applicability. Second, integrating data across different modes, period ranges, and geological settings could enable more robust inversion approaches and improve transferability across regions. Third, incorporating additional real observational data to construct datasets suitable for hybrid or transfer learning would further enhance model generalization in field applications. Finally, including higher-mode dispersion curves and complementary geophysical observables would support more comprehensive multi-modal and multi-physics inversion strategies. We envision OpenSWI as a long-term, evolving community resource that will continue to drive data-driven advances in surface wave inversion and geophysical imaging.

5 Conclusions

In this study, we present OpenSWI, the first AI-ready benchmark dataset at the tens-of-millions scale specifically designed for surface wave dispersion curve inversion, along with a complete data generation toolkit, SWIDP. The dataset encompasses both shallow and deep subsurface velocity structures across a wide range of geological settings. Its large scale, geological diversity, and standardized formats for velocity profiles and dispersion curves provide a robust foundation for evaluating machine learning–based inversion methods. Experimental results show that models trained entirely on synthetic data from OpenSWI can generalize effectively to real-world observations, highlighting the dataset's practical value in improving the robustness and applicability of data-driven inversion approaches. Future developments will focus on expanding the dataset's geographic and geological coverage, incorporating additional geophysical observables to support more complex joint inversion tasks, and explore deeper integration with real observational data. We expect OpenSWI to serve as an open, continuously evolving community resource that promotes reproducible research and supports the broader application of machine learning methods in geophysical imaging.

6 Code and data availability

All codes, datasets, and experimental results in this study are publicly available to ensure reproducibility, validation, and further development. The Python toolkit `SWIDP`, available at https://doi.org/10.5281/zenodo.16884901 (Liu, 2025b) and https://github.com/liufeng2317/OpenSWI, provides modules for 1-D velocity profile extraction and augmentation, layer parameter conversion, dispersion curve computation, and 2-D velocity model augmentation using diffusion models. The `OpenSWI` dataset, comprising `OpenSWI-shallow`, `OpenSWI-deep`, and `OpenSWI-real`, is released in a unified format with complete metadata, accessible via https://doi.org/10.5281/zenodo.16874111 (Liu, 2025a) and https://huggingface.co/datasets/LiuFeng2317/OpenSWI. Deep learning training codes, pretrained model weights, and experimental results are also openly shared to support future research and applications.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Appendix A: Illustrative Code Examples for the OpenSWI-shallow Generation Workflow with SWIDP

```
# Import all core functions from the SWIDP package
from SWIDP import *
# initialize the model
model = SWIModel()
# Step 1: Data extraction and duplication removal
# 1-1: Load velocity model
model.load_openfwi_velocity_model()
# 1-2: Extract velocity profiles
model.get_velocity_profiles()
# 1-3: Convert units (e.g., m/s to km/s)
model.convert_unit()
# 1-4: Remove duplicates
model.unique_profiles()
# 1-5: Convert vp to vs
model.transform_vp_to_vs()
# Step 2: Data augmentation
# 2-1: Merge adjacent layers with similar vs
model.combine_same_vs()
# 2-2: Remove thin layers
model.remove_thin_layer()
for i in range(augment_times):
    # 2-3: Perturb velocity and layer thickness
    model.perturb_vs_depth()
    # 2-4: Interpolate to original depth
    model.interpolate_profile()
    # 2-5: Generate full velocity model
    model.transform_vs_to_vel_model()
# Step 3: Compute dispersion curves
# 3-1: Generate period samples
model.generate_mixed_samples()  # uniform, random, and logarithmic sampling
# 3-2: Calculate dispersion curves
model.calculate_dispersion()
# Step 4: Save data
model.save_velocity_model()      # [depth, vp, vs, density]
model.save_dispersion_curves()  # [period, phase velocity, group velocity]
```

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Appendix B: Illustrative Code Examples for the OpenSWI-deep Generation Workflow with SWIDP

```
# Import all core functions from the SWIDP package
from SWIDP import *
# initialize the model
model = SWIModel()
# 1-1 load velocity profiles
model.extract_velocity_profiles()
# 1-2 interpolate velocity profiles
model.interpolate_velocity_profiles()
# 1-3 combine thin layers (remove extremely thin layers)
model.combine_thin_sandwich()
# 1-4 smooth velocity profiles (optional)
model.smooth_vs_by_node_interp()
# 2. find moho depth
model.find_moho_depth()
# 3-1 augment velocity model (Crust-Moho-Mantle)
model.augment_crust_moho_mantle()
# 3-2 transform velocity model to velocity model
model.transform_vs_to_vel_model()
# 4-1 generate mixed samples
model.generate_mixed_samples() # uniform, random, and  logarithmic sampling
# 4-2 calculate dispersion curves
model.calculate_dispersion()
# 5. save velocity model and dispersion curves
model.save_velocity_model() # [depth, vp, vs, density]
model.save_dispersion_curves() # [period, phase velocity, group velocity]
```

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Appendix C: Diffusion Probabilistic Models for Continually Augmenting the OpenSWI-shallow Subsets

C1 Introduction to Denoising Diffusion Probabilistic Models (DDPMs)

Denoising Diffusion Probabilistic Models (DDPMs) are a class of powerful generative models that progressively refine noisy data to generate realistic outputs (Ho et al., 2020; Taufik et al., 2024). The core principle of DDPMs involves a two-step diffusion process: a forward process in which noise is progressively added to the data, and a reverse process in which the model learns to remove the noise and recover the original data distribution. In this study, DDPMs are applied to model and augment geological structures within the OpenFWI dataset (Deng et al., 2021), which consists of five subsets: FlatVel-A, FlatFault-A, CurveVel-A, CurveFault-A, and Style-A. These subsets represent various subsurface geophysical features, and by learning their distribution characteristics, DDPMs are capable of generating new, physically plausible velocity models that exhibit complex geological features such as faults, folds, and field-style structures.

C2 Core Principle of DDPM

DDPMs are based on two main processes:

- **Forward diffusion**: Starting from an input data point, Gaussian noise is progressively added in multiple steps, transforming the data into pure noise.
- **Reverse diffusion**: The model learns to reverse this process, starting from random noise and progressively denoising it to recover the underlying data distribution.

The reverse denoising process is learned by training a neural network to predict the noise added at each diffusion step. The objective is to minimize the difference between the predicted noise and the actual noise, enabling the model to generate realistic data that follows the original distribution. Formally, the training loss function is defined as:

$$L(\theta) = \mathbb{E}_{q(\mathbf{x}_0)} \left[\sum_{t=1}^{T} \|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon_t\|^2 \right] \tag{C1}$$

where ϵ_θ is the predicted noise at step t, and ϵ_t is the actual noise added during the forward diffusion process.

For further details on the DDPM methodology, please refer to Ho et al. (2020). Additionally, an implementation of DDPM in PyTorch is available at https://github.com/lucidrains/denoising-diffusion-pytorch.

C3 Model Architecture and Training Configuration

The DDPM model used in this study follows a U-Net architecture with the following key components:

- **U-Net architecture**: A convolutional neural network with an encoder-decoder structure. The encoder reduces the spatial resolution, and the decoder restores it to the original resolution (64×64). The architecture includes residual blocks and batch normalization.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

- **Noise schedule**: A linear noise schedule is applied during the forward diffusion process, where the variance of the Gaussian noise increases progressively with each step (total 1000 steps).
- **Optimizer**: Adam optimizer with a learning rate of $1e^{-6}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.
- **Training duration**: The model was trained for 5000 epochs with a batch size of 256.

The training objective is to minimize the difference between the predicted and actual noise added during the forward diffusion process, as described by the loss function in the previous section.

C4 DDPM sampling and OpenSWI-shallow datasets Generation

After training, the DDPM model is used for continuous data augmentation by generating new velocity models. This process involves sampling Gaussian noise and running the reverse diffusion process to produce realistic velocity models. The generated models reflect a variety of subsurface features, such as faults and complex sedimentary structures, ensuring physical plausibility.

To facilitate the integration of the DDPM-generated models into the OpenSWI-shallow dataset, we provide a set of tools in the SWIDP pipline. These tools enable the extraction and conversion of the DDPM sampling results into 1-D velocity models, as required by OpenSWI-shallow. The process includes the following key steps:

- **DDPM sampling**: The DDPM model generates new velocity models by progressively denoising random Gaussian noise.
- **Denormalization**: The generated models, initially in normalized form, are denormalized to match the required velocity range.
- **Profile extraction and rationalization**: The velocity models are then extracted into 1-D velocity profiles and rationalized to ensure geological consistency.
- **Dispersion curve calculation**: The rationalized 1-D velocity profiles are used to calculate the corresponding dispersion curves, which are essential for surface wave inversion tasks.

By continually generating new data and performing the above operations, the OpenSWI-shallow dataset is augmented with a diverse set of realistic velocity profiles, further expanding the dataset's coverage and variability for improved inversion model robustness.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Open Access Earth System Science Data Discussions

Appendix D: Transformer-based Network Architecture for Different Datasets

Table D1. Transformer-based Network Architecture for Different Datasets

Dataset	Input Shape	Embedding Dim.	Transformer Blocks	Attention Heads	Output Shape
OpenSWI-shallow	3×100	64	3	8	1×70
OpenSWI-deep	3×300	128	3	8	1×300

Note: The input shape consists of three features: period, phase velocity, and group velocity. The output shape corresponds to the shear-wave velocity (v_s).

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Author contributions. F. LIU: Conceptualization, Data Curation, Methodology, Software, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Visualization
S. ZHAO: Conceptualization, Data Curation, Methodology, Formal Analysis, Writing - Review & Editing
X. GU: Resources, Supervision, Writing - Review & Editing, Visualization
F. LING: Resources, Supervision, Writing - Review & Editing
P. ZHUANG: Resources, Supervision, Writing - Review & Editing
Y. LI: Supervision, Writing - Review & Editing, Project Administration
R. SU: Supervision, Writing - Review & Editing, Funding Acquisition
L. Fang: Writing - Review & Editing
L. Zhou: Writing - Review & Editing
J. HUANG: Writing - Review & Editing
L. BAI: Supervision, Writing - Review & Editing, Funding Acquisition

Competing interests. The authors declare no competing interests.

Acknowledgements. This work was supported by the Shanghai Artificial Intelligence Laboratory, and . We also acknowledge the Shanghai Artificial Intelligence Laboratory for providing computational resources and the Science Discovery Platform (Intern-Discovery; available at https://discovery.intern-ai.org.cn/) for offering the testing and demonstration environment.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

References

Aleardi, M. and Stucchi, E.: A Hybrid Residual Neural Network–Monte Carlo Approach to Invert Surface Wave Dispersion Data, Near Surface Geophysics, 19, 397–414, https://doi.org/10.1002/nsg.12163, 2021.

Bensen, G. D., Ritzwoller, M. H., Barmin, M. P., Levshin, A. L., Lin, F., Moschetti, M. P., Shapiro, N. M., and Yang, Y.: Processing Seismic Ambient Noise Data to Obtain Reliable Broad-Band Surface Wave Dispersion Measurements, Geophysical Journal International, 169, 1239–1260, https://doi.org/10.1111/j.1365-246X.2007.03374.x, 2007.

Berg, E. M., Lin, F.-C., Allam, A., Schulte-Pelkum, V., Ward, K. M., and Shen, W.: Shear Velocity Model of Alaska via Joint Inversion of Rayleigh Wave Ellipticity, Phase Velocities, and Receiver Functions across the Alaska Transportable Array, Journal of Geophysical Research: Solid Earth, 125, https://doi.org/10.1029/2019jb018582, 2020.

Blom, N., Gokhberg, A., and Fichtner, A.: Seismic Waveform Tomography of the Central and Eastern Mediterranean Upper Mantle, Solid Earth, 11, 669–690, https://doi.org/10.5194/se-11-669-2020, 2020.

Brocher, T. M.: Empirical Relations between Elastic Wavespeeds and Density in the Earth's Crust, Bulletin of the Seismological Society of America, 95, 2081–2092, https://doi.org/10.1785/0120050077, 2005.

Cai, A., Qiu, H., and Niu, F.: Semi-Supervised Surface Wave Tomography With Wasserstein Cycle-Consistent GAN: Method and Application to Southern California Plate Boundary Region, Journal of Geophysical Research: Solid Earth, 127, e2021JB023 598, https://doi.org/10.1029/2021JB023598, 2022.

Cao, R., Earp, S., De Ridder, S. A. L., Curtis, A., and Galetti, E.: Near-Real-Time near-Surface 3D Seismic Velocity and Uncertainty Models by Wavefield Gradiometry and Neural Network Inversion of Ambient Seismic Noise, GEOPHYSICS, 85, KS13–KS27, https://doi.org/10.1190/geo2018-0562.1, 2020.

Chen, X., Xia, J., Feng, J., Cheng, F., Pang, J., and Hong, Y.: Why Choose Deep Learning for Surface-Wave Inversion, Surveys in Geophysics, 46, 695–722, https://doi.org/10.1007/s10712-025-09882-y, 2025.

Colli, L., Fichtner, A., and Bunge, H.-P.: Full Waveform Tomography of the Upper Mantle in the South Atlantic Region: Imaging a Westward Fluxing Shallow Asthenosphere?, Tectonophysics, 604, 26–40, https://doi.org/10.1016/j.tecto.2013.06.015, 2013.

Çubuk-Sabuncu, Y., Taymaz, T., and Fichtner, A.: 3-D Crustal Velocity Structure of Western Turkey: Constraints from Full-Waveform Tomography, Physics of the Earth and Planetary Interiors, 270, 90–112, https://doi.org/10.1016/j.pepi.2017.06.014, 2017.

Deng, C., Feng, S., Wang, H., Zhang, X., Jin, P., Feng, Y., Zeng, Q., Chen, Y., and Lin, Y.: OpenFWI: Large-scale Multi-Structural Benchmark Datasets for Full Waveform Inversion, Neural Information Processing Systems, 2021.

Feng, S., Wang, H., Deng, C., Feng, Y., Liu, Y., Zhu, M., Jin, P., Chen, Y., and Lin, Y.: EFWI Multiparameter Benchmark Datasets for Elastic Full Waveform Inversion of Geophysical Properties, Advances in Neural Information Processing Systems, 36, 23 701–23 713, 2023.

Fichtner, A. and Villaseñor, A.: Crust and Upper Mantle of the Western Mediterranean – Constraints from Full-Waveform Inversion, Earth and Planetary Science Letters, 428, 52–62, https://doi.org/10.1016/j.epsl.2015.07.038, 2015.

Fichtner, A., Bunge, H.-P., and Igel, H.: The Adjoint Method in Seismology, Physics of the Earth and Planetary Interiors, 157, 86–104, https://doi.org/10.1016/j.pepi.2006.03.016, 2006.

Fichtner, A., Kennett, B. L. N., Igel, H., and Bunge, H.-P.: Full Seismic Waveform Tomography for Upper-Mantle Structure in the Australasian Region Using Adjoint Methods, Geophysical Journal International, 179, 1703–1725, https://doi.org/10.1111/j.1365-246x.2009.04368.x, 2009.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Fichtner, A., Kennett, B. L., Igel, H., and Bunge, H.-P.: Full Waveform Tomography for Radially Anisotropic Structure: New Insights into Present and Past States of the Australasian Upper Mantle, Earth and Planetary Science Letters, 290, 270–280, https://doi.org/10.1016/j.epsl.2009.12.003, 2010.

Fichtner, A., Trampert, J., Cupillard, P., Saygin, E., Taymaz, T., Capdeville, Y., and Villaseñor, A.: Multiscale Full Waveform Inversion, Geophysical Journal International, 194, 534–556, https://doi.org/10.1093/gji/ggt118, 2013.

Fichtner, A., van Herwaarden, D.-P., Afanasiev, M., Simutė, S., Krischer, L., Çubuk-Sabuncu, Y., Taymaz, T., Colli, L., Saygin, E., Villaseñor, A., Trampert, J., Cupillard, P., Bunge, H.-P., and Igel, H.: The Collaborative Seismic Earth Model: Generation 1, Geophysical Research Letters, 45, 4007–4016, https://doi.org/10.1029/2018gl077338, 2018.

Foti, S., Comina, C., Boiero, D., and Socco, L.: Non-Uniqueness in Surface-Wave Inversion and Consequences on Seismic Site Response Analyses, Soil Dynamics and Earthquake Engineering, 29, 982–993, https://doi.org/10.1016/j.soildyn.2008.11.004, 2009.

Foti, S., Lai, C., Rix, G. J., and Strobbia, C.: Surface Wave Methods for Near-Surface Site Characterization, CRC Press, 0 edn., ISBN 978-0-429-17853-5, https://doi.org/10.1201/b17268, 2014.

Fu, L., Pan, L., Li, Z., Dong, S., Ma, Q., and Chen, X.: Improved High-resolution 3D vs Model of Long Beach, CA: Inversion of Multimodal Dispersion Curves from Ambient Noise of a Dense Array, Geophysical Research Letters, 49, e2021GL097 619, https://doi.org/10.1029/2021GL097619, 2022.

Gan, Y., Yang, Z., Pan, L., Sun, Y.-C., Zhang, D., Gao, Y., and Chen, X.: Deep Learning-Based Dispersion Spectrum Inversion for Surface Wave Exploration, IEEE Transactions on Geoscience and Remote Sensing, 62, 1–11, https://doi.org/10.1109/TGRS.2024.3399033, 2024.

Gao, H., Wu, X., Sun, X., Hou, M., Gao, H., Wang, G., and Sheng, H.: cigFacies: A Massive-Scale Benchmark Dataset of Seismic Facies and Its Application, Earth System Science Data, 17, 595–609, https://doi.org/10.5194/essd-17-595-2025, 2025.

Haskell, N. A.: The Dispersion of Surface Waves on Multilayered Media, in: Vincit Veritas: A Portrait of the Life and Work of Norman Abraham Haskell, 1905–1970, edited by Ben-Menahem, A., vol. 43, pp. 86–103, American Geophysical Union, Washington, D. C., ISBN 978-0-87590-762-8, https://doi.org/10.1785/BSSA0430010017, 1953.

Herrmann, R. B.: Computer Programs in Seismology: An Evolving Tool for Instruction and Research, Seismological Research Letters, 84, 1081–1088, https://doi.org/10.1785/0220110096, 2013.

Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, https://doi.org/10.48550/arXiv.2006.11239, 2020.

Hu, J., Qiu, H., Zhang, H., and Ben-Zion, Y.: Using Deep Learning to Derive Shear-Wave Velocity Models from Surface-Wave Dispersion Data, Seismological Research Letters, 91, 1738–1751, https://doi.org/10.1785/0220190222, 2020.

Huang, X., Yu, Z., Wang, W., and Wang, F.: JointNet: A Multimodal Deep Learning-Based Approach for Joint Inversion of Rayleigh Wave Dispersion and Ellipticity, Bulletin of the Seismological Society of America, 114, 627–641, https://doi.org/10.1785/0120230199, 2024.

Jiang, Y., Ma, J., Ning, J., Li, J., Wu, H., and Bao, T.: One-Fit-All Transformer for Multimodal Geophysical Inversion: Method and Application, Journal of Geophysical Research: Machine Learning and Computation, 2, e2024JH000 432, https://doi.org/10.1029/2024JH000432, 2025.

Krischer, L., Fichtner, A., Boehm, C., and Igel, H.: Automated Large-scale Full Seismic Waveform Inversion for North America and the North Atlantic, Journal of Geophysical Research: Solid Earth, 123, 5902–5928, https://doi.org/10.1029/2017JB015289, 2018.

Lin, F.-C., Li, D., Clayton, R. W., and Hollis, D.: High-Resolution 3D Shallow Crustal Structure in Long Beach, California: Application of Ambient Noise Tomography on a Dense Seismic Array, Geophysics, 78, Q45–Q56, https://doi.org/10.1190/geo2012-0453.1, 2013.

Liu, F.: OpenSWI-dataset, https://doi.org/10.5281/zenodo.16874111, 2025a.

Liu, F.: OpenSWI-toolbox, https://doi.org/10.5281/zenodo.16884901, 2025b.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Liu, F., Li, J., Fu, L., and Lu, L.: Multimodal Surface Wave Inversion with Automatic Differentiation, Geophysical Journal International, 238, 290–312, https://doi.org/10.1093/gji/ggae155, 2024.

Liu, F., Deng, B., Su, R., Bai, L., and Ouyang, W.: DispFormer: Pretrained Transformer for Flexible Dispersion Curve Inversion from Global Synthesis to Regional Applications, https://doi.org/10.48550/ARXIV.2501.04366, 2025.

Lu, Y., Stehly, L., Paul, A., and AlpArray Working Group: High-Resolution Surface Wave Tomography of the European Crust and Uppermost Mantle from Ambient Seismic Noise, Geophysical Journal International, 214, 1136–1150, https://doi.org/10.1093/gji/ggy188, 2018.

Luo, Y., Huang, Y., Yang, Y., Zhao, K., Yang, X., and Xu, H.: Constructing Shear Velocity Models from Surface Wave Dispersion Curves Using Deep Learning, Journal of Applied Geophysics, 196, 104 524, https://doi.org/10.1016/j.jappgeo.2021.104524, 2022.

Merrifield, T. P., Griffith, D. P., Zamanian, S. A., Gesbert, S., Sen, S., De La Torre Guzman, J., Potter, R. D., and Kuehl, H.: Synthetic Seismic Data for Training Deep Learning Networks, Interpretation, 10, SE31–SE39, https://doi.org/10.1190/INT-2021-0193.1, 2022.

Michelini, A., Cianetti, S., Gaviano, S., Giunchi, C., Jozinović, D., and Lauciani, V.: INSTANCE – the Italian Seismic Dataset for Machine Learning, Earth System Science Data, 13, 5509–5544, https://doi.org/10.5194/essd-13-5509-2021, 2021.

Mousavi, S. M., Sheng, Y., Zhu, W., and Beroza, G. C.: STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI, IEEE Access, 7, 179 464–179 476, https://doi.org/10.1109/ACCESS.2019.2947848, 2019.

Park, C. B., Miller, R. D., and Xia, J.: Multichannel Analysis of Surface Waves, GEOPHYSICS, 64, 800–808, https://doi.org/10.1190/1.1444590, 1999.

Pasyanos, M. E., Masters, T. G., Laske, G., and Ma, Z.: LITHO1.0: An Updated Crust and Lithospheric Model of the Earth, Journal of Geophysical Research: Solid Earth, 119, 2153–2173, https://doi.org/10.1002/2013JB010626, 2014.

Reid, A., Olivier, G., and Jones, T.: Ambient Noise Tomography: A Sensitive, Rapid, Passive Seismic Technique for Mineral Exploration, SEG Discovery, pp. 17–26, https://doi.org/10.5382/SEGnews.2025-140.fea-01, 2025.

Rickers, F., Fichtner, A., and Trampert, J.: The Iceland–Jan Mayen Plume System and Its Impact on Mantle Dynamics in the North Atlantic Region: Evidence from Full-Waveform Inversion, Earth and Planetary Science Letters, 367, 39–51, https://doi.org/10.1016/j.epsl.2013.02.022, 2013.

Shapiro, N. M. and Campillo, M.: Emergence of Broadband Rayleigh Waves from Correlations of the Ambient Seismic Noise, Geophysical Research Letters, 31, 2004GL019 491, https://doi.org/10.1029/2004GL019491, 2004.

Shapiro, N. M. and Ritzwoller, M. H.: Monte-Carlo Inversion for a Global Shear-Velocity Model of the Crust and Upper Mantle, Geophysical Journal International, 151, 88–105, https://doi.org/10.1046/j.1365-246X.2002.01742.x, 2002.

Shen, W., Ritzwoller, M. H., and Schulte-Pelkum, V.: A 3-D Model of the Crust and Uppermost Mantle beneath the Central and Western US by Joint Inversion of Receiver Functions and Surface Wave Dispersion, Journal of Geophysical Research: Solid Earth, 118, 262–276, https://doi.org/10.1029/2012JB009602, 2013.

Shen, W., Ritzwoller, M. H., Kang, D., Kim, Y., Lin, F.-C., Ning, J., Wang, W., Zheng, Y., and Zhou, L.: A Seismic Reference Model for the Crust and Uppermost Mantle beneath China from Surface Wave Dispersion, Geophysical Journal International, 206, 954–979, https://doi.org/10.1093/gji/ggw175, 2016.

Simutė, S., Steptoe, H., Cobden, L., Gokhberg, A., and Fichtner, A.: Full-waveform Inversion of the Japanese Islands Region, Journal of Geophysical Research: Solid Earth, 121, 3722–3741, https://doi.org/10.1002/2016jb012802, 2016.

Socco, L. and Strobbia, C.: Surface-wave Method for Near-surface Characterization: A Tutorial, Near Surface Geophysics, 2, 165–185, https://doi.org/10.3997/1873-0604.2004015, 2004.

https://doi.org/10.5194/essd-2025-502
Preprint. Discussion started: 5 November 2025

Taufik, M. H., Wang, F., and Alkhalifah, T.: Learned Regularizations for Multi-Parameter Elastic Full Waveform Inversion Using Diffusion Models, Journal of Geophysical Research: Machine Learning and Computation, 1, e2024JH000 125, https://doi.org/10.1029/2024JH000125, 2024.

Thomson, W. T.: Transmission of Elastic Waves through a Stratified Solid Medium, Journal of Applied Physics, 21, 89–93, https://doi.org/10.1063/1.1699629, 1950.

Wang, F., Song, X., and Li, J.: Deep Learning-Based H-κ Method (HkNet) for Estimating Crustal Thickness and Vp/Vs Ratio From Receiver Functions, Journal of Geophysical Research: Solid Earth, 127, e2022JB023 944, https://doi.org/10.1029/2022JB023944, 2022.

Wang, F., Huang, X., and Alkhalifah, T. A.: A Prior Regularized Full Waveform Inversion Using Generative Diffusion Models, IEEE Transactions on Geoscience and Remote Sensing, 61, 1–11, https://doi.org/10.1109/tgrs.2023.3337014, 2023a.

Wang, F., Song, X., and Li, M.: A Deep-Learning-Based Approach for Seismic Surface-Wave Dispersion Inversion (SfNet) with Application to the Chinese Mainland, Earthquake Science, 36, 147–168, https://doi.org/10.1016/j.eqs.2023.02.007, 2023b.

Wang, G., Wu, X., and Zhang, W.: cigChannel: A Large-Scale 3D Seismic Dataset with Labeled Paleochannels for Advancing Deep Learning in Seismic Interpretation, Earth System Science Data, 17, 3447–3471, https://doi.org/10.5194/essd-17-3447-2025, 2025.

Wathelet, M., Jongmans, D., and Ohrnberger, M.: Surface-wave Inversion Using a Direct Search Algorithm and Its Application to Ambient Vibration Measurements, Near Surface Geophysics, 2, 211–221, https://doi.org/10.3997/1873-0604.2004018, 2004.

Wen, L., Yu, S., Department of Geosciences, State University of New York at Stony Brook, Stony Brook, NY 11794, USA, Laboratory of Seismology and Physics of Earth's Interior; School of Earth and Space Sciences, University of Science and Technology of China, Hefei 230026, China, and Department of Earth Sciences, National Natural Science Foundation of China, Beijing 100085, China: The China Seismological Reference Model Project, Earth and Planetary Physics, 7, 521–532, https://doi.org/10.26464/epp2023078, 2023.

Xia, J., Miller, R. D., and Park, C. B.: Estimation of Near-surface Shear-wave Velocity by Inversion of Rayleigh Waves, GEOPHYSICS, 64, 691–700, https://doi.org/10.1190/1.1444578, 1999.

Xiao, X., Cheng, S., Wu, J., Wang, W., Sun, L., Wang, X., Ma, J., Tong, Y., Liang, X., Tian, X., Li, H., Chen, Q.-F., Yu, S., and Wen, L.: CSRM-1.0: A China Seismological Reference Model, Journal of Geophysical Research: Solid Earth, 129, e2024JB029 520, https://doi.org/10.1029/2024JB029520, 2024.

Xie, J., Chu, R., and Yang, Y.: 3-D Upper-Mantle Shear Velocity Model beneath the Contiguous United States Based on Broadband Surface Wave from Ambient Seismic Noise, Pure and Applied Geophysics, 175, 3403–3418, https://doi.org/10.1007/s00024-018-1881-2, 2018.

Xin, H., Zhang, H., Kang, M., He, R., Gao, L., and Gao, J.: High-resolution Lithospheric Velocity Structure of Continental China by Double-difference Seismic Travel-time Tomography, Seismological Research Letters, 90, 229–241, https://doi.org/10.1785/0220180209, 2019.

Yablokov, A., Lugovtsova, Y., and Serdyukov, A.: Uncertainty Quantification of Multimodal Surface Wave Inversion Using Artificial Neural Networks, GEOPHYSICS, 88, KS1–KS11, https://doi.org/10.1190/geo2022-0261.1, 2023.

Yablokov, A. V., Serdyukov, A. S., Loginov, G. N., and Baranov, V. D.: An Artificial Neural Network Approach for the Inversion of Surface Wave Dispersion Curves, Geophysical Prospecting, 69, 1405–1432, https://doi.org/10.1111/1365-2478.13107, 2021.

Yang, Y. and Ritzwoller, M. H.: Characteristics of Ambient Seismic Noise as a Source for Surface Wave Tomography, Geochemistry, Geophysics, Geosystems, 9, 2007GC001 814, https://doi.org/10.1029/2007GC001814, 2008.