JENNIFER-H2: Joint Embedding Neural Network Interpolation For Earth Resources – Hydrogen

1 Excellence

1.1 State of the art, knowledge needs and project objectives

Europe's recent energy security crisis has underscored the continent's vulnerability to supply disruptions and the urgent need to develop domestic, low-carbon energy sources. Yet Europe faces a critical paradox: the continent possesses vast subsurface energy resources, including deep geothermal reservoirs, natural hydrogen systems generated by serpentinization and radiolysis, and offshore oil and gas reserves, yet it cannot fully characterize or access them at scale [1]. Natural (or "white") hydrogen, a zero-carbon primary energy source generated continuously by geochemical reactions in the Earth's crust, represents the most transformative opportunity, with recent assessments suggesting a global subsurface endowment on the order of trillions of metric tons [2, 3]. Even a small recoverable fraction of this resource could sustain global energy demand for centuries. Within Europe there is a high potential for natural hydrogen, with high-potential zones identified across Hungary, Denmark, Poland, Serbia, and the failed rift basins of the North Sea [4]. Current exploration methods, including expensive seismic imaging, sparse drilling campaigns, and manual interpretation of fragmented datasets, cannot identify or characterize these hydrogen systems at the speed and scale required by the energy transition [2]. At the same time, enormous datasets from the European Plate Observing System (EPOS), the Norwegian DISKOS database, and the International Ocean Discovery Program (IODP) describe subsurface processes at petabyte scale [5–7], but no unified framework exists to integrate them into a coherent view of where hydrogen is generated, accumulates, migrates, and can be extracted economically and safely. **This is a problem that demands both geological expertise and large-scale Artificial Intelligence (AI) .**

Natural hydrogen is generated in the Earth's crust through several well-documented but incompletely quantified mechanisms. Serpentinization, the hydration of iron-bearing minerals in ultramafic rocks such as olivine and pyroxene, produces molecular hydrogen as ferrous iron is oxidized during the formation of serpentine, magnetite, and brucite [8]. This process is most vigorous at temperatures between 200 and 350°C and occurs wherever ultramafic lithologies interact with circulating water, including mid-ocean ridges, ophiolite complexes, and cratonic basement [9]. Radiolysis of water by ionizing radiation from uranium, thorium, and potassium decay in crystalline basement rocks provides a second, pervasive source of hydrogen that operates independently of temperature and tectonic setting [10]. Additional generation pathways include thermal decomposition of organic matter and iron-bearing carbonates at elevated temperatures in sedimentary basins [10]. The relative contributions of these mechanisms vary enormously across geological provinces, and in many settings multiple processes operate simultaneously, making source attribution a persistent challenge for exploration. Unlike petroleum systems, where hydrocarbons migrate once from a mature source rock into a static trap, hydrogen systems are dynamic: generation is ongoing, reservoirs may charge and discharge on timescales of years to millennia, and surface seeps can appear, migrate laterally, and vanish seasonally [1, 11]. These characteristics make hydrogen exploration fundamentally more difficult than conventional hydrocarbon exploration and create a compelling case for AI and foundation models. A foundation model trained on large, multi-modal subsurface datasets can address these challenges by learning latent representations that integrate geophysical, geochemical, and geological signals across scales, capturing subtle correlations between, for example, magnetic anomaly patterns, heat flow gradients, and fault network geometry that would be invisible to manual interpretation. The sparsity and heterogeneity of hydrogen-relevant data, rather than being obstacles, are precisely the conditions under which self-supervised models that learn to reconstruct masked inputs can extract the most value, because they force the model to internalize the underlying physical relationships rather than memorize surface-level patterns.

Recent advances in AI and machine learning have demonstrated that pretraining self-supervised models on very large data sets substantially increases the ability for models to generalize across tasks. Such models are known as "foundation" models [12]. A critical motivation for foundation models is the "platonic representation

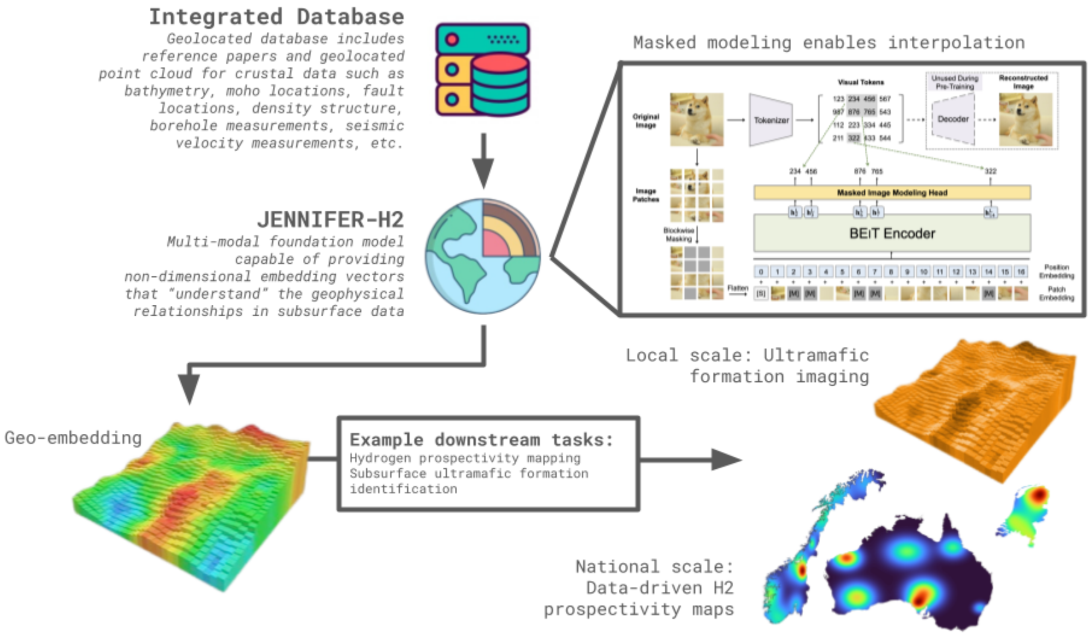

Figure 1: The JENNIFER model pipeline. This project will create both an open source integrated database and an AI model.

hypothesis" [13], which asserts that AI models' latent representations are converging. That is, the latent vector representations that are used to predict and generate text and images in AI models are learning effectively the same correlation structures. This has been demonstrated for text [14] and vision model tasks [15]. An analogous "platonic hypothesis" for subsurface modeling posits that beneath observable geophysical measurements lies a lower-dimensional manifold of latent geological processes that can be discovered through self-supervised learning. Thus, independently of particular choices of architecture and dataset (which are often arbitrary), a sufficiently-large model ought to converge to true geophysical principles, because they are the only true route to generalization at scale. For geophysical models, these "geophysical" embeddings will capture geologically meaningful relationships not explicit in individual measurement modalities, analogous to how AlphaEarth, a land-surface foundation model (not a subsurface model), produces meaningful embeddings from diverse Earth observation data [16]. A foundation model of the Earth's subsurface would thus improve subsurface characterization and imaging fidelity. A subsurface foundation model could serve as a backbone for current robotic decision-making frameworks, particularly probabilistic approaches like the partially-observable Markov decision process, which have proven superior at quantifying spatial uncertainty around mineralization targets [17, 18] and could also be used for mapping subsurface ultramafic formations that would be ideal hydrogen generation zones. Contemporary borehole datasets, exemplified by the LILY database [7], or the Norwegian DISKOS database [6], provide unprecedented training resources for such systems. Yet these data remain in data repositories and have not been used together to train any "embeddings" style AI expert model.

We propose to create the Joint Embedding Neural Network Interpolation For Earth Resources – Hydrogen (JENNIFER-H2) foundation model, an masked joint embedding architecture adapted for subsurface data. Inspired by the Joint Embedding Predictive Architecture [19], JENNIFER will be trained on a constructed dataset based on the petabytes of aggregated geophysical data from boreholes and seismic surveys coupled with global surveys and models. The model will learn to "infill" randomly masked sections of subsurface data, thereby acquiring a deep understanding of wave physics and rock property distributions without requiring explicit PDE solvers. A similar self-supervised approach is what enabled early language models to reach sufficient scale necessary to acquire nuanced understanding of linguistic principles [20]. Once trained, JENNIFER enables probabilistic zero-shot inversion. Given sparse, noisy data whether from seismic studies or boreholes, the model generates a posterior distribution of subsurface characteristics, a task that would require weeks using traditional methods or otherwise be impossible. Current state-of-the-art foundation models for the subsurface are uni-modal, trained on only a single kind of data, typically taken from the output of numerical models [21, 22]. **This will represent the first time that a single foundation model capable of producing embedding vectors will be trained on multi-modal, real-world data from the subsurface.**

Embedding vectors are the latent vector representations that are common to language and vision AI tasks (e.g., [20, 23]). Embeddings are continuous vector spaces calculated by neural networks that describe the complex interactions between data for a specific observation typically produced by a neural network [24]. The latent space of an neural network represents the non-dimensional "understanding" the model possesses that describes the relationships inherent to the training data. These encoder/decoder structures have been used to great effect to build modern Large Language Models such as ChatGPT [25]. They have also been extended to solve a great number of geoscience problems, for example, earthquake identification [26], weather and climate modeling [27], and converting geological maps into 3D structural diagrams [28, 29]. Understanding this embedding space can also provide better estimations of the state space and predictability of chaotic geosystems (e.g.,[30]). Furthermore, these models can interpolate missing data and classify previously unseen observations, a critical problem across geosciences. We will leverage these properties of neural networks to build a foundation model of the Earth's crust.

This project will pursue two scientific objectives, framed around natural hydrogen systems and their subsurface expression:

- **SO1:** Develop a subsurface foundation model capable of robust zero-shot prediction of hydrogen system components (sources, reservoirs, seals, and migration pathways) with quantified uncertainty across diverse geological settings.
- **SO2:** Create universal geophysical embeddings that enable transfer learning across hydrogen-related geological tasks (for example hydrogen prospectivity mapping, geothermal assessment, and critical mineral exploration) with minimal task-specific fine-tuning, and rigorously benchmark these capabilities.

These scientific objectives both capture the difficulty of training a model such as JENNIFER to learn domain-invariant representations of hydrogen systems and highlight the real world applications that such a trained model could provide to both academic and industrial problem spaces. The central challenge is integrating heterogeneous subsurface data at scale.

While this proposal targets natural hydrogen as its primary application, the JENNIFER model is designed as a general-purpose subsurface foundation model. Hydrogen exploration is chosen as the first benchmark precisely because it presents the greatest density of unsolved problems: sparse observations, ephemeral signals, and no established exploration workflow. The same learned representations are directly applicable to geothermal resource assessment, critical mineral exploration, carbon storage site characterization, and conventional subsurface imaging. Hydrogen exploration serves as a rigorous test case because success requires the model to integrate heterogeneous data under conditions of extreme sparsity, and any model that performs well under these conditions will naturally generalize to better-constrained subsurface problems.

1.2 Research questions, theoretical approach, methodology, and work packages

RQ1: Under what conditions can a subsurface foundation model achieve robust zero-shot prediction of hydrogen system components (sources, reservoirs, seals, and migration pathways) with quantified uncertainty across diverse geological settings?

Traditional supervised learning fails catastrophically under distribution shifts. Systems that are extreme, such as high-flux natural hydrogen seeps, overpressured hydrogen-rich reservoirs, or deeply rooted degassing faults relative to background crustal conditions, become difficult to predict. JENNIFER's joint-embedding masked autoencoder architecture learns abstract representations in latent space rather than "pixel-level" reconstructions. This is critical because subsurface data exhibits extreme variability. Sedimentary basins differ fundamentally from Precambrian cratons or ultramafic massifs, yet all may host hydrogen generation, accumulation, and leakage pathways. The "joint" in joint embeddings refers to the shared, common vector space where representations of different data inputs, often from multiple data types (e.g., seismic data, wireline logs, gravity and magnetic fields, heat flow, and even text-based geological descriptions), are mapped and aligned so that their semantic relationships can be compared directly. This mirrors the integrated geological and geophysical reasoning that geoscientists already use to infer hydrogen sources, reservoirs, seals, and migration pathways.

The key scientific question is whether learned embeddings can capture domain-invariant geophysical and geochemical principles (e.g., wave propagation physics, rock property relationships, fluid–rock interaction regimes) that transfer across geological settings while remaining predictive for hydrogen system components.

Recent work demonstrates that most “out-of-distribution” tasks in materials science are actually representationally in-distribution when viewed in learned embedding spaces [31]. This suggests JENNIFER’s embeddings may achieve genuine zero-shot capability for hydrogen system prediction if they learn fundamental physical relationships rather than domain-specific correlations. However, this requires rigorous uncertainty quantification: the model must confidently abstain when extrapolating beyond its training manifold. Ultimately we will need to answer questions such as: What is the minimum training data diversity, in terms of hydrogen-relevant geological environments (serpentinizing ultramafics, radiolytic cratons, rift basins, salt-sealed reservoirs), required for embeddings to capture cross-domain invariances? Can neural network ensembles in latent space provide calibrated uncertainty for zero-shot predictions of hydrogen sources, reservoirs, seals, and migration pathways? How do embedding dimensionality and masking strategies affect out-of-distribution generalization when inferring hydrogen system properties from sparse or noisy observations?

Uncertainty quantification is critical for two reasons. First, hydrogen exploration and drilling decisions based on JENNIFER involve multi-million euro investments; confidently abstaining when extrapolating prevents costly dry wells, mis-sited infrastructure, and mischaracterized storage or production projects. Second, real observational data contain genuine geological variability that constrains the posterior distribution, unlike synthetic PDE simulations, making it essential that uncertainty estimates reflect both data coverage and true subsurface complexity rather than model artefacts.

SO2 – Universal geophysical embeddings for hydrogen-related tasks

RQ2: How can learned embeddings serve as “universal geophysical descriptors” for hydrogen-related processes that enable transfer learning across fundamentally different geological tasks with minimal task-specific fine-tuning, and how will we assess these capabilities?

JENNIFER’s embeddings should function as universal geophysical descriptors, abstract representations that capture fundamental subsurface physics and geological relationships that control hydrogen generation, migration, storage, and leakage, independent of the specific end-use application. This parallels how large language models and vision models capture domain knowledge through pre-training [32]. For example, ultramafic complexes undergoing active serpentinization, radiolytic hydrogen systems in Precambrian cratons, and deep fault-controlled degassing in rift basins each produce characteristic combinations of seismic velocity structure, density, magnetic anomalies, heat flow, and fluid properties that should cluster in embedding space. Likewise, hydrogen-charged reservoirs sealed by evaporites or shales should occupy different regions of embedding space than highly fractured, poorly sealed systems that preferentially leak to the surface.

A central question that JENNIFER can address is therefore: will embeddings learned from abundant “barren” subsurface data, where no significant hydrogen accumulations are yet known, transfer to rare but economically critical hydrogen-rich systems with minimal fine-tuning (for example, few-shot learning on limited well, seep, or flux data)? This requires embeddings to capture fundamental processes such as fluid–rock interaction, fracture-controlled flow, basin-scale migration pathways, and seal integrity rather than deposit-specific signatures. Recent work demonstrates successful few-shot learning in geoscience contexts [33], and end-to-end AI workflows have already improved targeting in complex subsurface exploration problems [34]. However, these approaches rely on task-specific models rather than general-purpose embeddings.

For hydrogen, the core question is how JENNIFER will exploit training on petabytes of diverse subsurface data to capture subtle multivariate patterns that are invisible to human experts and expose them as reusable feature vectors for a small number of high-impact hydrogen tasks, such as: (i) regional to continental-scale screening of hydrogen prospectivity, and (ii) prioritization of exploration wells within already identified plays. Other applications, including monitoring and leakage risk assessment for storage or production projects, geothermal assessment, and critical mineral exploration, are expected longer-term beneficiaries of the same embedding space but lie beyond the scope of the present project.

To assess these capabilities within the constraints of the project, we will construct a focused benchmark suite (WP3, described below) in which simple downstream models (e.g., shallow classifiers or regressors) trained on fixed JENNIFER embeddings are compared against strong task-specific baselines on a small number of held-out hydrogen-relevant provinces. Performance will be evaluated primarily in the zero-shot setting, with limited few-shot experiments (for example, 1–10 labelled wells or seep locations per region) to quantify the data-efficiency gains provided by the embeddings, rather than attempting to exhaustively cover all possible geological settings.

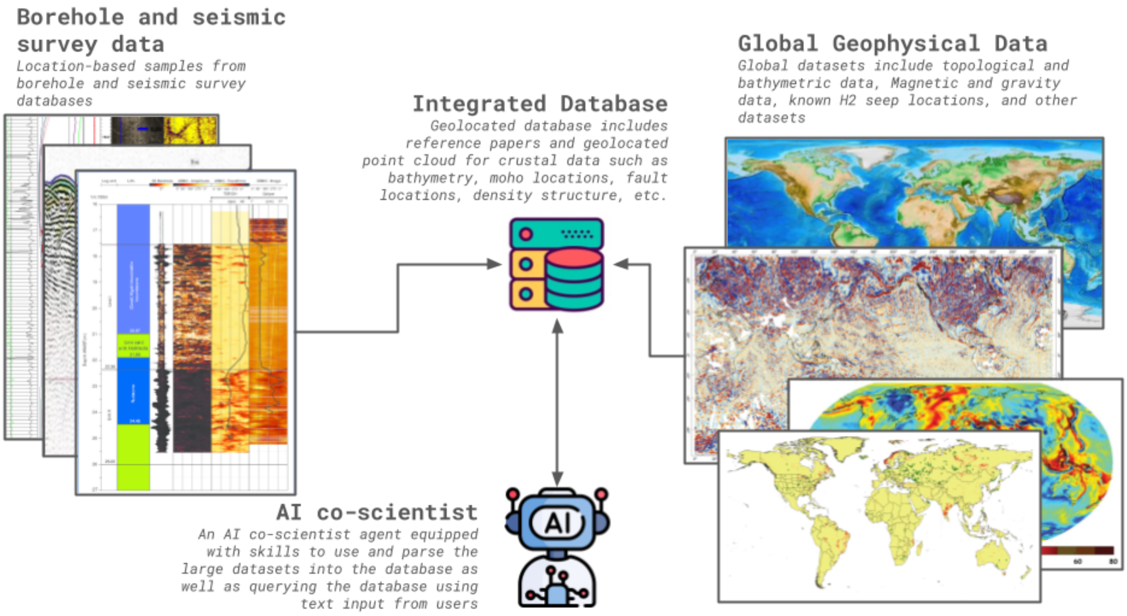

Figure 2: JENNIFER will be trained on both global scale geophysical data and local borehole and seismic survey data.

Project organization

The project is organized into three interconnected Work Packages (Fig. **??**). Work Package 1 (WP1) will construct the multi-modal training database by aggregating and harmonizing data from IODP, DISKOS, NAMSS, and EPOS. Work Package 2 (WP2) will design and train the JENNIFER foundation model on this database. Work Package 3 (WP3) will benchmark the trained model on hydrogen-specific prediction tasks. WP1 and WP2 will run partially in parallel, with WP2 beginning once an initial data subset is available in the second half of Year 1. WP3 will commence in Year 2 using preliminary model outputs and will intensify as the model matures. Each Work Package addresses both scientific objectives: WP1 provides the data diversity required by SO1 and the cross-domain coverage required by SO2; WP2 develops the model architecture and uncertainty quantification central to SO1; and WP3 directly evaluates the transferability and zero-shot capabilities defined by SO2.

Work Package 1: Building a multi-modal database of the Earth's subsurface

The foundation of any successful foundation model lies in high-quality, diverse training data. For JENNIFER, this involves aggregating multiple complementary databases. For example, the DISKOS database has in it's public repository 12,601 wellbores and 5817 seismic surveys. We have identified multiple target databases (IODP [7], DISKOS [6], NAMSS [35], EPOS [5]), providing an unprecedented training set for supervised cross-modal learning. Collectively, these datasets span the full spectrum of data maturity: from raw field recordings and archived paper logs to processed seismic volumes and interpreted geological models, covering diverse geological settings globally.

Implementing automated data quality pipelines is critical, as data preprocessing consumes a large amount of foundation model training time. We will implement automated quality control (completeness, consistency, temporal alignment, and physics-based plausibility checks) and modality-specific preprocessing for seismic and borehole data. A crucial first step will be data filtering for quality control and normalization to reduce the total amount of data, 10s of petabytes, to data comparable to modern training data sets for multi-modal vision foundation models (100s of terabytes). Unlike computer vision, where augmentations like random rotations are universal, geophysical data augmentation must respect physical constraints. This is critical to how masked encoding (Fig. 1) will be implemented. Boreholes have depth profiles based on how sediments form, seismic data has polarity, and text data could be converted to a domain specific language [29] to better target training.

A core component of this work package will be the development of an AI agent for automated data ingestion, handling, and transformation of this complex, multi-modal dataset. For JENNIFER, we will develop

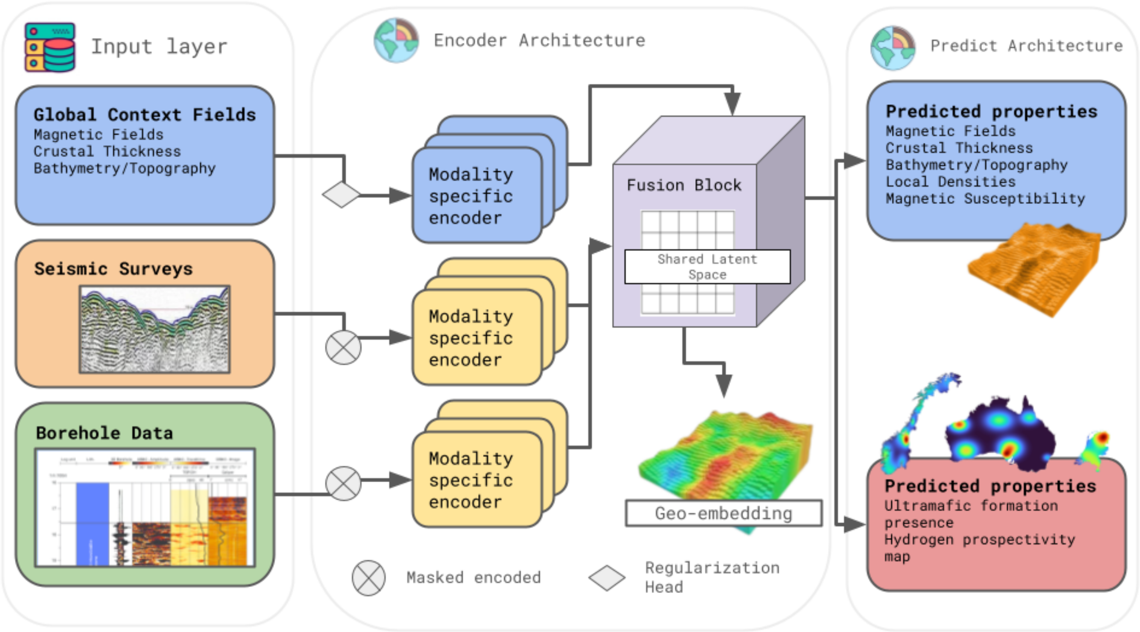

Figure 3: A diagram of the neural network architecture for the JENNIFER model.

a specialized agent that can ingest data from the target repositories (IODP LIMS, DISKOS, NAMSS, EPOS), identify and parse format variants (SEG-Y, LAS, DLIS, miniSEED, CSV), perform quality control checks, and flag anomalies for human review. This will be a necessary component of this project as the scale and heterogeneity of the target databases (up to tens of petabytes across dozens of formats, vintages, and coordinate systems) exceeds what manual or even traditional scripted pipelines (e.g., [36]) can reliably handle. The agent will serve as a persistent interface between researchers and the constructed database, enabling natural language queries such as "find all boreholes within 5 km of a 3D seismic survey in the North Sea with sonic and density logs deeper than 2000m." This lowers the barrier to entry for domain scientists who may lack the programming expertise to interact with the raw data directly, and ensures that the database remains accessible and usable beyond the lifetime of the project.

A key component of the JENNIFER model is that it should be able to provide embeddings for any location in the Earth's. Thus, JENNIFER requires globally continuous geophysical fields that provide context at any location on Earth, including regions with no subsurface sampling. Examples of this data include include satellite gravity anomalies [37], the EMAG2v3 global magnetic anomaly grid [38], topographic and bathymetric models [39], crustal thickness and structure models [40].

Deliverables for WP1: The primary deliverable of WP1 will be the constructed training database for the JENNIFER model and `python` libraries developed to ingest, sort, and filter the large datasets. Additionally, a primary deliverable will be the agent itself.

Work Package 2: Training the JENNIFER model

JENNIFER's core architecture builds on masked self-supervised learning, which has proven highly effective across domains (e.g., [41, 42]), including recent applications to borehole datasets by the PI using Masked Autoencoders [43]. While MAE learns by reconstructing masked input in data space, JENNIFER adopts the Joint Embedding Predictive Architecture (JEPA) framework [44], which instead predicts the latent representations of masked regions. Both approaches use masking to force the model to learn robust representations of underlying structure rather than memorizing local patterns, but JEPA's latent-space objective avoids overfitting to acquisition noise and survey artifacts. In our setting, this must be done on *heterogeneous* subsurface data spanning colocated borehole–seismic pairs (Fig. 1) and globally continuous geophysical fields (Fig. 2), at scales from meters to hundreds of kilometers. Jointly training across these regimes is technically challenging and will require careful architectural design, staged training, and extensive experimentation.

A core component of the model design is the asymmetry between (i) supervised, high-resolution borehole and seismic survey data available only at a finite number of locations and (ii) globally complete but lowerresolution geophysical context fields from satellite and mapping products. This asymmetry motivates JENNIFER's design. The model learns the mapping from globally available context fields to detailed subsurface properties using colocated borehole–seismic pairs as supervised training signal. At inference time, the global fields serve as the primary input, and the model predicts what a virtual borehole or seismic survey would reveal at any target location, together with calibrated uncertainty that reflects the distance in feature space from the nearest training examples. This transforms colocated pairs from a hard constraint on spatial coverage into a constraint on training set size where performance scales with the volume and diversity of data rather than being bounded by geography. As additional borehole, seismic, and hydrogen-relevant datasets are ingested, the JENNIFER model will naturally improve. JENNIFER does not require labeled hydrogen data during pre-training. The model learns subsurface representations from the full multi-modal dataset. Known hydrogen occurrences, including documented seep locations, geochemical measurements from hydrogen-bearing wells, and mapped serpentinization zones, enter exclusively as evaluation targets in WP3, where they test whether the learned representations capture hydrogen-relevant geological relationships without having been explicitly trained on them.

Architecture Design: We will implement an asymmetric encoder–predictor architecture (Fig. 3) where the encoder processes only visible (unmasked) patches while the predictor predicts the latent representations of the masked regions [45]. The encoder will be substantially larger than the lightweight predictors, because the encoder's latent representations are what transfer to downstream tasks [45]. Determining the optimal masking strategy for JENNIFER is non-trivial: for natural images, a 75% masking ratio is standard, while for video data optimal ratios reach 90% due to higher temporal redundancy [41]. Within Earth science, sea surface temperature reconstruction tolerates masking ratios up to 80% with minimal degradation [42], whereas MAE-based reconstruction of the LILY borehole dataset becomes unstable above 20% masking [43]. Given this large spread, we will conduct systematic ablation studies on representative subsets of the training data (e.g., hydrogen-fertile ultramafic provinces) to identify masking schedules that balance prediction difficulty, training efficiency, and stability. It is unlikely that a single masking ratio will suffice across all modalities; instead, we anticipate adopting curriculum learning, starting from lower masking in structurally complex or data-poor regimes and gradually increasing masking as the encoder stabilizes.

Multi-modal fusion and curriculum: Figure 3 will schematically illustrate the full encoding–predict process across the data hierarchy shown in Figs. 1 and 2. Concretely, we will implement (i) modality-specific encoders (for example, separate patch/token embeddings for seismic traces, borehole logs, gridded gravity/magnetic fields, and auxiliary text-based lithology descriptions), followed by (ii) fusion blocks that operate in the shared latent space to align these representations. In early training, we will focus on relatively "easy" prediction tasks (e.g., predicting masked borehole logs from unmasked logs and local seismic context) and only later introduce more difficult cross-scale tasks (e.g., predicting missing borehole features from global gravity and magnetic fields plus sparse seismic). This curriculum is necessary because training JENNIFER to simultaneously solve all inverse problems from the start is likely to be unstable given the petabyte-scale, highly heterogeneous input data.

Abstract representation space for JENNIFER: As introduced above, JENNIFER predicts in an abstract representation space rather than raw data space [44]. Building on the embedding concepts introduced in Section 1.1, we hypothesize that JENNIFER's latent vectors will encode hydrogen-relevant features such as source, reservoir, and seal likelihood, which will be tested through the benchmark tasks in WP3. This abstract representation space is a high-dimensional, continuous latent space in which each spatial location in the crust (for example, a voxel at given latitude, longitude, and depth) is represented by a fixed-length vector encoding the model's best estimate of its geophysical and geological state.

JEPA-style training then consists of predicting the latent representation of a "target" block (e.g., a masked depth interval, a missing segment of a seismic line, or an unsampled location in a basin) from the representations of surrounding "context" blocks, with a latent variable z capturing residual uncertainty. This design is crucial for geophysical applications where high-entropy acquisition noise and survey artifacts can obscure the predictable structural signal: by predicting in latent space rather than raw data space, JENNIFER is encouraged to learn stable, physically meaningful descriptors instead of overfitting to survey-specific noise.

Deliverables for WP2: The primary deliverable of WP2 is the JENNIFER model itself, including the `python` codes developed, trained model weights (to be hosted on HuggingFace), and documentation of the training curricula, masking strategies, and fusion architectures. Given the technical difficulty and computational demands (requiring significant computational resources [43, 44]), we explicitly scope WP2 to delivering (i) at least one full-scale multi-modal model trained on the integrated database of Figs. 3 and 2 and (ii) a suite of smaller ablation models used to quantify the impact of masking ratio, fusion strategy, and JEPA vs. standard MAE training on downstream hydrogen-relevant tasks. Crucially, we will also explore options to release the learned embeddings themselves as a global subsurface feature dataset, mirroring the AlphaEarth embeddings release [16], so that other groups can build hydrogen and subsurface applications on top of JENNIFER without retraining the foundation model.

Work Package 3: Benchmarking JENNIFER on hydrogen-relevant subsurface prediction tasks

A foundation model is only as valuable as its transferability [21]. WP3 will evaluate JENNIFER's learned representations on a small number of focused benchmark tasks designed to demonstrate zero-shot and few-shot capability for hydrogen system characterization.

Benchmark Task 1: Continental-scale hydrogen prospectivity mapping. Building on the prospectivity context outlined in Section 1.1, the benchmark task will be: given JENNIFER embeddings at a grid of locations across, e.g., Europe and the United States, can a simple downstream classifier (e.g., logistic regression or shallow neural network) predict the composite chance-of-sufficiency (COS) score for hydrogen prospectivity at each location?

We will compare three conditions: (i) baseline using only the raw geological indicators available in the published prospectivity assessments (e.g., proximity to ultramafic rocks, heat flow, fault density, sedimentary basin presence), (ii) JENNIFER embeddings alone (zero-shot), and (iii) JENNIFER embeddings plus 1, 5, or 10 labeled high-prospectivity locations per country (few-shot). Performance will be evaluated using spatial cross-validation, withholding entire countries or geological provinces during training and testing on held-out regions. The metric is area under the receiver operating characteristic curve (AUC) for binary classification (high vs. low prospectivity) and mean absolute error for continuous COS prediction. If JENNIFER embeddings improve upon the raw indicator baseline, this demonstrates that the model has learned latent geophysical relationships relevant to hydrogen systems that are not explicit in the input features alone.

This benchmark directly addresses SO2 by testing whether embeddings learned from general subsurface data (most of which comes from hydrocarbon basins, not hydrogen plays) transfer to the specialized task of hydrogen prospectivity prediction with minimal task-specific supervision.

Benchmark Task 2: Predicting hydrogen seep locations from subsurface embeddings. Natural hydrogen seeps and documented hydrogen occurrences (compiled in [3, 4, 46]) provide point observations where hydrogen reaches the surface or accumulates in drilled wells. The benchmark task is: given JENNIFER embeddings in a region surrounding a known seep or occurrence, can a simple model predict the *location* of the seep based only on the spatial pattern of embeddings, without access to the raw geophysical or geological data?

We will frame this as a binary spatial classification problem: for each documented seep location, extract embeddings on a regular grid within a 50 km radius, label the grid cell containing the seep as positive and all others as negative, and train a shallow classifier (e.g., random forest or logistic regression) on a subset of seeps, then test on held-out seeps in geologically distinct settings (e.g., train on Balkan ophiolites, test on Scandinavian cratonic seeps; train on French sedimentary basins, test on Spanish rift zones). The baseline uses hand-crafted features derived from the raw input data (gravity, magnetics, proximity to faults, etc.). Success is measured by precision-recall AUC, as the task is highly imbalanced (most locations are not seeps).

Deliverables for WP3: The primary deliverable of WP3 is a peer-reviewed benchmark paper documenting the hydrogen prospectivity and seep location prediction tasks, including open-source code, trained baseline models, and withheld test sets to enable future subsurface foundation models to report performance on the same benchmarks. Secondary deliverables include at least one application paper demonstrating improved hydrogen exploration targeting using JENNIFER embeddings in a real-world case study region (e.g., Hungary, Denmark, or Poland, identified as high-priority in [4] but currently under-explored).

1.3 Ethical issues

A major issue across large scale AI models such as JENNIFER is their alignment. This means that do the models make decisions using reasoning connected to the data and systems under study or do they "cheat", i.e., find local minima that produce convincing results but for unrelated reasons. In this project, a core component of the work is mitigating this risk by developing an understanding of what embeddings are used for what purposes across the project. This overall reduces the ethical risk of developing models that only spuriously provide answers to research questions.

1.4 Novelty and ambition

The tens of petabytes of data cumulative residing in diverse databases like I Currently, there are no examples of multi-modal subsurface foundation models trained on heterogeneous, real-world data spanning multiple measurement modalities, nor are there any that do uncertainty quantification, identified as a must in [47], or ones that are focused on calculating useful latent embedding spaces, two features which JENNIFER will actively implement. The risk to develop this model is low. Most of the neural network techniques exist to develop such a model and can be used directly from libraries such as `pytorch`. What does not exist is a centralized database nor the computing infrastructure that will be specific to the earth science data here. Thus, this proposal will produce the first foundation model of the earth's crust and the first ERA5-like global dataset. While data harmonization at this scale presents engineering challenges, the PI's extensive experience with large-scale data pipelines substantially reduces this risk.

2 Impact

2.1 Potential for academic impact of the research project

The JENNIFER model has a strong potential to have a large impact on the Earth Science communities. While recent FRIPRO projects apply ML to specific subsurface challenges (e.g., the PI's current SerpRateAI project), this proposal addresses a core issue in subsurface geoscience: a lack of reusable multimodal representations. JENNIFER's joint embeddings will encode cross-modal physical relationships that can be leveraged to build more accurate subsurface maps of hydrogen containing ultramafic formations. This proposal will build stronger links between the research communities of Earth Science and AI. This proposal will provide a novel multimodal database and a foundation model of the Earth that is able to map and identify hydrogen locations. We will compile this data as an open source database following FAIR principles. This will also allow this project to have impact on the new ICDP and IODP missions which are currently in planning stages as it can help shape the future of data management for such projects.

We will engage the community in the following ways. First, we have partnered with ExploreTech, a Silicon Valley-based startup that produces AI technology for mining applications. Second, both Utrecht University (partner: Charlie Beard) and UiO (host institution) have excellent applied data science and computational science programs where we will recruit Master students. We will organize a small conference (50+ participants) that is focused on the goals and developments of foundation models for subsurface Earth Science. The model's dual focus on reducing environmental impacts while improving assessment precision creates measurable pathways toward SDG 7 (Affordable and Clean Energy), SDG 9 (Industry, Innovation, and Infrastructure), SDG 12 (Responsible Production), and SDG 13 (Climate Action).

2.2 Measures for communication and exploitation

We plan to disseminate our results to the scientific communities through peer-reviewed articles, participation in international conferences, and organization of an international conference on AI and earth system dynamics. We plan to write four papers representing both novel results and novel methodology generated from this project. These papers will be submitted to journals such as the Journal of Geophysical Research - Solid Earth and leading AI venues such as the main track NeurIPS conference proceedings. In addition to the anticipated publications and workshops, this project will produce large, normalized, and easy-to-use data sets and an AI model. The PI has extensive experience producing public data sets for less-technical users to analyse [36]. The database will be available to access through the NIRD data management system, and model weights will be accessible to the public by posting on HuggingFace, the leading AI model repository.

3 Implementation

3.1 Project manager and project group

Project Manager: Dr. John M. Aiken: Dr. Aiken's research background has focused on integrating AI within the problem spaces of CO_2 sequestration and subsurface geophysics, rock physics, glaciology, and previously educational data mining, with a strong track record building large-scale, multi-modal databases for AI models. Dr. Aiken was previously the PI of the RCN funded project SerpRateAI.
Project Team: The project team consists of **Dr. Thorsten Becker** (University of Texas at Austin, Institute for Geophysics (UTIG)), **Dr. Dunyu Liu** (UTIG), **Dr. William Gilpin** (UT, Department of Physics), and **Dr. Charlie Beard** (Utrecht University, Department of Geosciences), plus a postdoc with a strong computing background to be hired. This highly interdisciplinary team covers geophysics, hydrogen, critical minerals, and AI/ML, essential for a cross-domain foundation model.

3.2 Project organisation and management

This project will connect four internationally recognized research centers creating a strong connection using AI. The project team will meet weekly for project updates, while the PI and postdoc will meet more frequently as they co-develop the JENNIFER model. A core component of this work will be the support of the Texas Advanced Computing Center (TACC) which sits next to UTIG. TACC has recently built the new GPU supercomputer, Horizon, which has 4000 Nvidia GPUs. JENNIFER will require significant compute resources to test and train, a sum of hours that will overwhelm most computer resources in Norway. This will greatly leverage both these GPUs and the interconnects that will allow it to access all of the data. This will also strengthen the relationship between UTIG at UT-Austin and Njord Centre at UiO. Compute resources for training will be secured through complementary pathways. Collaborator Becker will apply for a dedicated allocation on the TACC Horizon supercomputer, and the team will pursue an NVIDIA Academic Grant (Call for Proposals: Simulation and Modeling) providing up to 30,000 H100 GPU hours.

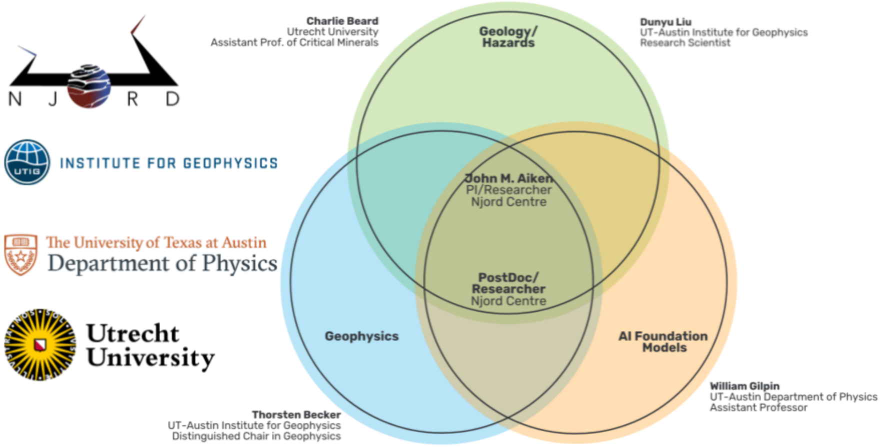

Figure 4: Project team expertise and institutional partnerships.

References

[1] Geoffrey S. Ellis and Sarah E. Gelman. In: *Sci. Adv.* 10.50 (2024), eado0955. DOI: 10.1126/sciadv.ado0955. URL: https://doi.org/10.1126/sciadv.ado0955.

[2] Sarah E. Gelman et al. USGS Professional Paper 1900. Version 1.2, January 22, 2025. U.S. Geological Survey, 2025. DOI: 10.3133/pp1900. URL: https://doi.org/10.3133/pp1900.

[3] Victor Zgonnik. In: *Earth-Sci. Rev.* 203 (2020), p. 103140. DOI: 10.1016/j.earscirev.2020.103140. URL: https://doi.org/10.1016/j.earscirev.2020.103140.

[4] Florian H. J. Willemsen and Johannes M. Miocic. Preprint, EarthArXiv. 2025. URL: https://eartharxiv.org/.

[5] Daniele Bailo et al. In: *Ann. Geophys.* 66.2 (2023), R0210. DOI: 10.4401/ag-9050.

[6] Norwegian Offshore Directorate. [DATASET]. 2023. URL: https://www.sodir.no/en/diskos/.

[7] Laurel B Childress et al. In: *Geochem. Geophys. Geosyst.* 25.2 (2024), e2023GC011287.

[8] Shaowen Mao et al. In: *Energy Environ. Sci.* 18 (2025), pp. 9991–10035. DOI: 10.1039/D5EE02910D.

[9] Peter B. Kelemen and Jürg Matter. In: *Proc. Natl. Acad. Sci.* 105.45 (2008), pp. 17295–17300. DOI: 10.1073/pnas.0805794105.

[10] Quanyou Liu et al. In: *Sci. Adv.* 11 (2025), eadr6771. DOI: 10.1126/sciadv.adr6771.

[11] Isabelle Moretti et al. In: *Int. J. Hydrogen Energy* 46.5 (2021), pp. 3615–3628. DOI: 10.1016/j.ijhydene.2020.11.026.

[12] Rishi Bommasani et al. 2022. arXiv: 2108.07258 [cs.LG].

[13] Minyoung Huh et al. 2024. arXiv: 2405.07987 [cs.LG].

[14] Rishi Jha et al. 2025. arXiv: 2505.12540 [cs.LG].

[15] Prakhar Kaushik et al. 2025. arXiv: 2512.05117 [cs.LG].

[16] C. F. Brown et al. Preprint arXiv:2507.22291. 2025. DOI: 10.48550/arXiv.2507.22291.

[17] John Mern and Jef Caers. In: *Geosci. Model Dev. Discuss.* 2022 (2022), pp. 1–50.

[18] John Mern et al. 2024. arXiv: 2410.10610 [cs.AI].

[19] Mahmoud Assran et al. 2023. arXiv: 2301.08243 [cs.CV].

[20] Jacob Devlin et al. 2019. arXiv: 1810.04805 [cs.CL].

[21] Alexandre Lacoste et al. In: *Adv. Neural Inf. Process. Syst.* 36 (2023), pp. 51080–51093.

[22] Hanlin Sheng et al. In: *Geophysics* 90.2 (2025), pp. IM59–IM79.

[23] Daniel Bolya et al. 2025. arXiv: 2504.13181 [cs.CV].

[24] Ian Goodfellow et al. http://www.deeplearningbook.org. MIT Press, 2016.

[25] Ashish Vaswani et al. 2023. arXiv: 1706.03762 [cs.CL].

[26] S. Mostafa Mousavi et al. In: *Nature Comm.* 11 (2020), p. 3952.

[27] Cristian Bodnar et al. In: *arXiv preprint arXiv:2405.13063* (2024).

[28] Margaret A. Goldman et al. In: *Appl. Comput. Geosci.* 27 (2025), p. 100274. ISSN: 2590-1974. DOI: https://doi.org/10.1016/j.acags.2025.100274.

[29] William Davis. URL: https://github.com/williamjsdavis/geo-lm.

[30] Francesco Regazzoni et al. In: *Nature Comm.* 15 (2024), p. 1834.

[31] Kangming Li et al. In: *Commun. Mater.* 6.1 (2025), p. 9. DOI: 10.1038/s43246-024-00731-w.

[32] Alec Radford et al. 2021. arXiv: 2103.00020 [cs.CV].

[33] Hanyue Li and Chao Shi. In: *Geodata and AI* 2 (2025), p. 100010. ISSN: 3050-483X. DOI: https://doi.org/10.1016/j.geoai.2025.100010.

[34] Jack Muir et al. 2024. arXiv: 2403.15095 [physics.geo-ph].

[35] P. J. Triezenberg et al. [DATASET]. 2016. DOI: 10.5066/F7930R7P.

[36] John M Aiken et al. In: *J. Geophys. Res.: Mach. Learn. Comput.* 2.2 (2025), e2025JH000666.

[37] D. P. Chambers. Version 5.0. Dataset accessed [05-02-2026]. PO.DAAC, CA, USA, 2012. DOI: 10.5067/TEOCN-0N005.

[38] Brian Meyer et al. Version 3. 05-02-2026. NOAA National Centers for Environmental Information, 2017. DOI: 10.7289/V5H70CVX.

[39] M. MacFerrin et al. In: *Earth System Science Data* 17.5 (2025), pp. 1835–1849. DOI: 10.5194/essd-17-1835-2025.

[40] Michael E. Pasyanos et al. In: *J. Geophys. Res.: Solid Earth* 119.3 (2014), pp. 2153–2173. DOI: https://doi.org/10.1002/2013JB010626.

[41] Christoph Feichtenhofer et al. In: *Adv. Neural Inf. Process. Syst.* Ed. by S. Koyejo et al. Vol. 35. 2022, pp. 35946–35958. URL: http://bit.ly/40PPiB7.

[42] E. Goh et al. In: *Ocean Science* 20.5 (2024), pp. 1309–1323. DOI: 10.5194/os-20-1309-2024.

[43] John M Aiken et al. Tech. rep. Copernicus Meetings, 2026.

[44] Mahmoud Assran et al. In: *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. June 2023, pp. 15619–15629.

[45] Jeongwoo Shin et al. In: *Adv. Neural Inf. Process. Syst.* Ed. by A. Globerson et al. Vol. 37. 2024, pp. 58929–58954. DOI: 10.52202/079017-1879.

[46] Giuseppe Etiope. In: *Int. J. Hydrog. Energy* 78 (2024), pp. 368–372. ISSN: 0360-3199. DOI: https://doi.org/10.1016/j.ijhydene.2024.06.292.

[47] Engineering National Academies of Sciences and Medicine. Washington, DC: The National Academies Press, 2025. ISBN: 978-0-309-99500-9. DOI: 10.17226/29212.