PROCEEDINGS, 51st Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 9-11, 2026
SGP-TR-230

Rapid Simulation of Aquifer Thermal Energy Storage Using Adaptive Physics Transformer

Hadrian Fung[1], Issac Ju[2, 3], Carl Jacquemyn[1], Meissam Bahlali[1], Gege Wen[1, 3] and Matthew Jackson[1]

[1] Department of Earth Science and Engineering, Imperial College London

[2] Department of Energy Science and Engineering, Stanford University

[3] EarthFlow AI

Keywords: Aquifer thermal energy storage, Machine learning, Adaptive mesh, Numerical modeling

ABSTRACT

Aquifer Thermal Energy Storage (ATES) offers sustainable, low-carbon heating and cooling to the built environment. Optimising the design and operation of ATES installations requires numerical simulation of groundwater flow and heat transport in heterogeneous aquifers. These simulations are typically computationally expensive. Machine Learning (ML) offers a rapid alternative to conventional numerical simulation of complex subsurface flow and transport processes. Here, we introduce the Adaptive Physics Transformer, a transformer-based ML approach, implemented purely data-driven, to significantly increase simulation efficiency whilst retaining accuracy. The APT simulation alternative is trained on outputs from our in-house Imperial College Finite Element Reservoir Simulator (IC-FERST), an advanced code that employs dynamic mesh optimization to achieve high solution accuracy at lower computational cost. The practical consequence is that the mesh varies across solution snapshots used for training. Conventional Convolutional Neural Network (CNN) or Neural Operator-based models often require a fixed mesh. Instead, APT can natively learn from the dynamic mesh simulation. Our results suggest a promising approach to rapid ATES simulation, reducing simulation times from tens of hours to a few minutes.

1. INTRODUCTION

Aquifer Thermal Energy Storage (ATES) offers sustainable, low-carbon heating and cooling to the built environment (Jackson et al., 2024). Optimising the design and operation of ATES installations requires numerical simulation of groundwater flow and heat transport in heterogeneous aquifers. These simulations are typically computationally expensive: high spatial resolution is required to accurately resolve pressure, flow, and temperature fields; moreover, high temporal resolution may be necessary to mitigate numerical diffusion and/or resolve rapid changes in injection flow rate and temperature. Simulations of systems that use multiple boreholes, or that require capturing interactions between neighbouring systems, are particularly challenging. Multiple simulations may be required to quantify the impact of aquifer heterogeneity uncertainty. Yet, the time available for aquifer modelling in many commercial projects is very limited. Rapid yet accurate approaches for simulating subsurface flow and heat transport in ATES and other shallow geothermal deployments are urgently required.

With the goal of delivering rapid but accurate simulations of complex subsurface flow and transport processes, here we introduce a machine learning (ML) model framework on a pure data-driven basis, which significantly increases simulation efficiency while retaining accuracy. The ML-based simulation alternative is trained on outputs from our in-house simulator (the Imperial College Finite Element Reservoir Simulator (IC-FERST, available at https://multifluids.github.io; Salinas et al. 2021; Regnier et al., 2022), an advanced code that solves the complex subsurface flow and transport processes on an unstructured tetrahedral mesh with dynamic mesh optimization (DMO). The simulator delivers numerical solutions with higher accuracy than conventional approaches that use fixed meshes or grids, at a comparatively manageable computational cost.

A practical consequence of DMO is that the mesh varies across the solution snapshots used for training. Conventional models based on Convolutional Neural Networks (CNN) and Graph Neural Networks (GNN) require a fixed input and output mesh. No existing ML model architecture can be applied directly to an unstructured, adaptive mesh.

Here, we introduce the Adaptive Physics Transformer (APT) as an ML architecture that learns the underlying physics directly from unstructured, adaptive-mesh training data. The novelty of APT as an ML-based alternative lies in (1) significant speed-up in modeling time while retaining high accuracy on the temperature and pressure prediction; (2) flexibility to train on any unstructured, adaptive mesh data, and (3) transformer architecture that enables parallel training and inferencing with hardware computational optimizations. With the unique strengths APT could bring to ATES simulation, it unlocks further possibilities in the workflow for optimising ATES design, including Monte Carlo sensitivity analysis to constrain the impact of uncertainties, in which multiple simulations can be run in parallel, yielding results for an optimised ATES setup.

2. BACKGROUND AND RELATED WORK

2.1 Numerical Simulation by IC-FERST

Conventional numerical simulations employ fixed grids or meshes, which means that high mesh resolution may be imposed where it is not required to capture rapid changes in pressure, velocity, temperature, or other solution metrics of interest; conversely, high-resolution

solution features may propagate into regions with coarse mesh and fail to be accurately captured. DMO, as implemented in IC-FERST, removes these issues by adapting the mesh between timesteps to resolve solution fields of interest (Salinas et al., 2021). However, despite the use of DMO, each simulation in the training dataset used here required approximately 7 h to complete.

2.2 Data-Driven Simulation Alternatives

Approaches to ML-based alternatives for coupled systems include graph neural networks (GNNs) on fixed meshes and convolutional or Fourier-based neural operators on structured grids. While these models can accelerate inference, they require either mapping adaptive mesh data to a reference grid, which incurs interpolation error and overhead, or retraining for each mesh configuration. Extensions to point-cloud neural operators exist, but they scale poorly in terms of time complexity $O(N^2)$ and lack mechanisms to exploit mesh adaptivity.

2.3 Transformer-based Machine Learning Model Architecture in Physical Modelling

Physics transformer models, with their attention mechanisms, offer O(N) variants (e.g., linear attention) that can handle large point sets (Alkin et al., 2024; Wu et al., 2024). However, to the best of our knowledge, no existing models have been applied to unstructured adaptive meshes with direct querying of arbitrary output points. We demonstrate that the APT can tokenize mesh nodes and query points and uses sparse attention projections to encode spatial structure and temporal dynamics, thereby enabling direct training on intrinsically adaptive mesh data.

3. METHODOLOGY

3.1 Numerical Simulation by IC-FERST

Our ML model is trained on a synthetic dataset of 840 different scenarios. The key parameters varied across scenarios in the synthetic dataset include well spacing, well depth (warm and cold), screen length, and vertical permeability (Figure 1). Throughout the one-step training of the ML model, the temperature of the initial timestep, well injection rate, well injection temperature, and vertical permeability are the key input features of node attributes to feed into the ML model; the output is the temperature field of the target timestep on the output mesh.

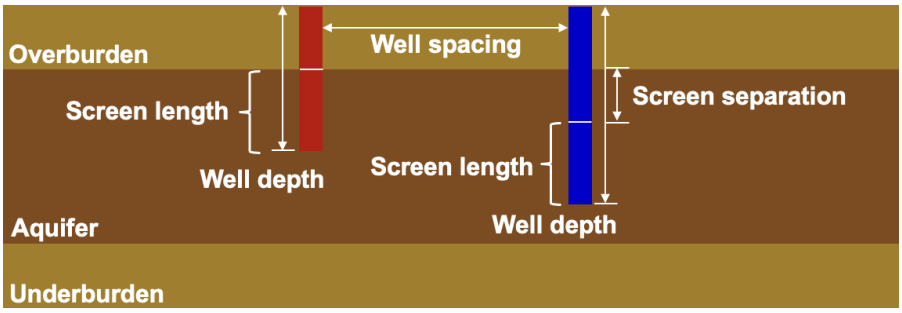

Figure 1: Schematic showing the parameters varied in the training dataset simulations following a simple Monte-Carlo sampling strategy.

The injection and extraction cycle consists of 8 timesteps of cold well injection for waste cool/warm well extraction for heating during winter, followed by 4 timesteps of resting period, then 8 timesteps of warm well injection for waste cool/cold well extraction for cooling during summer; followed by another 4 timesteps resting period (Figure 2).

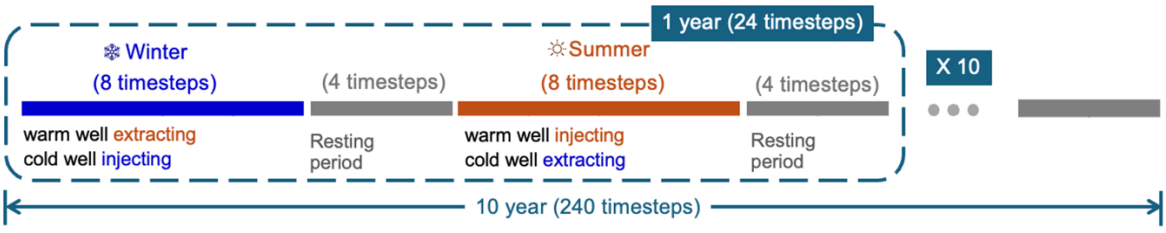

Figure 2: Schematic for the simulated operation of the ATES system.

We simulate cyclic seasonal operation over 10 years with 240 timesteps (2 timesteps per month). Each annual cycle consists of winter operation (cold well injection / warm well extraction), transitional resting periods (no injection/extraction in both wells), summer operation (warm well injection / cold well extraction), and another resting period.

The synthetic dataset contains 840 scenarios with varied well configurations and vertical heterogeneity. Key scenario variables are sampled within the following ranges: well spacing [12.5, 500] m; well depth (warm/cold) [−150, −60] m; screen length [10, 100] m;

screen separation [0, 90] m; and vertical permeability $K_z \in [1, 1000]$ (units as in the simulator setup). We split scenarios into 80% training, 10% validation, and 10% test.

3.2 Adaptive Physics Transformer (APT) Architecture

Our Adaptive Physics Transformer (APT) is a transformer-based autoregressive model comprising an encoder, a transformer, and a decoder block, as shown in Figure 3.

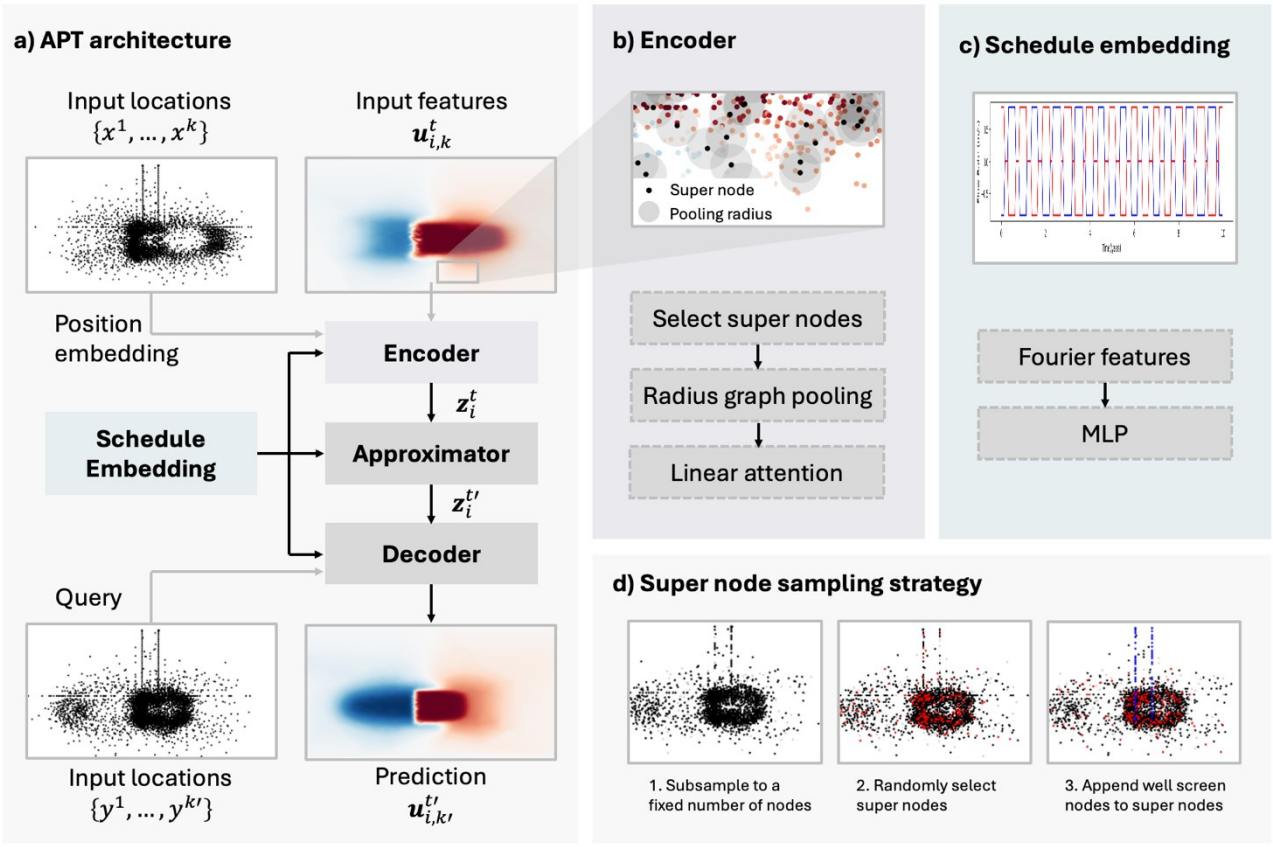

Figure 3: Schematic of the Adaptive Physics Transformer (APT) architecture, showing the encoder, transformer, and decoder blocks with cross-linear attention mechanisms.

Encoder

The encoder block accepts input features as node attributes, including the temperature at the current time step, well injection rate, well injection temperature, and vertical permeability. The encoder block then iteratively, over training epochs, randomly selects discrete, independent sets of supernodes from the mesh's input nodes, allowing the model to gradually learn and generalize the underlying continuous field from the discretized input mesh. Throughout super-node selection, the randomness of the process yields the same probabilistic distribution when the model selects a node from the mesh, regardless of the density of nodes across the mesh's localities, thereby efficiently compressing spatial information into a small tokenized latent representation.

Throughout the encoder, the model learns the mapping between the input features and the coordinates of input mesh by cross linear attention instead of conventional attention block being used in other current transformer model, which traditional attention compares every token with every other token—taking quadratic time dependence $O(N^2)$—whereas linear attention uses kernel or low-rank approximations to process each token in a single pass in each locality of the super-node, reducing complexity to linear time complexity $O(N)$. Cross linear attention maps the input feature and the input mesh coordinates jointly and preserves all spatial information in the output of the encoder block.

Transformer

The latent representation from the encoder block is then passed to the transformer block, where the stack of transformer blocks learns the temporal progression, mapping from the current timestep to the next in the latent space efficiently.

Decoder

The next latent representation is then recovered into the physical space of the reservoir by the decoder. The decoder block of the model takes a user-defined output query mesh, which is independent of the input mesh and is flexible in the number of nodes and the spatial distribution of the output mesh. The decoder block learns to query the latent representation from the previous transformer block with exact temperature prediction directly at the location of each node by the cross-linear attention mechanism, leveraging the full flexibility of the adaptive mesh on our synthetic dataset from IC-FERST with dynamic mesh optimization (DMO), without the necessity to perform interpolation between fixed mesh and adaptive mesh. The decoder block will yield the temperature at the next time step on the output mesh, which is readily available for the next pass of the ML model in an autoregressive manner.

4. MODEL RESULTS

4.1 Result visualisations

Figure 4 compares 3 scenarios and presents cross-sections of the output at several selected timesteps.

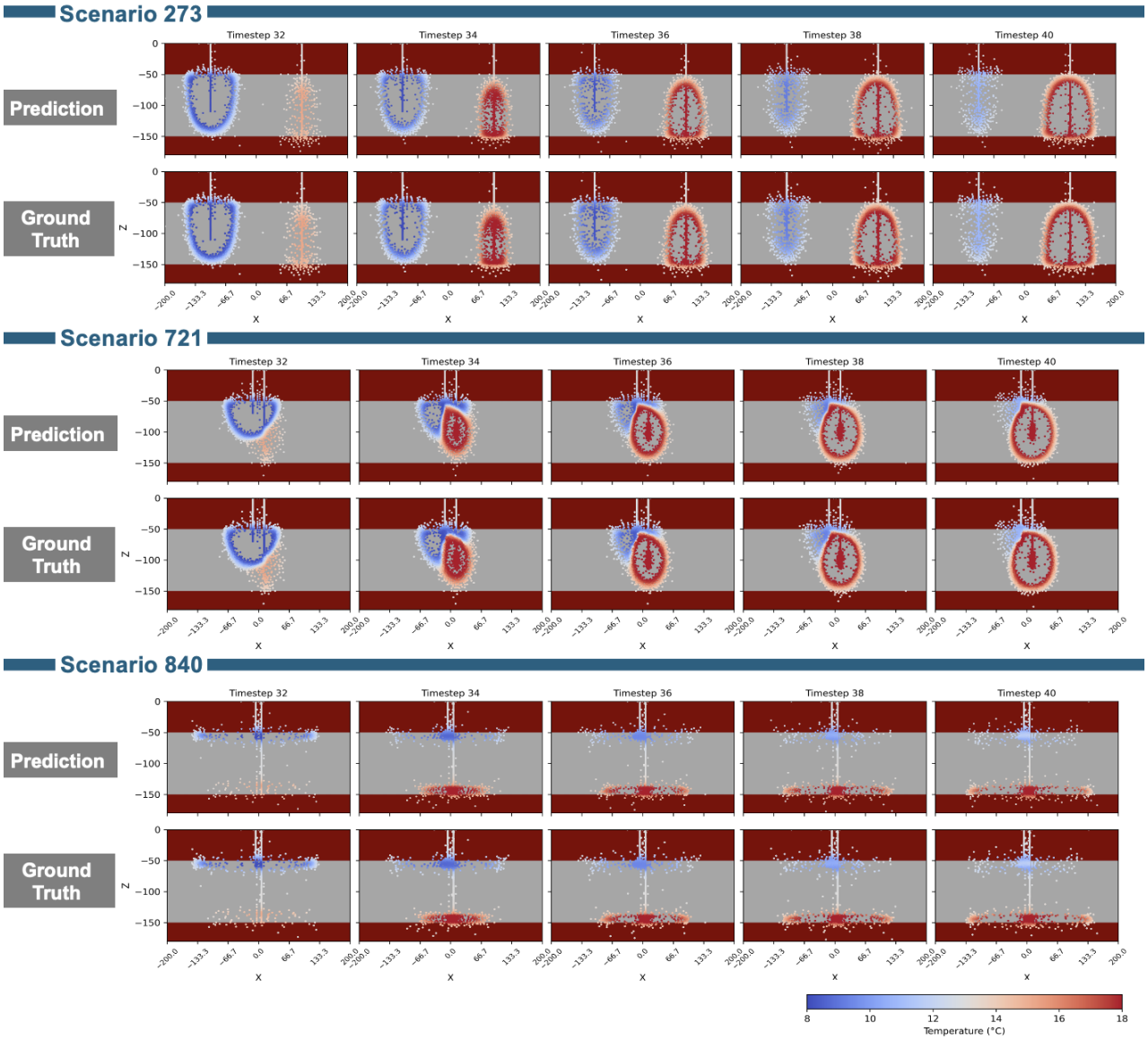

Figure 4: Comparison plots showing XZ cross-sections from 3 different scenarios at selected timesteps. Upper row: ML prediction. Lower row: Numerical Simulation Ground Truth.

The model can capture the plume expansion pattern and the exact temperatures at the well and at the plume fringe. As no interpolation is needed for the APT architecture, a direct comparison can be made between the output prediction from our ML model and the ground truth data from the numerical simulation of IC-FERST.

The visualization shows that the model is not only able to capture the alternating pattern in different injecting/extracting phases but also illustrates the capability to generalize the pattern in different ATES systems with different well locations, screen depth, and spacing. Some of the simulations show thermal breakthroughs and thermal interference, in which the warm plume and cold plume interact with each other (Scenario 840). These simulations are often harder for the ML model to generalize due to the abrupt high-temperature gradient at the contact front of the plumes. The visualization results show that the APT architecture can capture the short-circuited pattern accurately as well.

4.2 Model Performance and Error Metrics

Table 1: Error metrics

	R2 score	Rel L2
Train set	99.2 %	0.008
Test set	99.0 %	0.011

Table 1 shows the error metrics of the training dataset and the unseen test dataset of the model, including the R^2 score and Relative L2 error across the whole dataset.

Consistently high R2 scores on both the training and test data indicate strong overall generalization without overfitting, with low error across all scenarios. APT architecture is capable of capturing the heat plume contraction-expansion pattern under constant injection-extraction cycles, performing particularly well at later timesteps (generally after the 3rd year/3rd cycle).

4.3 Inferencing speed

The average inference model run time of the APT architecture for a single full scenario rollout of 240 timesteps (10 years of simulation time) is under 5 seconds upon training, while the average runtime of numerical simulation with DMO on IC-FERST takes up to 7 hours (Figure 5).

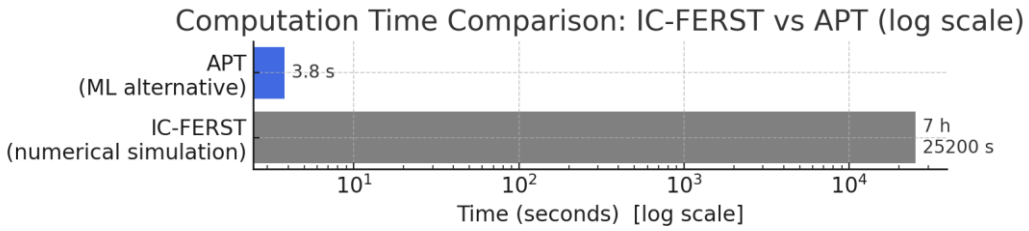

Figure 5: Graph showing the computational time comparison between the average run time of IC-FERST and APT.

5. DISCUSSION

5.1 Model Strength: Mesh Adaptivity

The novelty of the APT architecture for rapid ATES simulation focuses on its flexibility to allow variable input and output formats on an adaptive mesh. Conventional ML-based models strictly require a fixed input and output shape and format to be constant across the dataset, while APT uses linear cross attention to map between input features and input mesh coordinates, allowing variable input shape and number of nodes across the training dataset and subsequent usage, as linear attention does not rely on a global Query-Key-Value (QKV) matching. The same advantage applies to the output mesh as well, which naturally allows extra flexibility for adaptive mesh natively.

As no interpolation is required for the adaptive mesh to be adopted, the error comes purely from the model prediction, which the ML model could refine and learn the underlying physics much more accurately without the interpolation error involved in the representation of ground truth label throughout the training.

5.2 Model Strength: Efficient Latent Representation

As temporal progression is done on the small latent representation (token length) instead of the entire physical space of the full reservoir, the efficient latent representation allows a faster rollout in prediction.

Furthermore, cross-linear attention is efficiently done on each locality of the super-node instead of a global attention field of the whole mesh. This allows extra acceleration to the rollout in prediction when comparing with numerical simulations.

6. CONCLUSIONS AND FUTURE WORK

We introduced the Adaptive Physics Transformer (APT) as an ML alternative to numerical simulations for Aquifer Thermal Energy Storage (ATES), significantly accelerating the simulation by approximately 6600x while retaining high accuracy, with an R^2 greater than 0.99 on both training and test datasets.

The key innovations of APT include: (1) the ability to natively learn from unstructured, adaptive mesh data generated by IC-FERST with dynamic mesh optimization, eliminating the need for interpolation between fixed and adaptive meshes; (2) the use of cross-linear attention mechanisms that achieve O(N) time complexity, enabling efficient training and inference on large point sets; and (3) a flexible encoder-transformer-decoder architecture that allows variable input and output mesh configurations.

Our results demonstrate that APT can accurately capture complex thermal plume dynamics, including expansion-contraction patterns during injection-extraction cycles, thermal breakthroughs, and thermal interference between warm and cold plumes. The model generalizes well across diverse ATES configurations with varying well locations, screen depths, and spacing parameters.

The dramatic reduction in simulation time from approximately 7 hours to under 5 seconds opens new possibilities for ATES design optimization workflows. Potential applications include Monte Carlo sensitivity analysis to quantify the impact of aquifer heterogeneity uncertainty, real-time operational optimization, and rapid screening of multiple design alternatives.

Future work will focus on extending the APT framework to handle more complex multi-well ATES systems, incorporating additional physical processes such as geochemical reactions, and validating the approach against field data from operational ATES installations. We also plan to explore transfer learning strategies to enable rapid adaptation of pre-trained models to new geological settings with limited training data.

REFERENCES

Alkin B., Fürst A., Schmid S., Gruber L., Holzleitner M. & Brandstetter J. (2024) Universal Physics Transformers: A Framework for Efficiently Scaling Neural Operators. arXiv preprint arXiv:2402.12365. 2402.12365v5.pdf]

Regnier G., Salinas P., Jacquemyn C. and Jackson M.D.: Numerical simulation of aquifer thermal energy storage using surface-based geologic modelling and dynamic mesh optimisation. Hydrogeology Journal (2022) 30, 1179–98. http://dx.doi.org/10.1007/s10040-022-02481-w.

Jackson, M. D., Regnier, G., and Staffell, I. Aquifer thermal energy storage for low carbon heating and cooling in the united kingdom: Current status and future prospects. Applied energy, 376:124096, 2024.

Salinas, P., Regnier, G., Jacquemyn, C., Pain C.C. and Jackson, M.D., Dynamic mesh optimisation for geothermal reservoir modelling, Geothermics (2021) 94, 10289. https://doi.org/10.1016/j.geothermics.2021.102089.

Wu, H., Luo, H., Wang, H., Wang, J., and Long, M. Transolver: A fast transformer solver for pdes on general geometries. arXiv preprint arXiv:2402.02366, 2024.