Advances in Geo-Energy Research

Vol. 21, No. 1, p. 13-28, 2026

Original article

Physics-guided self-supervised upscaling of three-dimensional digital rock models from multi-resolution micro-CT data

Batyrkhan Gainitdinov[1]*, Rinat Prochii[1], Denis Orlov[1], Maxim Sharaev[1,2], Yili Ren[3], Dmitry Koroteev[1]

[1]*Skolkovo Institute of Science and Technology, Moscow 121205, Russia*

[2]*Biomedically Informed Artificial Intelligence Laboratory, University of Sharjah, Sharjah 27272, United Arab Emirates*

[3]*PetroChina Research Institute of Petroleum Exploration Development, Beijing 100083, P. R. China*

Keywords:
Self-supervised learning
upscaling
digital rock physics
reservoir characterization
darcy-scale rock models

Cited as:
Gainitdinov, B., Prochii, R., Orlov, D., Sharaev, M., Ren, Y., Koroteev, D. Physics-guided self-supervised upscaling of three-dimensional digital rock models from multi-resolution micro-CT data. Advances in Geo-Energy Research, 2026, 21(1): 13-28.
https://doi.org/10.46690/ager.2026.07.04

Abstract:
Digital Rock Physics relies on multiscale upscaling workflows that bridge pore-scale imaging and Darcy-scale flow simulation in heterogeneous, low-permeability reservoirs. Existing approaches use convolutional neural networks to transform low-resolution micro-computed tomography images into multiclass Darcy-scale models, but depend on supervised training with carefully curated labeled minicubes, which is costly and difficult to extend to new lithologies or scarce-label regimes. This work introduces a physics-guided self-supervised pretraining framework for three-dimensional digital rock models that combines volumetric contrastive learning with a permeability-aware regularization term. The encoder first learns volumetric representations by predicting contextual positions of three-dimensional image crops and then enforces consistency between embedding similarities and proxy permeabilities derived from percolation-based analysis of segmented pore space. After fine-tuning for rock typing, the encoder is integrated into an upscaling pipeline that maps low-resolution scans to Darcy-scale multiclass models used for single-phase flow simulations. The self-supervised model was compared with a purely supervised model with a purely supervised baseline in terms of rock-typing performance, visual fidelity of the upscaled models, and preservation of key petrophysical properties relative to laboratory and high-resolution numerical benchmarks. The results indicate that physics-guided self-supervised pretraining improves rock-typing accuracy, yields Darcy-scale models with more consistent connectivity of high-permeability channels and barriers, and reduces discrepancies in effective permeability, especially in low-label regimes. These findings suggest that self-supervised, physics-informed representation learning can enhance both classification robustness and the reliability of digital rock upscaling workflows for heterogeneous carbonate rocks.

1. Introduction

The accurate characterization of porous media remains a central challenge in geoscience and reservoir engineering, as macroscopic rock properties such as permeability, relative permeability, and elastic moduli are governed by the intricate microstructure of pores and fractures. This challenge extends beyond hydrocarbon recovery to encompass critical applications including groundwater management, geothermal energy production, underground hydrogen storage, and geological carbon sequestration, where an accurate understanding of pore-scale processes is essential for predicting large-scale behavior (Blunt et al., 2013, 2025; Spurin et al., 2025).

Yandy Scientific Press

*Corresponding author.
E-mail address: Batyrkhan.Gainitdinov@skoltech.ru (B. Gainitdinov); Rinat.Prochii@skoltech.ru (R. Prochii); D.Orlov@skoltech.ru (D. Orlov); M.Sharaev@skoltech.ru (M. Sharaev); renyili@petrochina.com.cn (Y. Ren); D.Koroteev@skoltech.ru (D. Koroteev).

Received April 20, 2026; revised May 19, 2026; accepted June 12, 2026; available online June 15, 2026.

Traditional laboratory methods for probing pore networks, including scanning electron microscopy (SEM), mercury intrusion porosimetry (MIP), and gas adsorption, provide valuable complementary information but suffer from well-known limitations such as destructive sample preparation, restriction to two-dimensional analysis, and inability to characterize three-dimensional pore connectivity (Blunt et al., 2013). Digital rock physics (DRP) addresses these challenges by integrating X-ray micro-computed tomography (μCT) with numerical simulations to infer macroscopic properties from pore-scale structure, enabling non-destructive "virtual experiments" (Blunt et al., 2013). However, even with μCT imaging, segmentation and preprocessing have decisive impact on porosity and permeability estimates and must be validated against laboratory measurements (Vidal et al., 2024).

Unfortunately, the capabilities of digital core technologies for flow simulation are currently limited to highly permeable reservoirs with relatively simple structures (Orlov et al., 2021). For low-permeable reservoir rocks, such as tight gas reservoirs in the United States and Canada formed in the Paleozoic era (Law et al., 1993), the formation of an organic-rich carbonate Tlyanchy-Tamakian of an upper Devonian Frasnian age in the European part of Russia (Mukhametdinova et al., 2020), or tight gas reservoirs in Chinese basins, which are mostly associated with coal strata (Zou et al., 2012), the use of DRP is not as straightforward as for high porosity sandstones. Natural geological formations consisting mainly of low-permeability sandstone or carbonate (permeability of < 0.1 mD) are the subject of many engineering projects, such as the development of oil and gas reservoirs and coal mines (Zheng et al., 2015). Hence, a comprehensive investigation of the transport and petrophysical properties of low-permeability rocks is necessary to ensure the efficiency and productivity of engineering projects in this type of geological environment.

A fundamental challenge in DRP is the resolution-volume trade-off: Sub-resolution features in large volumes significantly impact property predictions (Faisal et al., 2019; Behbahani et al., 2023; Zhou and Nie, 2024). The Pore Size Distribution of low permeable rocks is wide and characterized by pore sizes of several nanometers to hundreds micrometers (Mukhametdinova et al., 2020; Ebadi et al., 2022). The spatial resolution of μCT images is usually higher than 1 μm/voxel. This indicates that a significant portion of pores cannot be detected in the corresponding binary models, built during the conventional DRP procedure, thus there is no connected pathway for fluid flows calculations. Thus, for low-permeable rocks, the trade-off between spatial resolution of the data and representativeness of the model size becomes more important than for homogeneous or highly permeable rocks.

Three principal approaches address this limitation. Super-resolution techniques using deep learning (Jackson et al., 2022) enhance low-resolution μCT images, with methods like CinCGAN (Niu et al., 2020) enabling unpaired image translation. While effective for extending field of view, these approaches remain dependent on high-resolution training data. Dual-porosity modeling treats macro- and micro-porous regions as overlapping continua (Bultreys et al., 2015; Ruspini et al., 2021), capturing hierarchical heterogeneity but requiring extensive calibration. The upscaling approach explicitly resolves pores only at the finest available scale and represents smaller pores in an averaged Darcy-type manner at coarser scales (Faisal et al., 2019; Behbahani et al., 2023; Orlov et al., 2025). At the Darcy scale, sub-resolved porosity and permeability are encoded through additional continuum elements assigned to both pore and solid voxels, so that the effect of fine-scale heterogeneity is captured without explicitly resolving every pore. In contrast, dual-media models treat multiple pore systems explicitly, but quickly become computationally prohibitive when the scale separation between matrix and high-permeability features grows, which limits their applicability to strongly heterogeneous, low-permeability rocks.

In previous work on convolutional neural networks (CNN)-enhanced upscaling of digital cores, correlations between high-resolution and low-resolution μCT data were learned by partitioning both datasets into three-dimensional (3D) sub-volumes and automatically classifying each low-resolution fragment into a discrete digital rock type according to its petrophysical and transport properties (e.g., porosity and permeability) (Ebadi et al., 2022; Orlov et al., 2022). Once trained, the classifier maps a low-resolution scan into a Darcy-scale multi-class model in which each class is associated with pre-computed effective properties, enabling flow simulations on significantly coarsened grids while preserving the physical size of the core sample and its large-scale heterogeneity. This reduces memory and CPU requirements by orders of magnitude and allows massive digital cores to be used in routine flow modelling (Orlov et al., 2022).

However, this supervised CNN-based workflow critically depends on large, well-curated labelled datasets (Orlov et al., 2022; Behbahani et al., 2023). For each digital rock type, representative high-resolution fragments must be segmented, simulated (e.g., to obtain permeability tensors) and manually assigned to classes before training a classifier. This requires substantial expert time, extensive pore-scale simulations and careful balancing of classes, and it tightly couples the quality of the upscaled Darcy-scale model to the completeness and consistency of the training set. As a result, the supervised approach can be difficult to extend to new lithologies, to cores with limited high-resolution coverage, or to settings where only a small fraction of fragments can be reliably typed by experts.

To overcome these limitations of earlier supervised upscaling frameworks, the present study introduces a physics-guided self-supervised learning (SSL) approach that leverages large unlabeled μCT datasets and permeability-aware regularization to learn rock-typing representations before fine-tuning on a much smaller labelled set.

Importantly, this regularization is not purely data-driven: It embeds a known physical relationship between pore structure descriptors, such as porosity and connectivity, and permeability directly into the representation-learning objective. As a result, the encoder is guided not only by image statistics, but also by hydrodynamically meaningful similarity between rock fragments.

SSL has emerged as a powerful paradigm that addresses

these limitations by exploiting supervisory signals inherent in unlabeled data to learn transferable representations (Jing and Tian, 2020). Contrastive learning methods such as SimCLR and MoCo train encoders to maximize agreement between differently augmented views of the same instance while repelling representations of unrelated samples (Chen et al., 2020; He et al., 2020). This paradigm has proven particularly effective in low-label regimes, as representations learned from large unlabeled datasets can significantly boost performance on downstream tasks after fine-tuning with limited annotations (Azizi et al., 2021).

The advantages of SSL extend to 3D domains such as medical imaging, which faces similar challenges of annotation scarcity. Pre-training on unlabeled volumetric computed tomography and magnetic resonance imaging scans has yielded transferable 3D representations, significantly improving segmentation and classification performance with limited annotations (Zhou et al., 2019). Methods like Models Genesis introduced a unified self-supervised framework that learns from 3D medical images by restoring transformed image patches (Zhou et al., 2019), while VoCo proposed a volumetric contrastive learning method that leverages consistent contextual position priors in 3D medical images (Wu et al., 2024): A model is trained to predict the contextual position of randomly cropped sub-volumes by contrasting their features against a set of base crops that serve as position-dependent prototypes. This paradigm yields high-level semantic representations without requiring manual annotations and has been shown to outperform generic SSL baselines on a range of 3D medical tasks.

In DRP, initial steps toward SSL have only recently been reported. The research evaluated multiple SSL frameworks for 2D rock image classification, showing that self-supervised pre-training substantially improves performance compared to supervised baselines, particularly under severe label scarcity (Fourer et al., 2024). However, these studies remain limited to 2D rock image classification and do not address the unique challenges of 3D volumetric rock modeling and multiscale upscaling.

Despite recent progress in SSL for two-dimensional (2D) rock images and 3D medical volumes, to the best of our knowledge there is currently no framework that combines physics-guided SSL with multiscale upscaling in 3D digital rock models, nor a systematic comparison against supervised baselines in the Darcy-scale setting.

The present study alleviates the limitations of aforementioned approaches by introducing a novel physics-guided self-supervised pretraining stage based on a modified volumetric contrastive framework (Wu et al., 2024). During pretraining, the encoder is first optimized using a volume-contrast objective to predict contextual positions of 3D μCT crops, and then regularized with a permeability-aware loss that aligns similarities in the latent space with similarities in proxy permeability computed directly on binary masks. Concretely, each base crop is characterized by porosity and percolation-based proxy permeabilities along the principal axes, construct a similarity matrix in this proxy-permeability space, and penalize deviations between this matrix and the cosine-similarity matrix of the corresponding latent vectors. This encourages structurally and hydraulically similar sub-volumes to form clusters in the embedding space even before any supervised labels are used, and reduces the amount of labelled data required to train an accurate rock-type classifier.

The main objective of this study is to assess whether Darcy-scale digital core models built on top of self-supervised representations can match or outperform supervised baselines in terms of both classification quality and upscaled petrophysical properties. In this study, the focus is on a single, highly heterogeneous carbonate core imaged at two μCT resolutions (5 and 16.5 μm), which provides a challenging test for multiscale upscaling. Building on the previously established workflow, low-resolution data are coarsened by predicting digital rock types for non-overlapping 3D fragments, but here two variants of the classifier are compared: A purely supervised DenseNet-based model trained from scratch, and a classifier fine-tuned on top of a self-supervised VoCo encoder pre-trained with proxy-permeability-guided regularization (Orlov et al., 2025).

The scientific novelty also lies in qualitative and quantitative comparison of supervised and SSL approaches. Thus, validation is carried out along three complementary directions. First, the performance of the supervised and self-supervised classifiers is compared on the rock-typing task itself, using accuracy and macro-/micro-averaged F1-scores on labelled μCT mini-cubes. Second, Darcy-scale multi-class models are constructed from both classifiers and visually compare their spatial rock-type distributions with the original high-resolution μCT volume, focusing on the continuity of high-permeability channels, barriers and layered heterogeneities. Finally, single-phase absolute permeabilities are computed and porosities of the upscaled models and quantify their deviation from benchmark values obtained in the laboratory experiments, evaluating how the choice of representation learning (supervised vs. SSL with proxy-permeability loss) impacts the ability of Darcy-scale models to preserve key petrophysical properties. Additionally, both approaches are evaluated under severe label scarcity and on an independent 22 μm μCT sample that is excluded from both pretraining and fine-tuning, thereby probing their generalization to new resolutions and unseen volumes.

The results demosntrate that physics-guided self-supervised pretraining systematically improves rock-typing metrics, yields Darcy-scale models with more consistent representation of key heterogeneities, and reduces the discrepancy in effective permeability relative to reference values, particularly in low-label regimes. These findings indicate that incorporating self-supervised, physics-informed representation learning into Darcy-scale digital core workflows can enhance both classification robustness and the reliability of upscaled flow predictions in heterogeneous carbonate rocks.

2. Methodology

2.1 Upscaling problem formulation

The goal of the upscaling procedure is to construct a mapping $\mathscr{F}$ that converts a low-resolution 3D μCT volume into a Darcy-scale multi-class digital rock model. The input to $\mathscr{F}$ is a grayscale volume $\mathbf{I}_{LR} \in \mathbb{R}^{X \times Y \times Z}$ acquired at coarse

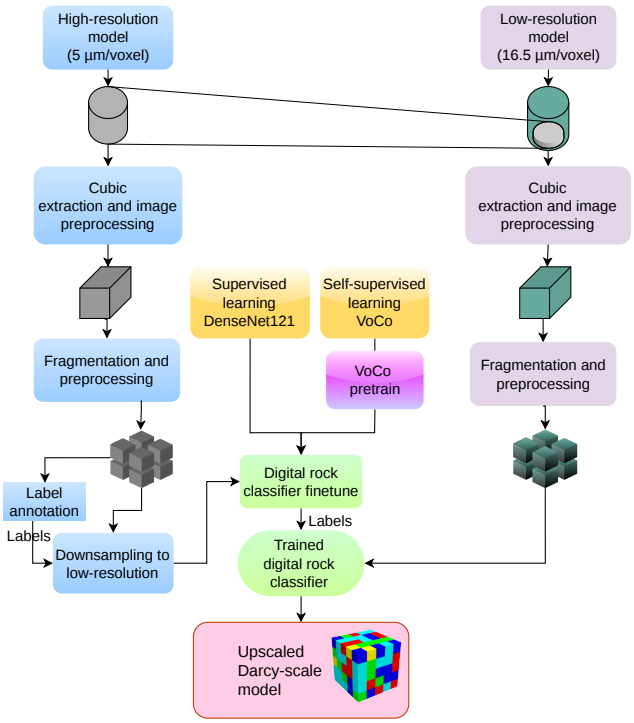

Fig. 1. The algorithm for constructing upscaled (multi-scale) digital model.

resolution (16.5 µm/voxel in this study). The output is a discrete-valued volume $\mathbf{M}_{Darcy} \in \{1,\ldots,K\}^{X'\times Y'\times Z'}$, where each voxel is assigned one of K predefined digital rock types. Every rock type k carries a set of effective petrophysical and transport properties (porosity ϕ_k, directional permeabilities k_k, etc.) pre-computed from high-resolution data. Because several input voxels are aggregated into a single Darcy-scale cell, the grid dimensions $(X',Y',Z') < (X,Y,Z)$, while the physical size of the model is preserved.

The mapping is realised by partitioning $\mathbf{I}_{LR}$ into non-overlapping 3D sub-volumes (minicubes) $\mathbf{p}^{(i)} \in \mathbb{R}^{L\times W\times H}$ and learning a classifier f_θ that predicts the rock-type label $\hat{y}^{(i)}$ for each sub-volume:

$$\hat{y}^{(i)} = f_\theta\left(\mathbf{p}^{(i)}\right), \qquad i = 1,\ldots,N_{\text{cubes}} \tag{1}$$

Here N_{cubes} denotes the total number of non-overlapping minicubes extracted from the low-resolution volume. The parameters θ are optimised on a labelled training set $\{(\tilde{\mathbf{p}}^{(j)},\ y^{(j)})\}_{j=1}^{N}$, in which $\tilde{\mathbf{p}}^{(j)}$ is a high-resolution minicube down-sampled to the low-resolution scale and $y^{(j)}$ is the rock-type label determined from the corresponding high-resolution binary model. The Darcy-scale model $\mathbf{M}_{Darcy}$ is then assembled by applying f_θ to every sub-volume in $\mathbf{I}_{LR}$ and mapping the predicted labels onto the coarsened grid.

The supervised baseline implements f_θ as a 3D DenseNet-121 (Hasan et al., 2021) trained from scratch using cross-entropy loss. This study additionally considers self-supervised alternative in which the encoder is first pre-trained on unlabelled minicubes using a modified physics-based volumetric contrastive framework (Section 3.1.2) and then fine-tuned on the same labelled set, yielding a second classifier $f_{\theta'}$. Both variants are evaluated within the identical upscaling pipeline described below.

Table 1. Averaged porosity and permeability values for each digital rock type.

Type	Fragments ratio (%)	Porosity (%)	Averaged permeability (mD)
1	63.0	4.5	0
2	11.8	12.2	0.109
3	11.9	18.3	0.903
4	10.3	31.1	6.625
5	2.9	59.8	55.81

2.2 Digital rock classifier pipeline

The classifier-based upscaling pipeline establishes a learned correlation between low-resolution µCT data and discrete digital rock types through a two-stage process: Encoder training followed by Darcy-scale model construction (Fig. 1). Two encoder training strategies are compared: Purely supervised DenseNet-121 and self-supervised SwinVT, within the identical pipeline structure.

2.2.1 Stage A: Classifier training (left contour in Fig. 1).

1) High-resolution (HR) µCT data (e.g., 5 µm/voxel) are acquired, denoised, and binarised (Section 2.4).
2) The HR volume is fragmented into non-overlapping minicubes of size (e.g., 32^3 voxels).
3) Each HR minicube is down-sampled to the low-resolution scale (e.g., 16.5 µm/voxel) and intensity-normalised, producing the input features $\tilde{\mathbf{p}}^{(j)}$ of the training set.
4) Following the methodology detailed in (Orlov et al., 2025), each HR mini-cube was assigned a discrete rock type label $y^{(j)} \in \{1,\ldots,K\}$ based on its simulated absolute permeability. Permeability values, calculated via direct numerical simulation of Stokes flow (see Section 2.5), were partitioned into K ranges corresponding to distinct petrophysical classes. The thresholds for these ranges were determined from the permeability distribution of the entire training dataset to ensure a representative number of samples per class, resulting in e.g., K=5 rock classes, where Type 1 represents impermeable microporous matrix and Type 5 represents highly permeable vuggy channels (see Table 1).
5) The encoder-classifier f_θ is trained on the labelled dataset $\{(\tilde{\mathbf{p}}^{(j)},\ y^{(j)})\}_{j=1}^{N}$. Two variants are compared:

- **Supervised baseline**: 3D DenseNet-121 trained from scratch using cross-entropy loss.
- **Self-supervised baseline**: SwinVT encoder pre-trained using the modified VoCo framework with proxy-permeability regularisation, followed by fine-tuning of a linear classification head.

Table 2. Overview of the multi-scale μCT datasets and laboratory measurements.

No.	Physical size (mm)	Voxel size (μm)	Porosity (%)	Permeability (mD)
LR	$\varnothing 30 \times 75$	16.5	13.3	54.2
HR	$\varnothing 15 \times 10$	5.0	/	/

2.2.2 Stage B: Darcy-scale model construction (right contour in Fig. 1).

1) Low-resolution (LR) μCT data (e.g., 16.5 μm/voxel) are acquired and preprocessed (cropping, denoising).
2) The LR volume is fragmented into non-overlapping minicubes e.g., size 11^3 voxels, forming the computational frame of the Darcy-scale model.
3) Each LR minicube $\mathbf{p}^{(i)}$ is intensity-normalised and passed through the trained classifier $f_\theta(\mathbf{p}^{(i)}) = \hat{y}^{(i)}$.
4) Predicted rock-type labels $\hat{y}^{(i)}$ are assembled into the coarsened Darcy-scale volume $\mathbf{M}_{Darcy} \in \{1, \ldots, K\}^{X' \times Y' \times Z'}$.
5) Continuum-scale (Darcy) flow simulations are performed on $\mathbf{M}_{Darcy}$ using the effective properties $\{\phi_k,\ k_k\}_{k=1}^{K}$ of each rock type to compute macroscopic absolute permeability and porosity.

This automated pipeline offers several key advantages over traditional multi-scale workflows: 1) Fully automated rock type mapping without manual intervention or physical subsampling; 2) ability to predict as many rock types as supported by the training data; 3) extraction of all necessary information solely from low-resolution μCT scans; and 4) applicability to arbitrarily large core samples while preserving physical size and large-scale heterogeneity.

Further, both classifiers - supervised DenseNet-121 and self-supervised physics-based VoCo - are tested on an unseen held-out 22 μm test sample (see Section 3.2.5). The dataset preparation pipeline is the same as for 16.5 μm sample. The volume is fragmented into minicubes of size 9^3 voxels and further normalized patches are passed through one of the two classifiers to predict a discrete digital rock type. The predicted labels are then assembled into two separate Darcy-scale multiclass models defined on the coarsened grid of the 22 μm core: One model derived from the DenseNet-121 classifier and one from the self-supervised encoder.

Given the lack of pore-scale ground truth and laboratory flow measurements for this particular sample, the evaluation on the 22 μm core focuses on qualitative, but physically motivated, visual comparisons. Two-dimensional slices and three-dimensional renderings of the two Darcy-scale models are analysed alongside the original 22 μm μCT data, with emphasis on the continuity and spatial organization of high-permeability channels, low-permeability barriers, and layered heterogeneities.

The 22 μm experiment therefore serves as a held-out qualitative test of how well the two trained classifiers can extrapolate their learned rock-typing criteria to a new resolution and an entirely unseen core volume, without any additional retraining or re-calibration.

The implementation of the proposed physics-guided self-supervised pipeline, including training scripts and configuration files, is available at https://github.com/Batr97/Upscaling_CTscans.

2.3 Data

2.3.1 Carbonate core and multi-scale μCT imaging

The supervised and upscaling stages are based on multi-scale X-ray μCT of a heterogeneous carbonate core. A standard core sample of 30 mm diameter and 75 mm length was scanned at coarse spatial resolution (16.5 μm), and a smaller core sample of 15 mm diameter and 10 mm length was extracted from the same core and imaged at 5 μm resolution. The high-resolution dataset is used to construct detailed benchmark binary models and to label the data with rock types via pore-scale flow simulations, whereas the low-resolution scan provides the frame for Darcy-scale model construction.

Table 2 summarises the physical dimensions, voxel sizes and experimentally measured single-phase properties. These petrophisycal properties are later used as reference values when assessing the Darcy-scale models assembled from DenseNet-121 and VoCo classifiers.

The raw μCT volumes consist of 4,587 and 1,564 16-bit grayscale slices for the low- and high-resolution scans, respectively. Each slice is cropped to exclude non-informative margins, then denoised using a bilateral filter to suppress high-frequency noise while preserving pore-grain interfaces. Segmentation into pore and solid phases is performed with the Random Walker algorithm applied independently along three orthogonal directions, and the three binary volumes are combined via voxel-wise median voting to obtain robust 3D pore-solid maps (Orlov et al., 2025).

2.3.2 Labelled dataset for supervised rock typing

The supervised classifier operates on 3D mini-cubes that mimic the effective resolution of the 16.5 μm volume while retaining the detailed physics from the 5 μm binary model. To construct the labelled dataset, the high-resolution binary core sample is run through fragmentation step, i.e. first partitioned into non-overlapping minicubes with size 32^3 voxels. For each fragment, a corresponding low-resolution representation is synthesised in three steps:

1) Downsample the 32^3 binary crop to an 11^3 volumetric image
2) Apply a 3D binomial blur to approximate partial-volume effects and scanner point-spread.
3) Upsample the blurred 11^3 volume back to 32^3 with trilinear interpolation.

This procedure produces synthetic low-resolution crops that closely match the grayscale statistics and visual appearance of the real 16.5 μm scan but preserve the one-to-one correspondence with the underlying 5 μm binary data and associated pore-scale simulations. For each original 32^3 frag-

Table 3. Crop sizes and corresponding μCT resolutions used for SSL pre-training.

Crop size (voxels)	Resolution (μm/voxel)
60^3	3.1
54^3	3.6
14^3	13.0
12^3	16.5

ment, porosity and a full permeability tensor along the three Cartesian directions are computed on the high-resolution binary cube using single flow simulations, see Section 2.5.

To alleviate severe class imbalance, augmentations are applied to the synthetic low-resolution crops, see supplementary material. After augmentation, the supervised dataset contains approximately 30000 labelled minicubes of size 32^3, which are used both to train DenseNet-121 from scratch and to fine-tune the classifier head on top of the pre-trained VoCo encoder.

During inference and Darcy-scale model construction, the real low-resolution μCT volume (16.5 μm) is partitioned into non-overlapping mini-cubes at the coarse scale, and each fragment is classified into a digital rock type by either DenseNet-121 or VoCo, generating two alternative multi-class upscaled models.

2.3.3 Unlabelled datasets for self-supervised pre-training

Self-supervised pre-training of the VoCo encoder leverages a heterogeneous collection of 3D carbonate μCT volumes acquired at different resolutions to expose the model to diverse pore-space morphologies and scale-dependent textures. In total, four datasets are used:

1) Estaillades carbonate at 3.1 μm with sizes 1,725×1,300×1,130 (Portal, 2020b).
2) Estaillades carbonate at 3.6 μm with sizes 761×1,000×1,000 (Portal, 2020a).
3) Skoltech carbonate scans at 13 μm with sized 2,650×1,500×1,500 and 16.5 μm with sizes 4,560×1,200×1,200 (Orlov et al., 2025).

From each volume, a large set of non-overlapping 3D sub-volumes is extracted, aka samples cropping step in Fig. 3, that serve as inputs for the VoCo pretext task. To ensure comparable physical coverage across datasets with different voxel sizes, the sub-volume dimensions are chosen such that the physical edge length is approximately constant, Table 3:

Within each sub-volume, a fixed grid of 16 non-overlapping base crops is defined, which act as prototypes for contextual position prediction in the VoCo framework, and 2 additional random crops are sampled at each training iteration. Across all four datasets, this sampling strategy yields approximately 180,000 sub-volumes, from which base and random crops are drawn throughout pre-training (Fig. 2).

Porosity is measured as the fraction of pore voxels, percolation analysis identifies the size of the spanning pore cluster along each principal axis, and directional proxy permeabilities are defined by combining porosity with the percolating pore fraction. These proxy values are then used to build a hydrodynamic similarity matrix that guides the organisation of latent embeddings during self-supervised training (see Section 3.1.2).

2.4 High-resolution image processing and binary model construction

The high-resolution μCT dataset, acquired at 5μm voxel size, serves as the foundation for generating ground-truth binary models required for rock typing and subsequent pore-scale simulations. The raw grayscale images undergo a series of preprocessing steps to enhance quality and enable accurate segmentation. First, each 2D slice is cropped to remove non-informative edge regions. A bilateral filter (Banterle et al., 2012) is then applied to reduce noise while preserving edge sharpness, which is critical for maintaining pore boundaries.

Segmentation into pore and solid phases is performed using the Random Walker algorithm (Grady, 2006). To improve robustness and reduce directional bias, the Random Walker segmentation is applied independently along the three orthogonal directions, and the resulting binary volumes are combined using median averaging. This multi-directional approach mitigates artifacts and yields a more reliable three-dimensional binary representation of the pore space.

The resulting binary model at 5μm resolution has a porosity of 12.3%, closely matching the experimentally measured value of 13.3% (Table 2), confirming the accuracy of the segmentation workflow.

2.5 Methods for single-phase flow calculations

a) Assembling the Darcy-scale model from predicted classes: For each low-resolution μCT volume, the trained classifier $f_{\boldsymbol{\theta}}$ (supervised or self-supervised) assigns a rock-type label $\hat{y}^{(i)} \in \{1,\ldots,K\}$ to every non-overlapping minicube $\mathbf{p}^{(i)}$ of size 11^3 voxels. These labels are then mapped onto a coarsened Cartesian grid, yielding the Darcy-scale multi-class model $\mathbf{M}_{Darcy} \in \{1,\ldots,K\}^{X'\times Y'\times Z'}$. Each cell of $\mathbf{M}_{Darcy}$ is treated as a homogeneous continuum block and is assigned the effective porosity ϕ_k and permeability tensor k_k of its corresponding rock type $k = \hat{y}^{(i)}$, pre-computed from pore-scale simulations on high-resolution minicubes. The resulting model preserves the physical size and large-scale heterogeneity of the original core sample, while reducing the total number of grid cells by more than an order of magnitude.

b) Single-phase flow on the Darcy-scale grid: To evaluate the impact of representation learning (supervised vs. self-supervised) on upscaled transport properties, this work computes single-phase absolute permeability on each Darcy-scale model by solving the Darcy equations on the multi-class grid. Fluid flow in the porous medium is described by the Darcy equations:

$$\nabla \cdot u = 0,\ u = -\frac{K(x)\nabla p}{\mu} \tag{2}$$

where u is the superficial velocity averaged over a macroscopic region, p is the pressure, μ is the dynamic viscosity, and $K(x)$ is the local permeability tensor, which is piecewise constant inside each rock-type region. For each Cartesian direction, a

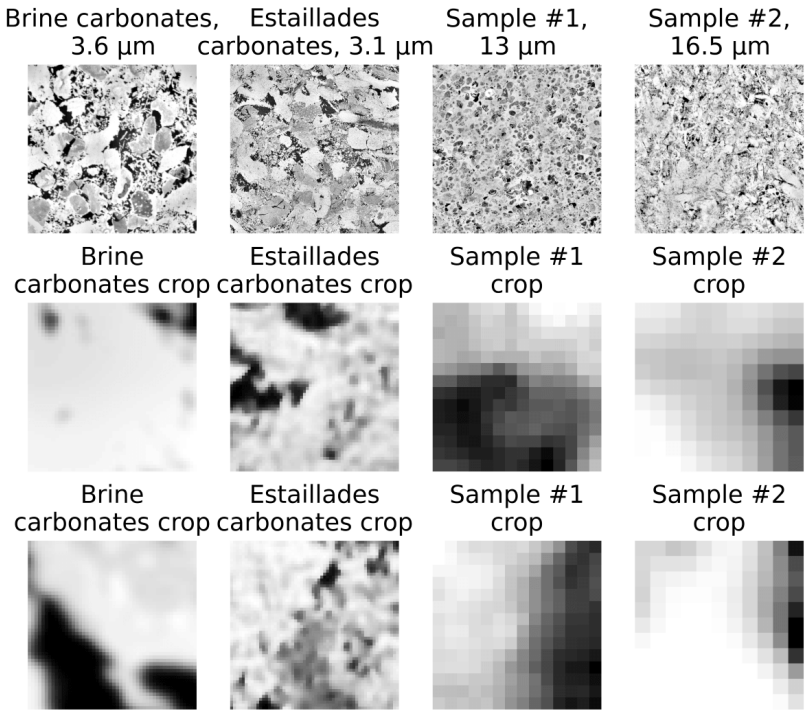

Fig. 2. Illustration of the self-supervised pre-training data: Example slices from the four carbonate datasets at different resolutions and corresponding base/random crops.

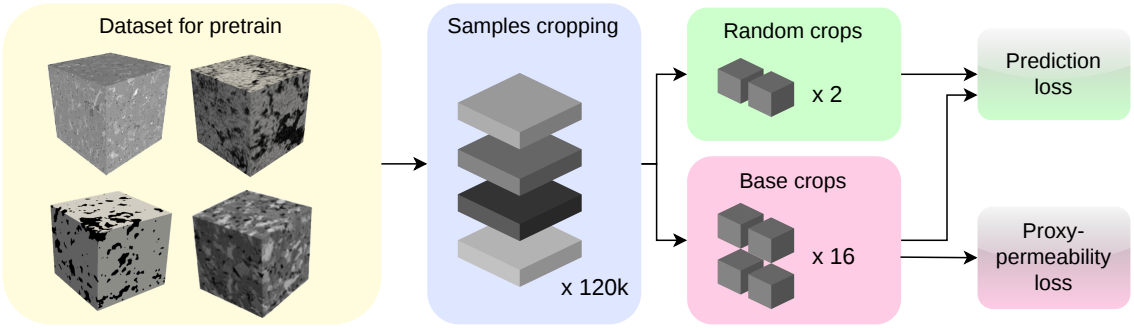

Fig. 3. SSL pretrain scheme.

fixed pressure drop is imposed between opposite faces of the sample and no-flow conditions on the remaining boundaries, and compute the volume-averaged flux $\langle u \rangle$. Here $\langle . \rangle$ denotes volume averaging over the entire Darcy-scale domain. The corresponding component of the effective permeability tensor K^{eff} is then obtained from the Darcy law using the known pressure gradient and viscosity.

The Darcy-scale flow problem is solved using the same finite-volume CFD solver (SigmaFlow) that is employed for pore-scale Stokes simulations, but with cell-wise heterogeneous permeability $K(x)$ instead of explicit pores. This study compares the effective permeabilities K^{eff}_{sup} and K^{eff}_{ssl} obtained from the supervised and self-supervised Darcy-scale models, respectively, against laboratory measurements and high-resolution direct numerical simulation benchmarks. This comparison provides a physically interpretable criterion for assessing how well each representation-learning strategy preserves large-scale transport properties.

During computation of permeability of the laboratory core the flow direction was aligned with the core axis. A uniform inlet velocity $U_{in} = 10^{-2}$ m/s was prescribed, while the outlet boundary was set to zero-gauge pressure; the lateral boundaries were treated as impermeable no-slip walls. Fluid properties were fixed to $\rho = 1{,}204$ kg/m^3 and $\mu = 1.834 \times 10^{-5}$ Pa·s

SigmaFlow uses a finite-volume discretization and a simple-type pressure-velocity coupling. Convergence was assumed when the relative iterative change of the velocity field dropped below 10^{-5}.

The effective permeability was extracted from the total pressure drop reported by SigmaFlow and converted to Darcy units. The same simulation setup was applied consistently for all compared upscaled models, and the resulting permeabili-

Table 4. Architecture comparison for 3D image analysis.

Architecture	Params (M)	Peak mem (GB)	Comment
ResNet-18	33.2	0.53	Residual baseline
DenseNet-121	11.2	0.50	Dense connectivity
3D U-Net	4.8	0.50	Segmentation-oriented

ties were validated against the laboratory core measurement reported in Table 2 and used as the reference in Table 9.

3. Experiments

3.1 SSL framework for 3D image classification

3.1.1 Supervised baseline: Classification with DenseNet-121

As a baseline model for supervised classification of core fragments, the DenseNet-121 architecture was employed. This choice is motivated by the specific requirements of the task: Processing 3D grayscale mini-cubes of limited size (32^3 voxels) with a need for computational efficiency and robust feature extraction.

DenseNet architectures are characterized by a dense connectivity pattern, where each layer receives the feature maps of all preceding layers as input. This design offers several key advantages for 3D image analysis. First, it promotes feature reuse and strengthens gradient flow throughout the network, mitigating the vanishing gradient problem and enabling the training of deeper architectures even with moderate dataset sizes. Second, DenseNet-121 incorporates bottleneck layers ($1\times1\times1$ convolutions) that reduce the number of input channels before the more expensive $3\times3\times3$ convolutions, significantly decreasing the number of parameters and computational cost while preserving representational capacity.

To adapt the network for volumetric data, all 2D operations were converted to their 3D counterparts. The input tensor shape was changed from (H,W) to (D,H,W), and all 2D convolutional and pooling layers were replaced with 3D versions using kernels of size (k,k,k). The kernel size in the initial convolutional layer was also reduced to better suit the smaller spatial dimensions of the input patches.

To assess the computational efficiency of different architectures for 3D μCT fragment classification, this study evaluates three widely used backbones available in the MONAI framework based on their parameter count and peak graphics processing unit (GPU) memory consumption during inference. Table 4 summarizes this comparison for input volumes of size 32^3 voxels with batch size 1 and FP32 precision on an RTX A4000 GPU. The architectures include a residual baseline (ResNet-18), a densely connected network (DenseNet-121), and a lightweight segmentation-oriented 3D U-Net variant with channel dimensions ranging from 16 to 256.

The results demonstrate that DenseNet-121 offers the most favorable trade-off between model capacity and computational efficiency. Compared to ResNet-18, it reduces the number of trainable parameters by approximately 66%, while its memory footprint during inference remains virtually identical to that of the compact 3D U-Net architecture (0.50 GB). Given its proven effectiveness in CT-based medical image analysis tasks (Hasan et al., 2021; Babu and Brindha, 2024; Molinski et al., 2024) and its parameter efficiency demonstrated here, DenseNet-121 was selected as the backbone for the supervised baseline in all subsequent experiments.

3.1.2 Proposed approach: Self-supervised VoCo with physics-guided loss

In the self-supervised branch a 3D Swin-based encoder is tailored for volumetric CT data. The backbone follows the Swin-UNETR design, where a 3D Swin Transformer (SwinVT) with shifted window self-attention and four hierarchical stages (depths, heads) is coupled with convolutional UnetrBasicBlock encoders operating at multiple resolutions. Feature maps from five scales (input convolutional stem, three intermediate transformer stages, and the deepest stage) are aggregated via global adaptive average pooling, concatenated, and passed through a three-layer projection head with batch-normalised linear blocks to obtain a 2,048-dimensional embedding used in the VoCo objective. The same SwinVT encoder is instantiated as a student-teacher pair updated with exponential moving average, so that contextual position prediction and permeability-regularised contrastive learning are simultaneously applied to stable volumetric representations during pretraining.

1) Volume contrastive Pre-training framework

The proposed self-supervised approach is built upon the concept of volume contrastive learning, originally introduced by (Wu et al., 2024) for 3D medical image analysis. The fundamental insight from VoCo is that volumetric data contains inherent contextual information-spatial relationships between different regions-that can serve as a supervisory signal without requiring labels. However, unlike medical images where organ positions are anatomically consistent, rock μCT images exhibit stochastic, heterogeneous structures. Therefore, the original framework was substantially modified to account for the unique characteristics of porous media and to incorporate physical knowledge about fluid flow by presenting SSL framework tailored for 3D μCT rock images. The core idea is to pretrain a feature extractor on unlabeled minicubes by solving a pretext task that encourages the model to learn representations sensitive to both spatial context and hydrodynamic similarity Fig. 3.

Given an input volume, firstly it was partitioned into a grid of n non-overlapping *base crops* that collectively cover the entire volume. These base crops serve as prototypes representing different spatial regions within the core sample. Their features, after passing through the backbone network and a projector head, are denoted as $\mathbf{q}_i \in \mathbb{R}^{1\times C}$, where C is the embedding dimension. Further the pipeline followed by random sampling of a *query crop* from the same volume and extract its feature representation $\mathbf{p} \in \mathbb{R}^{1\times C}$.

The pretext task is formulated as a similarity prediction problem. For each query crop, cosine similarity scores (l_i) were

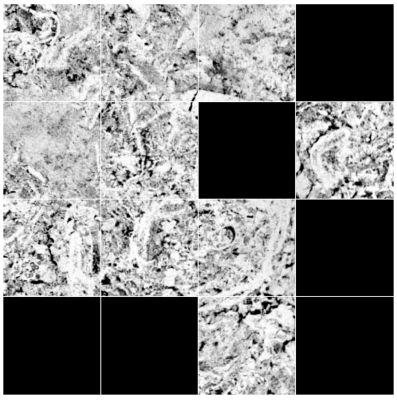

Fig. 4. Random masking of base crops during proxy-permeability calculation.

computed with all base crops:

$$l_i = \mathrm{CosSim}(\mathbf{p}, \mathbf{q}_i) = \frac{\mathbf{p} \cdot \mathbf{q}_i}{\|\mathbf{p}\| \|\mathbf{q}_i\|}, \quad i \in \{1, \ldots, n\} \tag{3}$$

Here, $\mathrm{CosSim}(\mathbf{p}, \mathbf{q}_i)$ denotes the cosine similarity between two vectors $\mathbf{p}$ and $\mathbf{q}_i$, defined as their inner product normalised by the product of their Euclidean norms.

To supervise this task, soft pseudo-labels y_i were generated based on the actual 3D overlap between the query crop and each base crop. Specifically, y_i is defined as the proportion of the query crop's volume that falls within the i-th base crop region. The similarity logits l_i are then mapped to soft probabilities via a temperature-scaled softmax before comparing with the pseudo-labels $\mathbf{y}$. The prediction loss (L_{pred}) is then computed as:

$$\mathscr{L}_{pred} = -\frac{1}{n} \sum_{i=1}^{n} \log(1 - |y_i - l_i|) \tag{4}$$

where the non-linear transformation $\log(1 - |y_i - l_i|)$ increasingly penalises large discrepancies between predicted similarities and geometric overlaps. This surrogate objective plays a role analogous to cross-entropy for soft labels but is numerically more stable in this setting. This formulation encourages the model to predict the spatial provenance of each query crop relative to the fixed base grid, effectively learning representations that encode volumetric context.

2) Proxy-permeability regularization

While the contextual prediction task captures spatial relationships, it does not explicitly account for the physical properties that govern fluid flow. To address this limitation, a novel regularization term is introduced, which leverages a physically meaningful proxy for permeability, computed directly from segmented high-resolution images.

Every mini-cube is characterized by a directional permeability proxy. First, porosity ϕ is calculated as the fraction of pore voxels. Then, percolation analysis is performed along each axis to determine the fraction of pores belonging to the percolating cluster $P_f = n_p/N$, where n_p is the number of pore voxels in the percolating cluster and N is the total number of pore voxels. The directional permeability proxy (k_{axis}) is then defined as:

$$k_{axis} = \phi \cdot P_f \tag{5}$$

This proxy captures both the total pore volume and its connectivity, providing a simple yet effective estimate of the sample's ability to transmit fluid.

For a pair of mini-cubes i and j, the work defines of *hydrodynamic similarity* (G_{ij}) based on the difference in their permeability proxies. The difference is converted into a similarity score using an exponential kernel with temperature parameter τ:

$$G_{ij} = \exp\left(-\frac{|k_i - k_j|}{\tau}\right) \tag{6}$$

Here, k_i and k_j are the permeability proxy values (averaged over the three axes for simplicity). For each fragment i, a scalar proxy permeability k_i is obtained by averaging the directional proxies k_{axis} over the three axes. G_{ij} ranges from 0 to 1, with values close to 1 indicating that the two fragments are hydrodynamically similar (small permeability difference), and values near 0 indicating dissimilarity.

Simultaneously, the cosine similarity (CosSim) is computed between the embeddings of the same two mini-cubes after processing by the backbone and projector.

$$S_{ij} = \mathrm{CosSim}(\mathbf{q}_i, \mathbf{q}_j) = \frac{\mathbf{q}_i \cdot \mathbf{q}_j}{\|\mathbf{q}_i\| \|\mathbf{q}_j\|} \tag{7}$$

The key insight of the proposed approach is to enforce consistency between the embedding similarity S_{ij} and the physically-derived similarity G_{ij}. The regularization loss (L_{perm}) is formulated as the Frobenius norm of their difference:

$$\mathscr{L}_{perm} = \|\mathbf{S} - \mathbf{G}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (S_{ij} - G_{ij})^2 \tag{8}$$

This loss term acts as a strong inductive bias, pulling together the embeddings of mini-cubes that are hydrodynamically similar (high G_{ij}) while pushing apart those with different flow properties (low G_{ij}). In this way, the loss gives the latent space a physical structure: Similarity between embeddings is encouraged to reflect similarity in permeability-related properties, rather than only similarity in image appearance. Crucially, this regularization is applied during pre-training on *unlabeled* data, as the permeability proxy is computed directly from the segmented images without any manual annotation.

Computing the percolation coefficient for all n base crops across the entire dataset is computationally expensive, as it requires analyzing connected pore clusters for each 3D fragment. To address this, a masking strategy is applied in which a subset of base crops is randomly sampled for permeability proxy calculation, as illustrated in Fig. 4.

Specifically, from the total set of base crops arranged in a grid, each crop is selected independently with probability $p = 0.2$. Only the selected crops undergo the full percolation analysis and proxy-permeability computation described above. The matrix $\mathbf{G}$ and the corresponding entries of $\mathbf{S}$ are therefore constructed only for the subset of base crops with available proxy permeabilities, and L_{perm} is evaluated on this submatrix. This stochastic masking strategy reduces the computational cost by approximately 80%, enabling efficient pre-training on

Table 5. Correlation between simulated absolute permeability and the proxy permeability for the three Cartesian directions. Pearson coefficients are computed for $\log_{10}$-transformed permeabilities.

Component	Pearson ($\log_{10}$)	Spearman
K_x	0.972	0.960
K_y	0.975	0.960
K_z	0.997	0.973
Mean	0.996	0.975

Table 6. Classification performance on labelled mini-cubes.

Model	Accuracy	F1-macro	F1-micro
DenseNet-121	0.758	0.759	0.778
VoCo + linear head	0.781	0.784	0.781

Table 7. Per-class metrics for self-supervised VoCo encoder.

Class	Precision	Recall	F1-score
0	0.827	0.795	0.804
1	0.813	0.796	0.798
2	0.791	0.773	0.778
3	0.780	0.760	0.769
4	0.792	0.761	0.771

large volumes while maintaining sufficient coverage of the spatial domain. The random selection also acts as a form of data-dependent regularization, preventing the model from over-relying on any fixed spatial location during the contrastive learning process.

This optimization is particularly important given the scale of the pre-training dataset and the need to compute percolation statistics for each sampled fragment. Thus, a balance is achieved between computational efficiency and the richness of the physical signal provided to the model.

3) Final objective function

The complete objective function for pre-training combines the contextual prediction task with the physics-guided regularization:

$$\mathscr{L} = \mathscr{L}_{pred} + \lambda \mathscr{L}_{perm} \tag{9}$$

where λ is a balancing hyperparameter that controls the relative contribution of the two loss terms. The first term ensures that the model learns representations sensitive to spatial context and volumetric structure, while the second term embeds physical knowledge about permeability into the embedding space. Through this combined objective, the model is encouraged to organize the latent space according to both geometric position and hydrodynamic function, providing a rich foundation for subsequent fine-tuning on the supervised classification task.

3.2 New framework validation using multiscale μCT date

3.2.1 Comparison of supervised and self-supervised models

Before comparing the supervised and self-supervised upscaled models, the permeability proxy used in the VoCo loss is first verified to be physically meaningful. For this purpose, a subset of about 500 binary crops was selected (see the dataset for pretraining VoCo in Section 2.3) and computed absolute permeability along the three Cartesian directions (K_x, K_y, K_z) (Section 2.5), treating these values as ground-truth permeabilities at the crop scale. On the same set of fragments, directional proxy permeabilities derived from porosity and percolation-based connectivity were evaluated (Section 3.1.2) and analysed the correlation between the two measures. Because the relationship spans several orders of magnitude and is strongly monotonic but not strictly linear, Both linear (Pearson) and monotonic (Spearman) correlations between the simulated and proxy permeabilities were quantified reported in Table 5, all associated p-values are below 10^{-3}.

For all directions, both the Pearson and Spearman coefficients exceed 0.96, indicating a very strong association between the proxy and simulated permeabilities, confirming that the proxy closely tracks the simulated absolute permeability over several orders of magnitude and can be reliably used as a physically grounded signal in the self-supervised regularisation term of VoCo.

Two alternative upscaled Darcy-scale models for the same 16.5 μm carbonate core are then compared: One populated by the supervised DenseNet-121 classifier and one populated by the VoCo-based model after self-supervised pre-training. Both classifiers are evaluated on the held-out set of labelled 32^3 mini-cubes constructed from the high-resolution μCT, using overall accuracy and macro/micro-averaged F1-scores as quality metrics. The VoCo-based model consistently outperforms DenseNet-121 in all three metrics (Table 6), with the largest gains observed in macro F1-score, indicating improved recognition of minority digital rock types that are critical for flow but under-represented in the labelled dataset.

Across all five rock types, the self-supervised model matches or slightly improves the performance of the purely supervised baseline, while providing a clear gain in macro F1-score (Tables 7 and 8). The improvement is most pronounced for the flow-critical intermediate classes 2 and 3. This indicates that the physics-guided self-supervised pretraining makes the classifier more robust to underrepresented morphologies without degrading performance on the dominant classes 0 and 1.

The differences observed at the mini-cube level translate into distinct spatial patterns in the Darcy-scale models. Fig. 5 shows representative 2D cross-sections of the two multi-class models. Both approaches reproduce the main layering of the carbonate sample, but the DenseNet-121-based model tends to fragment high-permeability channels into shorter, disconnected

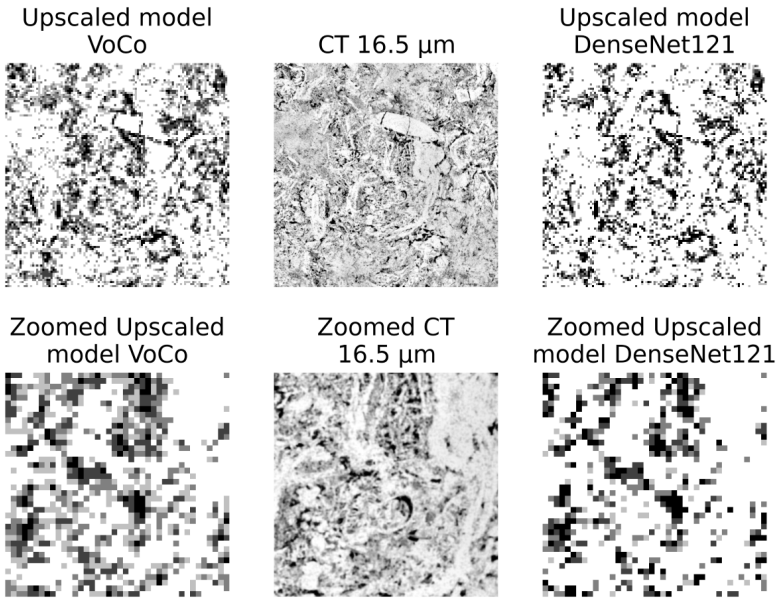

Fig. 5. Darcy-scale multi-class models for the 16.5 µm carbonate core constructed from DenseNet-121-based and VoCo-based models.

Table 8. Per-class metrics for supervised DenseNet-121 baseline.

Class	Precision	Recall	F1-score
0	0.800	0.772	0.782
1	0.795	0.763	0.774
2	0.750	0.727	0.732
3	0.750	0.723	0.733
4	0.792	0.761	0.774

Table 9. Darcy-scale porosity and absolute permeability for the laboratory core and two digital models.

Model	ϕ (%)	K_x (mD)	K_y (mD)	K_z (mD)	K_{avg} (mD)
Lab core	13.3	/	/	54.2	/
DenseNet-121-based	14.0	47.3	39.4	47.6	44.8
VoCo-based	14.2	55.0	45.8	54.5	52.0

segments and introduces isolated high-permeability patches within the tight matrix. This breaks the continuity of conductive pathways and creates artefacts that are not consistent with the underlying high-resolution structure. The VoCo-based model produces smoother class maps with fewer salt-and-pepper artefacts, preserves the connectivity and thickness of channel-like features, and maintains coherent low-permeability barriers, resulting in facies patterns that are visually closer to the geological structures observed in the original µCT images.

To quantify the impact of these structural differences on macroscopic properties, porosity and directional absolute permeability (K_x, K_y, K_z) are computed for each upscaled model, using pre-computed effective properties for each digital rock type. Table 9 reports the simulated Darcy-scale properties for each model, where K_{avg} denotes the arithmetic mean $(K_x + K_y + K_z)/3$. In terms of relative error with respect to the laboratory permeability, VoCo shows nearly 4%, while for the DenseNet-121 the error is nearly 17%, indicating that the VoCo-based Darcy-scale model provides a much closer match to the experimental permeability than the purely supervised baseline. Both models slightly overestimate porosity relative to the laboratory value, but the deviation remains modest (~ 7 % for VoCo and ~ 5 % for DenseNet-121), suggesting that the main difference lies in how connectivity of conductive facies is captured rather than in the overall pore volume. Overall, the consistent improvement of VoCo across classification metrics, visual facies continuity and Darcy-scale permeability demonstrates that incorporating a permeability-guided self-supervised pre-training stage leads to more physically faithful rock typing under the same labelling budget.

In addition to the experiments on the 16.5 µm low-resolution volume, a qualitative held-out test was performed on the independent 22 µm core scan described in Section 3.2.5. For this sample, Darcy-scale multiclass models were generated using both the supervised DenseNet-121 classifier and the physics-guided self-supervised encoder, and representative 2D slices were visually compared with the corresponding grayscale µCT slice. The zoomed fragments in Fig. 6 highlight that the self-supervised upscaled model reproduces the large-scale zonation of more conductive and more resistive regions in closer agreement with the intensity patterns of the original image, preserving the continuity of dominant high-conductivity channels while retaining surrounding lower-

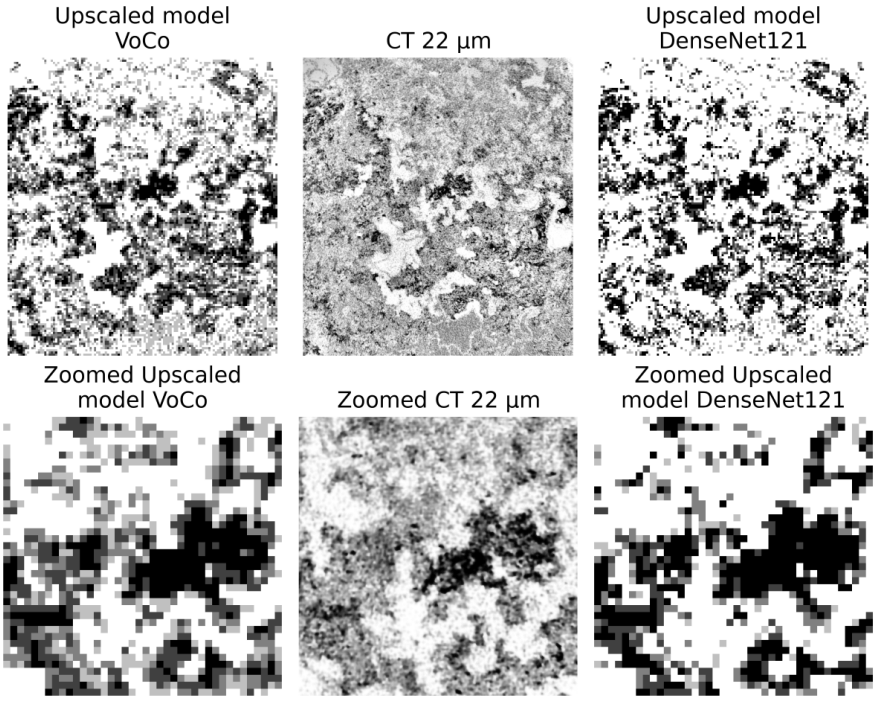

Fig. 6. Comparison of SSL and supervised approaches on a held-out 22 µm test sample.

conductivity facies.

In contrast, the Darcy-scale model obtained from the purely supervised DenseNet-121 classifier tends to over-emphasize high-permeability rock types across the same field of view. This manifests as a systematic expansion and merging of conductive regions at the Darcy scale, with low- and intermediate-conductivity facies being more frequently promoted into higher-conductivity classes and fine-grained background domains being mapped into the most resistive type. As a result, the supervised model appears to under-represent subtle low-permeability barriers and baffles that are visible in the 22 µm µCT data, whereas the physics-guided self-supervised model provides a more balanced representation of conductive pathways and blocking structures.

3.2.2 Robustness under limited labelled data

To evaluate the label efficiency of the two competing approaches, an additional experiment is conducted in which the amount of labelled data available for supervised learning is deliberately reduced. Starting from the full labelled dataset described in Section 2.3, three reduced training subsets are constructed containing 50%, 25%, and 10% of the original number of labelled minicubes. In all cases, the held-out test set remains unchanged, so that the results obtained at different label fractions are directly comparable.

The subsets are created by stratified sampling with respect to the digital rock classes in order to preserve, as closely as possible, the class distribution of the complete labelled dataset. This is particularly important in the present problem because several rock types are under-represented, yet they strongly affect the continuity of conductive pathways and, consequently, the quality of the resulting Darcy-scale model. The same augmentation strategy described in supplementary material is applied after subsampling, ensuring that both methods are compared under identical data conditions.

For the supervised baseline, DenseNet-121 (Section 3.1.1) is trained on each labelled-data configuration, including the full training set (100%) and the reduced subsets (50%, 25%, and 10%). For the SSL approach, the VoCo encoder pretrained on the full unlabelled dataset (Section 3.1.2) is fine-tuned on the same labelled-data configurations. Importantly, the supervised optimization protocol is kept identical for both models across all experiments, including the 100% labelled-data case: The same train/validation split, data preprocessing, augmentation pipeline, optimizer settings, learning-rate schedule, batch size, stopping criterion, and evaluation protocol are used throughout. This design ensures a controlled comparison between the two approaches, such that any performance differences can be attributed mainly to the effect of self-supervised pretraining. It also allows us to evaluate how both models behave as the amount of labelled training data is reduced.

Performance is evaluated on the same held-out labelled minicubes using overall accuracy together with macro- and micro-averaged F1-scores. The results of this experiment are summarized in Table 10. This comparison quantifies the robustness of both classifiers under progressive label scarcity and tests the central hypothesis of this work: Namely, that physics-guided self-supervised pretraining provides a clear advantage over purely supervised learning when only a small fraction of annotated rock fragments is available.

3.2.3 Latent representation analysis

To gain insight into what the VoCo encoder learns during the self-supervised stage, its latent space and intermediate act-

Table 10. Classification performance under limited labelled data.

Labelled data	Model	Accuracy	F1-macro	F1-micro
50%	DenseNet-121	0.734	0.736	0.734
	VoCo + linear head	0.740	0.743	0.740
25%	DenseNet-121	0.679	0.680	0.679
	VoCo + linear head	0.705	0.707	0.705
10%	DenseNet-121	0.625	0.627	0.625
	VoCo + linear head	0.683	0.685	0.683

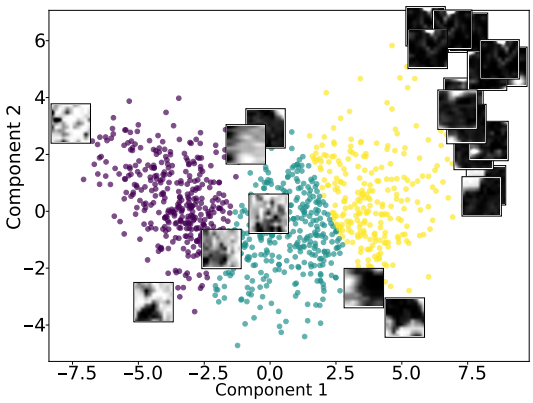

Fig. 7. 2D projection of VoCo latent embeddings for mini-cubes. Points correspond to individual fragments, and selected μCT slices are overlaid for representative clusters, illustrating that different regions of the latent space are associated with distinct carbonate facies and pore textures.

ivation are analysed. For a large set of 32^3 mini-cubes extracted from the 16.5 μm carbonate volume, the feature vectors from the penultimate layer of the pre-trained VoCo encoder are recorded. and project them to two dimensions using PCA. The resulting 2D embedding is shown in Fig. 7, where each point corresponds to a mini-cube; for several clusters, representative μCT slices from the corresponding fragments are overlaid. to illustrate the underlying pore-space morphology.

Visual inspection of the clusters and the associated μCT slices shows that these clusters are not mere reflections of the permeability-based labels, but correspond to distinct carbonate facies and pore textures. The left group is dominated by dense matrix with very low porosity and only a few isolated pores; The central group contains homogeneous, channel-like structures with elongated, well-connected throats; the right cluster corresponds to heterogeneous transitional zones where vugs and enlarged pores are embedded within a tighter background. In some cases, a single permeability class is split into multiple VoCo clusters that differ in morphology, indicating that the self-supervised encoder is sensitive to higher-order geometric patterns such as channel continuity and percolation paths, rather than only to the scalar permeability target used for rock typing. This latent organisation is consistent with the design of the proxy-permeability regularisation, which pulls together mini-cubes with similar porosity-percolation signatures and pushes apart fragments that differ in their ability to support flow even if their grayscale statistics are similar.

Further insight into how VoCo processes individual fragments is obtained by visualising activation maps at different depths of the network. Fig. 8 illustrates a central slice of a carbonate μCT fragment rescaled to the internal resolution of the model and the corresponding feature maps for the central channel after successive stages of downsampling in the encoder. In the first block, activations highlight sharp intensity transitions associated with grain boundaries and pore walls, indicating that the early layers act as generic edge and texture detectors. In the intermediate blocks, the model starts to aggregate local pores into larger structures: Activated regions trace clusters of connected pores and begin to suppress isolated speckles in the tight matrix. In the deepest block, after several downsampling steps, the activation pattern collapses onto a small number of contiguous high-response regions forming a continuous corridor that closely follows the main percolating pore network within the fragment. Thus, the encoder builds a hierarchical representation that progressively abstracts from local edges to mesoscale connectivity and ultimately to large-scale conductive pathways, in line with the physics of single-phase flow in the carbonate pore system.

Together, the clustering patterns in the VoCo latent space and the structure of the activation maps demonstrate that the self-supervised encoder internalises physically meaningful descriptors of the pore space. It organises mini-cubes according to textural facies and connectivity patterns that are directly relevant for permeability, providing a plausible explanation for the improved rock-typing accuracy and superior preservation of Darcy-scale flow properties.

3.2.4 Held-out test on a 22 μm CT sample

To assess the generalization capability of the supervised and self-supervised upscaling pipelines beyond the resolutions and samples used during training, both classifiers are additionally evaluated on an independent carbonate μCT scan acquired at 22 μm voxel size. This core plug has a physical height of approximately 60 mm and is not used at any stage of the workflow: It is excluded from the self-supervised pretraining, from the supervised fine-tuning, and from the calibration of effective porosity and permeability for the digital rock types. Consequently, it provides a strictly out-of-distribution test case for the learned representations and the resulting Darcy-scale models.

The 22 μm dataset undergoes the same preprocessing steps as the 16.5 μm low-resolution volume used in the main upscaling experiments. First, the raw grayscale μCT images are cropped to the region containing the core, and then a noise is reduced. The preprocessed volume is then partitioned into non-overlapping 3D sub-volumes of size $9 \times 9 \times 9$ voxels. This patch size is chosen so that the physical dimensions of each sub-volume match those of the minicubes used during pretraining and fine-tuning on the 5 μm/16.5 μm data, ensuring that

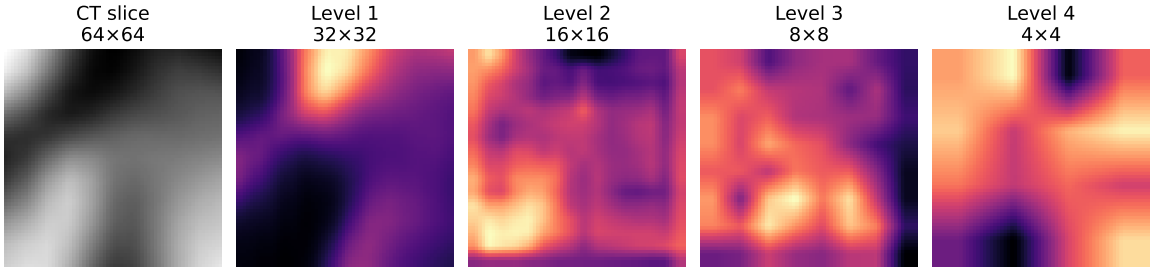

Fig. 8. Example of hierarchical feature extraction in the VoCo encoder.

the classifiers are applied at a consistent physical scale despite the different voxel size. Each 9^3 sub-volume is subsequently intensity-normalized using the same normalization scheme as for the training data.

No Darcy-scale single-phase flow simulations are performed for the 22 μm sample, and no quantitative petrophysical benchmarking is attempted, since effective porosity and permeability for this core have not been measured and the rock-type property calibration is derived exclusively from the 5 μm high-resolution dataset.

4. Discussion

The results of this study suggest that incorporating a physics-guided self-supervised pre-training stage into the digital rock upscaling workflow can improve both classification quality and the fidelity of Darcy-scale models compared to a purely supervised baseline trained from scratch for the considered carbonate core. At the fragment level, the VoCo-based encoder yields higher accuracy and macro-averaged F1-score than 3D DenseNet-121, with the largest gains observed for minority rock types that control the continuity of conductive pathways but remain under-represented in the labelled dataset. This improvement is consistent with earlier observations in medical imaging and 2D rock image analysis, where self-supervised features learned from large unlabeled corpora translate into better performance in low-label regimes. Although the improvement in classification metrics (accuracy, macro F1) achieved by the self-supervised VoCo model may appear moderate (2-3 percentage points), its practical significance is far greater when considering the overall upscaling workflow. The gain in macro F1 indicates better recognition of rare but flow-critical rock types, which is consistent with superior preservation of connected pathways and a fourfold reduction in permeability error (from 17% to 4%). Moreover, the SSL model demonstrates markedly better visual coherence of facies and higher robustness under limited labelled data. These combined advantages outweigh the modest increase in model complexity and strongly support physics-guided self-supervised pretraining as a promising approach for constructing reliable Darcy-scale digital rock models in similar settings.

Beyond standard classification metrics, the *benefit* of the physics-guided SSL stage is most evident at the level of upscaled Darcy-scale models. Visual inspection of multi-class maps shows that VoCo produces smoother facies distributions, maintains the connectivity and thickness of high-permeability channels, and avoids the salt-and-pepper artefacts and spurious high-permeability islands characteristic of the DenseNet-121-based model. These structural differences directly impact macroscopic transport properties: The VoCo-based model reproduces the experimental absolute permeability of the carbonate core with a relative error of about 4%, whereas the supervised baseline exhibits an error of order 17% despite a similar deviation in porosity. The fact that both models slightly overestimate porosity but differ strongly in permeability indicates that capturing connectivity of conductive facies is more critical for reliable upscaling than matching pore volume alone.

A key factor behind this behaviour is the permeability-aware regularisation introduced in the self-supervised loss. By enforcing consistency between cosine similarities in the latent space and a proxy permeability based on porosity and percolation analysis, the encoder is encouraged to group fragments that are not only texturally similar but also hydrodynamically equivalent. The strong Pearson and Spearman correlations above 0.96 between the proxy and simulated permeabilities confirm that this proxy provides a physically meaningful signal across several orders of magnitude and can therefore serve as a reliable guide for structuring the embedding space. Analysis of the latent representations further supports this interpretation: VoCo clusters mini-cubes into groups that correspond to distinct carbonate facies and connectivity patterns, and deep feature maps progressively focus on percolating corridors rather than isolated pores, suggesting that the encoder internalises physically relevant descriptors of flow pathways.

From a methodological standpoint, the proposed workflow shows that physics-guided SSL can be integrated into existing CNN-enhanced upscaling pipelines with minimal changes to the downstream stages. The same fragmentation, rock-typing criteria, and Darcy-scale flow solver are used for both the supervised and SSL-based variants, which isolates the effect of representation learning and simplifies the comparison. In contrast to earlier upscaling approaches that directly regress specific properties or rely on binary images only, the classification-based formulation operates on low-resolution grayscale CT data and predicts discrete rock types with precomputed effective properties, making it applicable even when pores are not fully resolved and when multiple property sets (e.g., elastic or thermal) must be attached to the same facies classes.

An important practical aspect of the comparison is compu-

tational cost. In the present study, the supervised DenseNet-121 baseline was trained directly on the labelled dataset, whereas the proposed SSL pipeline required an additional self-supervised pretraining stage on unlabeled carbonate volumes before fine-tuning on the same labelled data. Concretely, VoCo pretraining was performed for 250,000 optimisation steps, after which the encoder was fine-tuned for 60 epochs; the supervised DenseNet-121 baseline was trained for 50 epochs under the same labelled-data protocol. All experiments were run on a single NVIDIA L40S GPU. Thus, the main overhead of the proposed method lies in the one-time offline pretraining stage, while the downstream rock-typing workflow and Darcy-scale model construction remain unchanged.

At the same time, several limitations of the present study should be acknowledged. First, all experiments are conducted on a single, highly heterogeneous carbonate core, and the rock-typing criteria are deliberately kept simple with only five digital rock types defined by scalar porosity-permeability thresholds. This design facilitates controlled comparison but does not fully exploit the capacity of SSL encoders to separate more subtle textural facies or to handle strongly anisotropic permeability tensors; extending the framework to sandstones, tight gas rocks or mixed lithologies will require revisiting both the labelling scheme and the choice of proxy statistics. Second, the computation of proxy permeabilities and percolation-based similarity matrices is still relatively expensive, and although random masking of base crops reduces the cost by roughly 80%, further optimisation or approximate graph-based methods may be needed for very large pre-training sets.

Finally, the current implementation focuses on single-phase Darcy-scale permeability as the primary validation metric, while multi-phase flow behaviour is only indirectly constrained through the choice of rock types and their effective properties. Earlier studies on CNN-enhanced upscaling have demonstrated that relative permeability curves are more sensitive than absolute permeability to both coarsening and classification errors, especially in low-permeability rocks with complex wettability patterns. A natural extension of the proposed approach is therefore to augment the physics-guided loss with additional proxies related to capillary entry pressures, residual saturations or multiphase connectivity measures, and to evaluate VoCo-based Darcy-scale models against full sets of relative permeability curves and displacement experiments rather than only against absolute permeability and porosity. Such developments would move physics-guided SSL one step closer to a more general, lithology-agnostic tool for building multi-property digital twins of reservoir cores from multi-resolution CT data.

Acknowledgements

M. S. and R. P. were supported by Russian Science Foundation “Robust machine learning models creation” (No. 25-71-30008).

Supplementary file

https://doi.org/10.46690/ager.2026.07.04

Conflicts of interest

The authors declare no competing interest.

References

Azizi, S., Mustafa, B., Ryan, F., et al. Big self-supervised models advance medical image classification. Paper Presented at IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10-17 October, 2021.

Babu, P. P. S., Brindha, T. Deep learning fusion for intracranial hemorrhage classification in brain CT imaging. International Journal of Advanced Computer Science & Applications, 2024, 15(8): 884-894.

Banterle, F., Corsini, M., Cignoni, P., et al. A low-memory, straightforward and fast bilateral filter through subsampling in spatial domain. Computer Graphics Forum, 2012, 31(1): 19-32.

Behbahani, H., Azin, R., Osfouri, S. A comprehensive review of digital rock physics: From tomographic images to pore network modeling. Chemical Process Design, 2023, 2(1): 6-30.

Blunt, M. J., Bijeljic, B., Dong, H., et al. Pore-scale imaging and modelling. Advances in Water Resources, 2013, 51: 197-216.

Blunt, M. J., Sun, S., Boone, M. A., et al. Digital rock physics and fluid flow in the context of the energy transition. Advances in Geo-Energy Research, 2025, 18(3): 299-302.

Bultreys, T., De Boever, W., Van Hoorebeke, L., et al. A multi-scale, image-based pore network modeling approach to simulate two-phase flow in heterogeneous rocks. Paper SCA2015-027 Presented at International Symposium of the Society of Core Analysts, St. John’s Newfoundland and Labrador, Canada, 16-21 August, 2015.

Chen, T., Kornblith, S., Norouzi, M., et al. A simple framework for contrastive learning of visual representations. Paper Presented at International Conference on Machine Learning, Vienna, Austria, 12-18 July, 2020.

Ebadi, M., Orlov, D., Alekseev, V., et al. Lift the veil of secrecy in sub-resolved pores by xe-enhanced computed tomography. Fuel, 2022, 328: 125274.

Faisal, T. F., Islam, A., Jouini, M. S., et al. Numerical prediction of carbonate elastic properties based on multi-scale imaging. Geomechanics for Energy and the Environment, 2019, 20: 100125.

Fourer, D., Sidibé, D., Lecomte, J.-F., et al. A comparative evaluation of self-supervised methods applied to rock images classification. Paper Presented at 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Rome, Italy, 27-29 February, 2024.

Grady, L. Random walks for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1768-1783.

Hasan, N., Bao, Y., Shawon, A., et al. Densenet convolutional neural networks application for predicting covid-19 using ct image. SN Computer Science, 2021, 2(5): 389.

He, K., Fan, H., Wu, Y., et al. Momentum contrast for unsupervised visual representation learning. Paper Presented at IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June, 2020.

Jackson, S. J., Niu, Y., Manoorkar, S., et al. Deep learning of multiresolution X-ray micro-computed-tomography images for multiscale modeling. Physical Review Applied, 2022, 17(5): 054046.

Jing, L., Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(11): 4037-4058.

Law, B. E., Spencer, C. W., Howell, D. G. Gas in tight reservoirs – an emerging major source of energy. The Future of Energy Gases: US Geological Survey Professional Paper, 1993, 1570: 233-252.

Molinski, N. S., Kenda, M., Leithner, C., et al. Deep learning-enabled detection of hypoxic-ischemic encephalopathy after cardiac arrest in ct scans: A comparative study of 2D and 3D approaches. Frontiers in Neuroscience, 2024, 18: 1245791.

Mukhametdinova, A., Kazak, A., Karamov, T., et al. Reservoir properties of low-permeable carbonate rocks: Experimental features. Energies, 2020, 13(9): 2233.

Niu, Y., Wang, Y., Mostaghimi, P., et al. An innovative application of generative adversarial networks for physically accurate rock images with an unprecedented field of view. Geophysical Research Letters, 2020, 47(23): e2020GL089029.

Orlov, D., Ebadi, M., Muravleva, E., et al. Different methods of permeability calculation in digital twins of tight sandstones. Journal of Natural Gas Science and Engineering, 2021, 87: 103750.

Orlov, D., Gainitdinov, B., Koroteev, D. Darcy-scale digital core models for rock properties upscaling and computational domain reduction. Journal of Computational Science, 2025, 92: 102715.

Orlov, D. M., Alekseev, V. V., Pimanov, V. O., et al. Multiscale digital core model for complex carbonate reservoirs. Gas Science Bulletin, 2022, 3(52): 78-89. (in Russian).

Portal, D. P. M. Estaillades carbonate micro-CT dataset (3.6 μm voxel size), 2020a.

Portal, D. P. M. Portal, D. P. M. Estaillades carbonate micro-CT dataset (3.1 μm voxel size), 2020b.

Ruspini, L. C., Øren, P. E., Berg, S., et al. Multiscale digital rock analysis for complex rocks. Transport in Porous Media, 2021, 139(2): 301-325.

Spurin, C., Callas, C., Darraj, N., et al. The importance and challenges associated with multi-scale heterogeneity for geological storage. InterPore Journal, 2025, 2(1): IPJ260225-2.

Vidal, A. D., Neta, A. P. M., de Castro Vargas Fernandes, J., et al. Multi-resolution X-ray micro-computed tomography images of carbonate rocks from brazilian pre-salt. Scientific Data, 2024, 11(1): 1361.

Wu, L., Zhuang, J., Chen, H. Voco: A simple-yet-effective volume contrastive learning framework for 3D medical image analysis. Paper Presented at IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16-22 June, 2024.

Zheng, J., Zheng, L., Liu, H., et al. Relationships between permeability, porosity and effective stress for low-permeability sedimentary rock. International Journal of Rock Mechanics and Mining Sciences, 2015, 78: 304-318.

Zhou, Y., Nie, X. Identification and parameter characterization of pores and fractures in shales based on multi-scale digital core data. Paper Presented at Asia Petroleum Geoscience Conference and Exhibition 2024, Kuala Lumpur, Malaysia, 20-21 November, 2024.

Zhou, Z., Sodha, V., Rahman Siddiquee, M. M., et al. Models genesis: Generic autodidactic models for 3D medical image analysis. Paper Presented at International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13-19 October, 2019.

Zou, C., Zhu, R., Liu, K., et al. Tight gas sandstone reservoirs in china: Characteristics and recognition criteria. Journal of Petroleum Science and Engineering, 2012, 88: 82-91.