

A Comparative Benchmark of GAN, DDPM, and Flow Matching for 2D and 3D Porous Microstructures Synthesis.

Ali Aouf

ali.aouf@uclouvain.com

UCLouvain

Eric Laloy

Belgian Nuclear Research Centre

Christophe De Vleeschouwer

UCLouvain

Research Article

Keywords: 2D Microstructure Synthesis, 3D Microstructure Synthesis, Porous Media, Subsurface structures, Deep Generative Models, Denoising Diffusion Models, Generative Adversarial Networks, Flow Matching, Digital rock physics

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11001828/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Highlights

A Comparative Benchmark of GAN, DDPM, and Flow Matching for 2D and 3D Porous Microstructures Synthesis.

Ali Aouf, Eric Laloy, Christophe De Vleeschouwer

- We benchmark WGAN-GP, DDPM, and flow matching for porous microstructure synthesis using 2D and supervised 3D datasets of clay, carbonate, and sandstone.
- Model fidelity is assessed using phase statistics, microstructure-relevant structural descriptors, and visual comparison, with effective diffusivity additionally evaluated for the 3D benchmark.
- Flow matching is evaluated as an alternative generative framework for both 2D porous-media images and 3D volumetric microstructures.
- Flow matching provides a consistent quality–cost compromise across the benchmark, reproducing the 3D porosity–diffusivity relation particularly well while requiring substantially fewer sampling steps than DDPM.

A Comparative Benchmark of GAN, DDPM, and Flow Matching for 2D and 3D Porous Microstructures Synthesis.

Ali Aouf^a, Eric Laloy^b and Christophe De Vleeschouwer^c

^aali.aouf@sckcen.be — ali.aouf@uclouvain.be

^beric.laloy@sckcen.be

^cchristophe.devleeschouwer@uclouvain.be

ARTICLE INFO

Keywords:

2D Microstructure Synthesis
3D Microstructure Synthesis
Porous Media
Subsurface structures
Deep Generative Models
Denoising Diffusion Models
Generative Adversarial Networks
Flow Matching
Digital rock physics

ABSTRACT

Accurate representations of porous microstructures are essential for studying transport and effective properties in geomaterials such as clay, carbonate, and sandstone. Deep generative models provide a means of producing additional statistically representative microstructures, but different model families can differ substantially in their ability to reproduce morphology, population variability, and transport-relevant connectivity. In this work, we present a comparative benchmark of three generative-model families for porous microstructure synthesis: a Wasserstein generative adversarial network with gradient penalty (WGAN-GP), denoising diffusion probabilistic models (DDPMs), and flow matching models. The benchmark considers six microstructure datasets in two dimensionally consistent settings: two-dimensional models trained on 2D images and supervised three-dimensional models trained on 3D volumes. The comparison therefore addresses 2D-to-2D and 3D-to-3D synthesis rather than dimensional expansion from 2D to 3D. Model performance is assessed using visual comparison, phase statistics, and microstructure-relevant structural descriptors, with effective diffusivity additionally evaluated for the three-dimensional benchmark. Across the considered datasets, all three model families reproduce important aspects of the reference microstructures, although their relative performance depends on the quantity being assessed. Flow matching is comparatively consistent across the reported structural measures and, in three dimensions, reproduces the reference porosity–diffusivity relation most consistently. It also requires substantially fewer sequential sampling evaluations than the 1,000-step DDPM configuration considered here, while WGAN-GP retains the advantage of single-pass generation. Overall, the results identify flow matching as a competitive and computationally attractive approach for porous-microstructure synthesis, providing a favorable quality–cost compromise without universally outperforming the alternative models across every material and metric.

CRedit authorship contribution statement

Ali Aouf: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft. **Eric Laloy:** Conceptualization, Methodology, Formal analysis, Writing - review and editing. **Christophe De Vleeschouwer:** Conceptualization, Methodology, Formal analysis, Writing - review and editing.

1. Introduction

Modelling the behaviour of porous materials requires realistic representations of their microstructure. In clays, carbonates, and sandstones, pore geometry, phase arrangement, and connectivity govern effective transport properties and therefore influence the outcome of pore-scale simulations used in digital rock physics. Reliable microstructure representations are consequently required not only for visual reconstruction, but also for statistical and physics-based analyses.

Three-dimensional microstructure data are commonly obtained using imaging techniques such as X-ray computed

ORCID(s): 0009-0004-1820-9707 (.A. Aouf)

tomography (CT) and focused ion beam scanning electron microscopy (FIB-SEM) Zhang et al. (2020); Ibrahim et al. (2021). These techniques require substantial experimental effort and involve trade-offs between spatial resolution, field of view, acquisition time, and sample preparation Blunt et al. (2013); Gaboreau et al. (2016); Sezen (2016). For fine-scale porous media and multiphase materials, segmentation can introduce an additional source of uncertainty and can affect the structural statistics and transport properties derived from the images Salzer et al. (2014); Andrew (2018). Large ensembles of representative three-dimensional volumes are therefore difficult to obtain for many materials of interest.

Synthetic microstructure generation complements direct imaging by providing additional realizations that can be used for statistical analysis, uncertainty quantification, and physics-based simulation Torquato and Haslach (2002); Blunt et al. (2013). Classical reconstruction approaches, including simulated annealing and multiple-point statistics, have been used extensively, but their computational cost and ability to reproduce complex heterogeneous structures can become limiting at high resolution Yeong and Torquato (1998); Strebel (2002); Mariethoz et al. (2010); Aouf et al. (2025b). These limitations have motivated the use of deep generative models for porous-media and microstructure synthesis Rabbani et al. (2021); Lukas Mosser (2017); Aouf et al. (2025b); Laloy et al. (2018).

Generative adversarial networks (GANs) learn a data distribution through adversarial optimisation between a generator and a discriminator and have been applied to porous-media synthesis in both two and three dimensions Goodfellow et al. (2014); Lukas Mosser (2017). Wasserstein formulations modify the adversarial objective to improve optimisation behaviour, and the gradient-penalty variant provides a practical way to regularise the critic while avoiding weight clipping Arjovsky et al. (2017); Gulrajani et al. (2017). Adversarial models retain the important advantage of single-pass sampling, but their observed performance can depend strongly on training dynamics and checkpoint selection.

Denoising diffusion probabilistic models (DDPMs) provide an iterative alternative based on reversing a gradual noising process Ho et al. (2020); Nichol and Dhariwal (2021). Diffusion models have shown strong performance for porous microstructures, including clay and multiphase systems Aouf et al. (2025a,b). Their standard sampling procedure, however, requires many sequential network evaluations, which can become costly for large two-dimensional fields and especially for three-dimensional volumes Ho et al. (2020); Nichol and Dhariwal (2021).

Flow matching (FM) is a continuous-time generative formulation in which a neural network learns a velocity field that transports samples from a simple reference distribution to the data distribution Lipman et al. (2023); Holderrieth and Erives (2025). Like diffusion, the standard training objective is non-adversarial; unlike a long discrete diffusion chain, generation is performed by numerically integrating a learned ordinary differential equation, so the number of sampling evaluations can be substantially smaller when the learned trajectory is sufficiently regular Lipman et al. (2023); Albergo et al. (2025). Flow matching has received comparatively little attention in porous-media synthesis, particularly for volumetric microstructure generation.

In this work, we benchmark WGAN-GP, DDPM, and flow matching in two dimensionally consistent settings: two-dimensional models trained on two-dimensional images and supervised three-dimensional models trained on three-dimensional volumes. The benchmark retains six microstructure datasets spanning binary and multiphase illite, Savonnières carbonate, Boom Clay, Bandera Brown sandstone, and Bentheimer sandstone. The comparison is based on phase statistics, microstructure-relevant structural descriptors, visual assessment, and, for the three-dimensional benchmark, lattice-Boltzmann effective diffusivity. These quantities are interpreted jointly rather than collapsed into a single visual or statistical score, because they probe different aspects of the generated populations.

The remainder of the paper is organized as follows. Section 2 introduces the three generative model families and the experimental framework. Section 4 presents the two-dimensional and three-dimensional benchmark results and compares their structural, physical, and computational trade-offs. Section 6 summarizes the conclusions and limitations of the study.

2. Generative models

This section summarizes the three model families used in the benchmark. The emphasis is on the generation mechanism needed to interpret the comparison; dataset construction, phase encoding, sampling configurations, and evaluation procedures are reported separately below.

2.0.1. Generative Adversarial Networks (WGAN-GP)

Generative adversarial networks (GANs) formulate generation as an adversarial optimisation problem between a generator and a discriminator Goodfellow et al. (2014). The generator maps samples from a latent prior to synthetic microstructures, while the discriminator is trained to distinguish generated samples from real data. In the Wasserstein formulation, the discriminator is replaced by a real-valued critic and the objective is related to the Wasserstein distance between the generated and reference distributions Arjovsky et al. (2017).

The Wasserstein GAN with gradient penalty (WGAN-GP) regularises the critic by penalising deviations of its input-gradient norm from unity Gulrajani et al. (2017). When written as a loss minimised with respect to the critic parameters, the objective is

$$\mathcal{L}_D = \mathbb{E}_{z \sim P_z} [D(G(z))] - \mathbb{E}_{x \sim P_r} [D(x)] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{x}}} (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2, \quad (1)$$

where D denotes the critic, G the generator, P_r the real-data distribution, P_z the latent prior, and $P_{\hat{x}}$ the distribution of points interpolated between real and generated samples. The coefficient λ controls the strength of the gradient penalty. The corresponding generator is trained to increase the critic score assigned to generated samples.

WGAN-GP is used here as the single-pass adversarial baseline. Its main computational advantage is that, once

trained, each sample is produced by one generator evaluation. The benchmark evaluates its empirical training behaviour and generated statistics directly rather than assuming that the adversarial objective determines performance by itself.

2.0.2. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) generate samples by learning to reverse a gradual Gaussian noising process Ho et al. (2020). The forward process is a Markov chain

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (2)$$

where β_t specifies the noise level at timestep t . A time-conditioned neural network, implemented using a U-Net-style architecture, is trained to predict the noise component at randomly sampled timesteps Ronneberger et al. (2015); Ho et al. (2020). Generation starts from Gaussian noise and applies the learned reverse transitions sequentially until a data-space sample is obtained.

The standard DDPM objective is non-adversarial and has been associated with stable optimisation and strong reproduction of complex image statistics Nichol and Dhariwal (2021); Aouf et al. (2025a). Its principal practical cost is iterative sampling: a standard reverse chain can require many network evaluations for each generated sample. Accelerated samplers exist, but they are not the focus of the present benchmark Nichol and Dhariwal (2021); Karras et al. (2022).

2.0.3. Flow Matching

Flow matching learns a continuous-time transformation from a simple reference distribution, typically Gaussian noise, to the data distribution Lipman et al. (2023). Let $\mathbf{x}_0$ denote a noise sample and $\mathbf{x}_1$ a data sample. For the linear conditional path

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \quad t \in [0, 1], \quad (3)$$

the target conditional velocity is $\mathbf{x}_1 - \mathbf{x}_0$. A neural network $\mathbf{v}_\theta(\mathbf{x}_t, t)$ is trained by regression to match this velocity field.

At inference, the learned ordinary differential equation

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}_\theta(\mathbf{x}, t) \quad (4)$$

is integrated from a Gaussian sample at $t = 0$ to a generated data sample at $t = 1$ Lipman et al. (2023).

The velocity field is represented using time-conditioned convolutional U-Net-style networks for both image and volumetric data. The number of numerical integration steps used at inference is not fixed by the training objective. This

provides a direct accuracy–cost trade-off: sufficiently regular learned trajectories can be integrated using fewer network evaluations than a long standard diffusion chain, although the resulting accuracy still depends on the integration scheme and step count Albergo et al. (2025).

3. Experimental evaluation

3.1. Datasets and preprocessing

The benchmark contains six microstructure datasets spanning different lithologies, phase compositions, and levels of structural complexity. The sandstone data comprise Bentheimer and Bandera Brown sandstone from the Digital Rocks Portal Neumann et al. (2021). Savonnières carbonate, also obtained from the Digital Rocks Portal Bultreys (2016), provides a more heterogeneous pore geometry. Clay microstructures are represented by binary illite, three-phase illite, and Boom Clay data provided by SCK CEN Gaboreau et al. (2016); Honty et al. (2021); Zeelmaekers et al. (2015); Yu et al. (2013); Hemes et al. (2013). The multiphase illite dataset is included to test whether the model comparison extends beyond binary pore–solid segmentation.

For two-dimensional benchmarking, 5,000 crops of size 256^2 pixels are extracted from the available reference data for each dataset. For three-dimensional benchmarking, 2,500 sub-volumes of size 64^3 voxels are extracted for each dataset. The same reference phase definitions are used when comparing the three model families within an experiment.

Binary data are represented in one scalar channel, with pore encoded as -1 and solid as $+1$. The three phases of the multiphase illite data are also represented in one scalar channel, with pore encoded as -1 , clay matrix as 0 , and mineral as $+1$. The same scalar phase ordering is used in both the two-dimensional and three-dimensional benchmarks reported here. Generated outputs are mapped back to the corresponding discrete phases before phase statistics, structural descriptors, and transport quantities are evaluated.

Reported sampling configurations. For the sampling-cost comparison reported in Section 4, the DDPM uses 1,000 reverse timesteps and flow matching uses 100 numerical integration steps. WGAN-GP generates each sample in one forward pass. These are the sampling configurations used for the reported timing comparison; the timing values should therefore be interpreted for that configuration rather than as architecture-independent constants.

3.2. Evaluation methods

The models are evaluated using complementary measures of phase composition, spatial organisation, and three-dimensional transport. Visualisations are used to identify qualitative features and artefacts, but are not treated as sufficient evidence for ranking the model families.

Phase composition is evaluated first. For binary datasets, the empirical porosity distributions of generated crops are compared with the reference distribution. For multiphase illite, the three phase-fraction distributions are reported

separately. Both the location and width of these distributions are relevant: agreement in mean or median composition does not by itself establish that the generated population reproduces the variability of the reference data.

Morphological fidelity is evaluated using the two-point probability function $S_2(r)$, lineal-path function $L(r)$, chord-length distribution, and pore-size distribution Torquato and Haslach (2002). The two-point probability function measures the probability that two points separated by r belong to the same phase. The lineal-path function measures the probability that a complete segment of length r lies within a phase and therefore probes uninterrupted phase continuity. Chord-length distributions characterize contiguous phase-segment lengths, while pore-size distributions characterize the range of local pore scales. These descriptors are interpreted together because no single structural function provides a complete measure of microstructure fidelity or connectivity.

Structural descriptors are computed consistently for generated and reference data using PoreSpy Gostick et al. (2019). For two-dimensional images, directional evaluations are performed along the horizontal and vertical directions; for three-dimensional volumes, they are performed along the three orthogonal axes. Phase statistics are considered before the descriptor curves because several structural quantities, including S_2 , depend directly on phase fraction.

For the three-dimensional benchmark, functional relevance is assessed using effective gas diffusivity calculated with Yantra, the lattice-Boltzmann implementation developed at SCK CEN Patel et al. (2018). All compared samples are processed using the same numerical settings. The simulated effective diffusivity D_{eff} is interpreted comparatively rather than as a direct prediction of an experimental bulk coefficient. In particular, the generated porosity–diffusivity relation is compared with the reference population: differences at similar porosity provide evidence of connectivity differences that cannot be explained by pore fraction alone.

Finally, the benchmark should be interpreted as a test of consistency with the available acquisitions. Because multiple training and evaluation crops can originate from the same parent scan, the results do not by themselves establish generalisation to an independent physical specimen of the same material. This limitation applies to both the two-dimensional and three-dimensional comparisons.

4. Results

This section presents the WGAN-GP, DDPM, and flow-matching benchmark results separately for two-dimensional images and three-dimensional volumes. The comparison combines visual assessment with empirical phase statistics and structural descriptors; the three-dimensional benchmark additionally includes effective diffusivity. These quantities are interpreted as complementary evidence rather than combined into a single aggregate score.

4.1. Two-dimensional benchmark

Figure 1 compares generated 2D samples across the six benchmark datasets. All three model families reproduce the principal material-specific motifs at the displayed scale, including dispersed pore structures in illite, coarser connected void space in the sandstones, and the stronger heterogeneity of Savonnières carbonate. The differences are more local. WGAN-GP appears smoother than the reference in several cases; on Boom Clay, for example, the generated features are more rounded and less sharply defined. DDPM and flow matching retain more high-frequency texture, although local seams or block-like artefacts remain visible in some datasets. These images provide qualitative context but are not used on their own to rank the models.

Phase statistics are summarized in Figure 2. For the five binary datasets—binary illite, Savonnières carbonate, Boom Clay, Bandera Brown sandstone, and Bentheimer sandstone—the boxplots report crop-wise porosity distributions. WGAN-GP and flow matching generally place the median and interquartile range near the reference distributions, whereas DDPM more often produces broader distributions and more extreme outliers. The direction and magnitude of the DDPM shift are material dependent, including higher median porosity for Boom Clay and Bentheimer in the reported results. The phase statistics are therefore treated as a composition check before the spatial descriptors are interpreted.

For multiphase illite, Figure 2(f) reports the three phase-fraction distributions separately. The models reproduce the reference phase proportions to different degrees, and WGAN-GP shows a near-degenerate distribution for one phase in the reported experiment. This is a distinct population-level discrepancy and should be considered when interpreting the multiphase structural descriptors.

Figure 3 compares chord-length distributions, lineal-path functions, pore-size distributions, and two-point correlation functions. The mean curves of all three generators follow the reference trends for many material-descriptor combinations, while the population bands differ more strongly. DDPM frequently shows the broadest variability, particularly for pore-size and two-point statistics, whereas WGAN-GP and flow matching are more concentrated around their mean curves in several datasets. The WGAN-GP multiphase descriptors must, however, be read together with the phase-fraction discrepancy identified above. The 2D results therefore do not define a single model ordering: DDPM shows broader population variability in several cases, WGAN-GP has material-dependent artefacts and a specific multiphase composition failure, and flow matching is comparatively consistent across the reported phase and structural measures.

4.2. Three-dimensional benchmark

Figures 4 and 5 compare representative outputs from the three-dimensional benchmark across the six datasets. All three model families reproduce the dominant morphology visible at the displayed scale, including the dispersed

pore structure of illite, connected pore networks in the sandstones, and the pronounced heterogeneity of Savonnières carbonate. As in the 2D analysis, these visualisations are treated as qualitative context rather than as a sufficient ranking criterion.

The clearest visual differences occur at interfaces and in fine-scale texture. WGAN-GP exhibits local speckling and irregular pore–solid boundaries in several datasets, including binary illite and Bandera Brown sandstone, and its multiphase illite output appears more diffuse with less clearly separated phases. DDPM and flow matching generally preserve sharper interfaces and finer texture in the displayed examples. The three-dimensional comparison nevertheless requires the quantitative phase, structural, and transport results below, because visually plausible samples can still differ in population statistics or connectivity.

Phase statistics are summarized in Figure 6. For the five binary datasets, WGAN-GP and flow matching generally place the porosity medians near the reference values, while DDPM produces broader, material-dependent distributions with more pronounced outliers. Flow matching is comparatively narrow for some datasets. These statements refer specifically to the volumetric crop populations in Figure 6; the two-dimensional distributions are evaluated separately in Section 4.1.

For multiphase illite, Figure 6(f) shows the three volumetric phase-fraction distributions. The central phase proportions and the cross-sample spread differ between models. In particular, one phase has reduced variability for flow matching and DDPM in the reported experiment. This is interpreted as a difference in population spread rather than, by itself, as evidence of an incorrect mean phase proportion.

Figure 7 applies the same four descriptor families to the volumetric samples. The mean curves of the three generators follow the reference trends for many material–descriptor combinations, but the population spread and some material-specific deviations differ. DDPM generally shows the broadest uncertainty bands across several descriptors. Flow matching and WGAN-GP are more concentrated in many cases, although the WGAN-GP multiphase descriptors must be interpreted together with its phase-fraction and interface differences. As in two dimensions, the descriptor families provide complementary evidence and do not yield a universal ranking on their own.

Figure 8 compares effective diffusivity with porosity. All three generators recover the overall positive association between porosity and diffusivity, but they differ in the position and spread of the generated clouds relative to the reference. Flow matching follows the reference porosity–diffusivity relation most consistently across the reported datasets, whereas WGAN-GP and DDPM show larger material-dependent scatter or offsets, including deviations for illite and the sandstone datasets. Because this comparison conditions the interpretation of transport on porosity, it provides information about pore connectivity that is not contained in the phase statistics alone and is the clearest quantitative distinction between the model families in the 3D benchmark.

Taken together, the volumetric results again show that the model ordering depends on the assessed quantity. DDPM

reproduces many mean structural trends but often with broader sample-to-sample variation; WGAN-GP retains fast generation but shows material-dependent interface, phase, and transport discrepancies; and flow matching provides the most consistent agreement with the reference porosity–diffusivity relation in this experiment.

The reported three-dimensional training runs also differed in their observed dynamics. Flow matching produced stable improvements relatively early, whereas DDPM required more training before comparably well-defined structures emerged. WGAN-GP checkpoints varied more strongly over training. For the WGAN-GP comparison reported here, a post-training checkpoint was selected by visual inspection to reduce obvious artefacts, and a mild median filter was applied to mitigate residual speckle. The iterative models were evaluated without the same intervention. This asymmetry is a limitation of the comparison and is not used as evidence for an intrinsic difference between adversarial and iterative model families.

5. Discussion

Beyond reconstruction quality, the reported models differ substantially in sampling cost. Under the benchmark configuration used for the timing comparison, generating 100 samples required approximately 55 minutes for the 1,000-step DDPM, approximately 5 minutes for flow matching with 100 integration steps, and approximately 1 minute for WGAN-GP. These measurements reflect the specific implementation, hardware, and sampling configurations used here rather than universal model-family constants. The robust distinction is the number of sequential network evaluations: WGAN-GP is single-pass, while DDPM and flow matching are iterative, with flow matching using substantially fewer evaluations than the DDPM in the reported comparison.

The benchmark therefore does not support a universal ordering across every metric. Flow matching provides the most consistent overall compromise in the reported experiments because it combines close agreement with the reference porosity–diffusivity relation with substantially fewer sampling evaluations than the DDPM. WGAN-GP retains the clear advantage of single-pass inference, while DDPM remains a competitive non-adversarial baseline but incurs the highest sampling cost in the configuration studied here. These conclusions are specific to the evaluated architectures, datasets, and processing choices.

6. Conclusion

This work compared WGAN-GP, DDPM, and flow matching for dimensionally consistent porous-microstructure synthesis in two dimensions and supervised three dimensions. Across the six benchmark datasets, all three model families reproduced important aspects of the reference morphology, but the relative ordering depended on the quantity being assessed. Phase statistics, structural descriptors, qualitative appearance, and effective diffusivity did not provide identical rankings.

Table 1

Summary of the main advantages, limitations, and practical implications of the evaluated generative model families for porous microstructure synthesis.

Model	Main advantages	Main limitations	Practical implication
WGAN-GP	Fast sampling; single-pass generation; comparatively simple inference once trained.	Strong checkpoint sensitivity in the reported runs; local artefacts and material-dependent phase, structural, or transport deviations; mild post-processing was used in the reported 3D comparison.	Lowest inference cost, but training and checkpoint selection require additional care in this benchmark.
DDPM	Non-adversarial training; good visual quality; mean structural trends close to the reference in many cases.	Broad population distributions in several datasets; slow sampling with the 1,000-step configuration; highest inference cost of the three reported models.	Provides a strong iterative baseline when the additional sampling cost is acceptable.
Flow matching	Non-adversarial training; adjustable inference step count; lower sampling cost than the reported DDPM; most consistent reference porosity–diffusivity relation in the 3D benchmark.	Still requires iterative ODE integration; accuracy and cost depend on the numerical solver and number of sampling steps.	Provides the most consistent quality–cost compromise in the present benchmark, without dominating every individual metric.

In the two-dimensional benchmark, DDPM generally produced the broadest population distributions, while WGAN-GP showed material-dependent artefacts and a pronounced phase-fraction discrepancy in the multiphase illite case. Flow matching was comparatively consistent across phase and structural measures, although no single two-dimensional descriptor was sufficient to define a universal model ranking. In three dimensions, the porosity–diffusivity relation provided the clearest functional distinction: flow-matching samples followed the reference trend most consistently, whereas WGAN-GP and DDPM showed larger material-dependent scatter or offsets in parts of the benchmark.

The model families also differ substantially in sampling cost. WGAN-GP retains the advantage of single-pass generation. Under the reported sampling configuration, flow matching required markedly fewer network evaluations and less generation time than the 1,000-step DDPM while maintaining comparable or better agreement with the reference transport trend. The WGAN-GP comparison additionally involved checkpoint selection by visual inspection and mild median filtering in the reported three-dimensional experiment; this asymmetry is a limitation of the benchmark and should not be interpreted as evidence for an intrinsic difference between all adversarial and iterative generators.

Taken together, the results support flow matching as the most consistent compromise in this benchmark rather than as a universal winner across every metric and material. WGAN-GP remains attractive when single-pass sampling is the dominant requirement, while DDPM provides a non-adversarial iterative baseline at substantially higher sampling cost in the configuration studied here. Because the benchmark is based on crops from the available parent acquisitions, these conclusions establish consistency with those acquisitions rather than generalisation to independent specimens.

Code availability section

All models are trained on an NVIDIA A100 Tensor Core GPU, the software packages used are:

- Source code: <https://github.com/40uf411/modelsBench>
- Python 3.10 as a programming language.
- Training time of about 9 hours on average.
- PyTorch for models implementation, training, and sampling.
- Numpy for samples representation.
- scikit-image and OpenCV for image processing.
- ParaView for 3D data
- CherryMSA for samples analysis and visualization.

Funding

The authors received no specific funding for this work.

Ethics and Consent to Participate

Not applicable.

Declarations

Christophe De Vleeschouwer is a Research Director of the Fonds de la Recherche Scientifique – FNRS.

References

- Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., 2025. Stochastic interpolants: A unifying framework for flows and diffusions. URL: <https://arxiv.org/abs/2303.08797>, arXiv:2303.08797.
- Andrew, M., 2018. A quantified study of segmentation techniques on synthetic geological xrm and fib-sem images. Computational Geosciences 22, 1503–1512. URL: <https://doi.org/10.1007/s10596-018-9768-y>, doi:10.1007/s10596-018-9768-y.
- Aouf, A., Laloy, E., Rogiers, B., De Vleeschouwer, C., 2025a. 3d clay microstructure synthesis using denoising diffusion probabilistic models. Applied Computing and Geosciences 26, 100248. URL: <https://www.sciencedirect.com/science/article/pii/S2590197425000308>, doi:<https://doi.org/10.1016/j.acags.2025.100248>.
- Aouf, A., Laloy, E., Rogiers, B., et al., 2025b. Reconstruction of 3d multiphase microstructures using denoising diffusion models. Research Square URL: <https://doi.org/10.21203/rs.3.rs-7448759/v1>, doi:10.21203/rs.3.rs-7448759/v1. preprint (Version 1).

- 361 Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein gan. URL: <https://arxiv.org/abs/1701.07875>, arXiv:1701.07875.
- 362 Blunt, M.J., Bijeljic, B., Dong, H., Gharbi, O., Iglauer, S., Mostaghimi, P., Paluszny, A., Pentland, C., 2013. Pore-scale imaging and modelling. *Advances in Water Resources* 51, 197–216. URL: <https://www.sciencedirect.com/science/article/pii/S0309170812000528>, doi:<https://doi.org/10.1016/j.advwatres.2012.03.003>. 35th Year Anniversary Issue.
- 363
- 364
- 365 Bultreys, T., 2016. Savonnières carbonate. Digital Porous Media Portal. URL: <https://doi.org/10.17612/P7W88K>, doi:10.17612/P7W88K.
- 366 dataset.
- 367 Gaboreau, S., Robinet, J.C., Prêt, D., 2016. Optimization of pore-network characterization of a compacted clay material by tem and fib/sem imaging. *Microporous and Mesoporous Materials* 224, 116–128. URL: <https://www.sciencedirect.com/science/article/pii/S138718111500640X>, doi:<https://doi.org/10.1016/j.micromeso.2015.11.035>.
- 368
- 369
- 370 Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. URL: <https://arxiv.org/abs/1406.2661>, arXiv:1406.2661.
- 371
- 372 Gostick, J., Khan, Z.A., Tranter, T.G., Kok, M.D.R., Agnaou, M., Sadeghi, M., Jervis, R., 2019. Porespy: A python toolkit for quantitative analysis of porous media images. *Journal of Open Source Software* 4, 1296. URL: <https://doi.org/10.21105/joss.01296>, doi:10.21105/joss.01296.
- 373
- 374
- 375 Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein gans. URL: <https://arxiv.org/abs/1704.00028>, arXiv:1704.00028.
- 376
- 377 Hemes, S., Desbois, G., Urai, J., De Craen, M., Honty, M., 2013. Variations in the morphology of porosity in the boom clay formation: insights from 2d high resolution bib-sem imaging and mercury injection porosimetry. *Netherlands Journal of Geosciences - Geologie en Mijnbouw* 92, 275–300. doi:10.1017/S0016774600000214.
- 378
- 379
- 380 Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Red Hook, NY, USA.
- 381
- 382 Holderrieth, P., Erives, E., 2025. An introduction to flow matching and diffusion models. URL: <https://arxiv.org/abs/2506.02070>, arXiv:2506.02070.
- 383
- 384 Honty, M., Fredericks, L., Wang, L., De, C., Thomas, P., Moors, H., Jacobs, E., 2021. Boom clay pore water geochemistry at the mol site: Experimental data as determined by in situ sampling of the piezometers. *Applied Geochemistry* 136, 105156. doi:10.1016/j.apgeochem.2021.105156.
- 385
- 386
- 387 Ibrahim, E.R., Jouini, M.S., Bouchaala, F., Gomes, J., 2021. Simulation and validation of porosity and permeability of synthetic and real rock models using three-dimensional printing and digital rock physics. *ACS Omega* 6, 31775–31781. URL: <https://doi.org/10.1021/acsomega.1c04429>, doi:10.1021/acsomega.1c04429.
- 388
- 389
- 390 Karras, T., Aittala, M., Aila, T., Laine, S., 2022. Elucidating the design space of diffusion-based generative models. URL: <https://arxiv.org/abs/2206.00364>, arXiv:2206.00364.
- 391
- 392 Laloy, E., Hérault, R., Jacques, D., Linde, N., 2018. Training-image based geostatistical inversion using a spatial generative adversarial neural network. *Water Resources Research* 54, 381–406. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017WR022148>, doi:<https://doi.org/10.1002/2017WR022148>, arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2017WR022148>.
- 393
- 394
- 395
- 396 Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M., 2023. Flow matching for generative modeling. URL: <https://arxiv.org/abs/2210.02747>, arXiv:2210.02747.
- 397
- 398 Lukas Mosser, Olivier Dubrule, M.J.B., 2017. Reconstruction of three-dimensional porous media using generative adversarial neural networks.

- 399 CoRR abs/1704.03225. URL: <http://arxiv.org/abs/1704.03225>, arXiv:1704.03225.
- 400 Mariethoz, G., Renard, P., Straubhaar, J., 2010. The direct sampling method to perform multiple-point geostatistical simulations. *Water Resources*
401 *Research* 46. URL: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR007621>, doi:[https://doi.org/10.](https://doi.org/10.1029/2008WR007621)
402 [1029/2008WR007621](https://doi.org/10.1029/2008WR007621), arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008WR007621>.
- 403 Neumann, R.F., Barsi-Andreeta, M., Lucas-Oliveira, E., Barbalho, H., Trevizan, W.A., Bonagamba, T.J., Steiner, M.B., 2021. High accuracy
404 capillary network representation in digital rock reveals permeability scaling functions. *Scientific Reports* 11. URL: [http://dx.doi.org/10.](http://dx.doi.org/10.1038/s41598-021-90090-0)
405 [1038/s41598-021-90090-0](http://dx.doi.org/10.1038/s41598-021-90090-0), doi:10.1038/s41598-021-90090-0.
- 406 Nichol, A.Q., Dhariwal, P., 2021. Improved denoising diffusion probabilistic models, in: Meila, M., Zhang, T. (Eds.), *Proceedings of the 38th*
407 *International Conference on Machine Learning*, PMLR. pp. 8162–8171. URL: <https://proceedings.mlr.press/v139/nichol21a.html>.
- 408 Patel, R.A., Perko, J., Jacques, D., De Schutter, G., Ye, G., Van Breugel, K., 2018. A three-dimensional lattice boltzmann method based reac-
409 tive transport model to simulate changes in cement paste microstructure due to calcium leaching. *Construction and Building Materials* 166,
410 158–170. URL: <https://www.sciencedirect.com/science/article/pii/S0950061818301363>, doi:[https://doi.org/10.1016/](https://doi.org/10.1016/j.conbuildmat.2018.01.114)
411 [j.conbuildmat.2018.01.114](https://doi.org/10.1016/j.conbuildmat.2018.01.114).
- 412 Rabbani, A., Fernando, A.M., Shams, R., Singh, A., Mostaghimi, P., Babaei, M., 2021. Review of data science trends and issues in
413 porous media research with a focus on image-based techniques. *Water Resources Research* 57, e2020WR029472. URL: [https:](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029472)
414 [/agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029472](https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029472), doi:<https://doi.org/10.1029/2020WR029472>,
415 arXiv:<https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020WR029472>. e2020WR029472 2020WR029472.
- 416 Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. URL: [https://arxiv.org/](https://arxiv.org/abs/1505.04597)
417 [abs/1505.04597](https://arxiv.org/abs/1505.04597), arXiv:1505.04597.
- 418 Salzer, M., Thiele, S., Zengerle, R., Schmidt, V., 2014. On the importance of fib-sem specific segmentation algorithms for porous media. *Mate-*
419 *rials Characterization* 95, 36–43. URL: <https://www.sciencedirect.com/science/article/pii/S1044580314001582>, doi:[https:](https://doi.org/10.1016/j.matchar.2014.05.014)
420 [/doi.org/10.1016/j.matchar.2014.05.014](https://doi.org/10.1016/j.matchar.2014.05.014).
- 421 Sezen, M., 2016. Focused ion beams (fib) — novel methodologies and recent applications for multidisciplinary sciences, in: Janecek, M., Kral, R.
422 (Eds.), *Modern Electron Microscopy in Physical and Life Sciences*. IntechOpen, Rijeka. chapter 6, pp. –. URL: [https://doi.org/10.5772/](https://doi.org/10.5772/61634)
423 [61634](https://doi.org/10.5772/61634), doi:10.5772/61634.
- 424 Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology* 34, 1–21. URL:
425 <https://doi.org/10.1023/A:1014009426274>, doi:10.1023/A:1014009426274.
- 426 Torquato, Haslach, 2002. Random Heterogeneous Materials: Microstructure and Macroscopic Properties. *Ap-*
427 *plied Mechanics Reviews* 55, B62–B63. URL: <https://doi.org/10.1115/1.1483342>, doi:10.1115/1.1483342,
428 arXiv:https://asmedigitalcollection.asme.org/appliedmechanicsreviews/article-pdf/55/4/B62/5439035/b61_1.pdf.
- 429 Yeong, C.L., Torquato, S., 1998. Reconstructing random media. *Physical review E* 57, 495.
- 430 Yu, L., Rogiers, B., Gedeon, M., Marivoet, J., Craen, M.D., Mallants, D., 2013. A critical review of laboratory and in-situ hydraulic conductivity
431 measurements for the boom clay in belgium. *Applied Clay Science* 75-76, 1–12. URL: [https://www.sciencedirect.com/science/](https://www.sciencedirect.com/science/article/pii/S0169131713000446)
432 [article/pii/S0169131713000446](https://www.sciencedirect.com/science/article/pii/S0169131713000446), doi:<https://doi.org/10.1016/j.clay.2013.02.018>.
- 433 Zeelmaekers, E., Honty, M., Derkowski, A., Środoń, J., De Craen, M., Vandenberghe, N., Adriaens, R., Ufer, K., Wouters, L., 2015. Qualitative
434 and quantitative mineralogical composition of the rupelian boom clay in belgium. *Clay Minerals* 50, 249–272. doi:10.1180/claymin.2015.
435 050.2.08.
- 436 Zhang, L., Qi, S., Ma, L., Guo, S., Li, Z., Li, G., Yang, J., Zou, Y., Li, T., Hou, X., 2020. Three-dimensional pore characterization of intact

437 loess and compacted loess with micron scale computed tomography and mercury intrusion porosimetry. Scientific Reports 10. URL: [https:](https://doi.org/10.1038/s41598-020-65302-8)
438 [//doi.org/10.1038/s41598-020-65302-8](https://doi.org/10.1038/s41598-020-65302-8), doi:10.1038/s41598-020-65302-8.

List of Figures

1	Two-dimensional qualitative comparison across datasets. Columns correspond to the six microstructures and rows correspond to the reference data (ground truth) and the three generative models (WGAN-GP, DDPM, and flow matching). Each cell contains a 4×4 grid of 2D crops.	16
2	Distribution of porosity and phase fractions across datasets and generative models. For binary microstructures, boxplots represent the porosity (phase 0 fraction). For multiphase illite, the distribution of phase fractions (Phase 0, Phase 1, Phase 2) is shown instead. Boxes denote interquartile range, whiskers represent $1.5 \times \text{IQR}$, and markers indicate outliers.	17
3	Comparison of structural descriptors for all investigated 2D microstructures. Columns correspond to chord length, lineal path function, pore size distribution, and two-point correlation function. Rows correspond to Illite (binary), Illite (multiphase), Savonnières carbonate, Boom clay, Bandera Brown sandstone, and Bentheimer sandstone. For the structural descriptor subplots, the x -axis represents the relevant length scale: chord length for the chord-length distribution, path length for the lineal-path function, pore radius for the pore-size distribution, and lag distance for the two-point correlation function. The y -axis shows the corresponding descriptor value: probability density for the chord-length distribution, probability for the lineal-path function, cumulative pore-volume fraction for the pore-size distribution, and two-point correlation probability for S_2	18
4	Qualitative comparison of views from the three-dimensional benchmark. Columns correspond to the six microstructure datasets and rows to the reference data and the three generative models (WGAN-GP, DDPM, and flow matching). Each cell contains representative views of the corresponding volumetric samples.	19
5	Additional qualitative comparison for the three-dimensional benchmark. Columns correspond to the six microstructure datasets and rows to the reference data and the three generative models (WGAN-GP, DDPM, and flow matching). The figure complements Figure 4 with additional views of the volumetric samples.	20
6	Distribution of porosity and phase fractions across datasets and generative models. For binary microstructures, boxplots represent the porosity. For multiphase illite, the distribution of phase fractions are shown instead. Boxes denote interquartile range, whiskers represent $1.5 \times \text{IQR}$, and markers indicate outliers.	21
7	Comparison of structural descriptors for all investigated 3D microstructures. Columns correspond to chord length, lineal path function, pore size distribution, and two-point correlation function. Rows correspond to Illite (binary), Illite (multiphase), Savonnières carbonate, Boom clay, Bandera Brown sandstone, and Bentheimer sandstone. For the structural descriptor subplots, the x -axis represents the relevant length scale: chord length for the chord-length distribution, path length for the lineal-path function, pore radius for the pore-size distribution, and lag distance for the two-point correlation function. The y -axis shows the corresponding descriptor value: probability density for the chord-length distribution, probability for the lineal-path function, cumulative pore-volume fraction for the pore-size distribution, and two-point correlation probability for S_2	22
8	Effective gas diffusivity as a function of porosity for 150 generated samples per model and dataset. Diffusivity is computed using lattice Boltzmann simulations on the processed binary volumes. Results are shown for six materials: (a) Binary Illite, (b) Multi-phase Illite, (c) Bandera Brown, (d) Bentheimer, (e) Boom Clay, and (f) Savonnières Carbonate. Markers correspond to different data sources: ground truth (black), WGAN-GP (red), DDPM (blue), and flow matching (green).	23

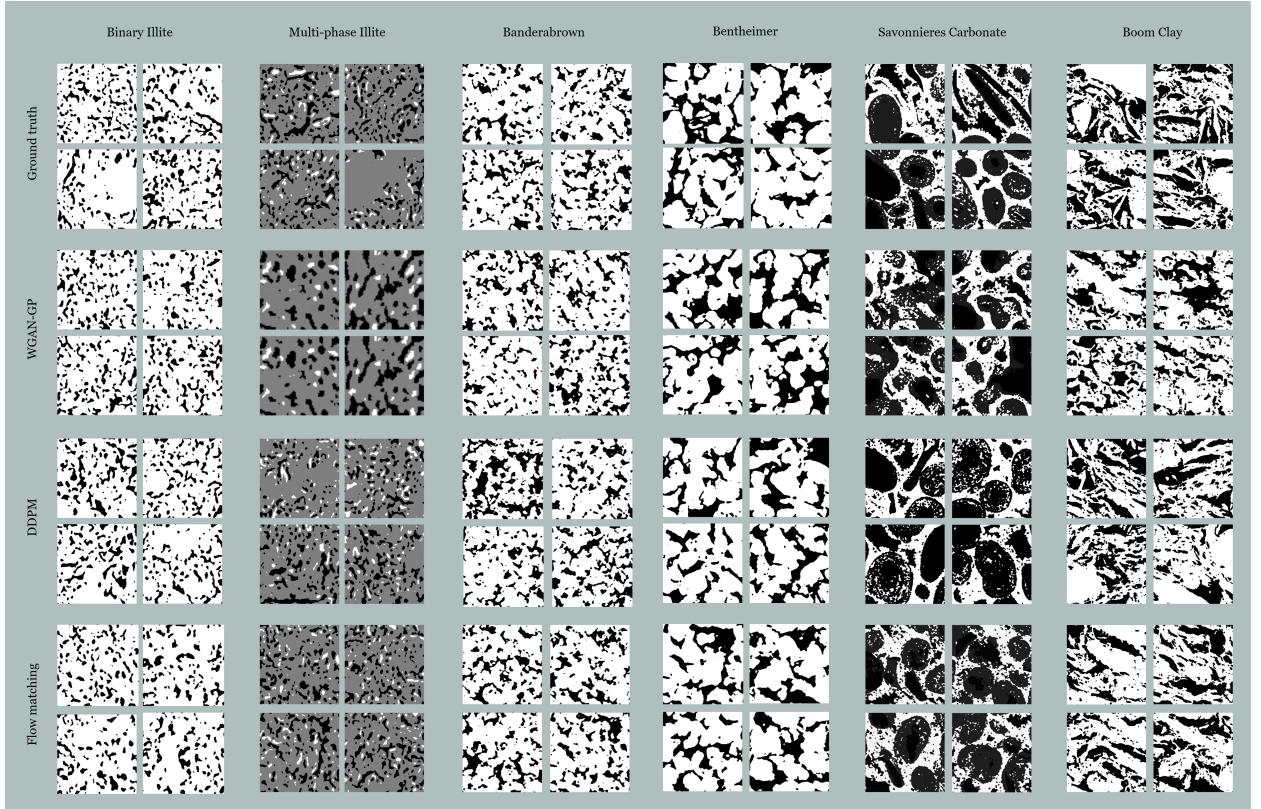


Figure 1: Two-dimensional qualitative comparison across datasets. Columns correspond to the six microstructures and rows correspond to the reference data (ground truth) and the three generative models (WGAN-GP, DDPM, and flow matching). Each cell contains a 4×4 grid of 2D crops.

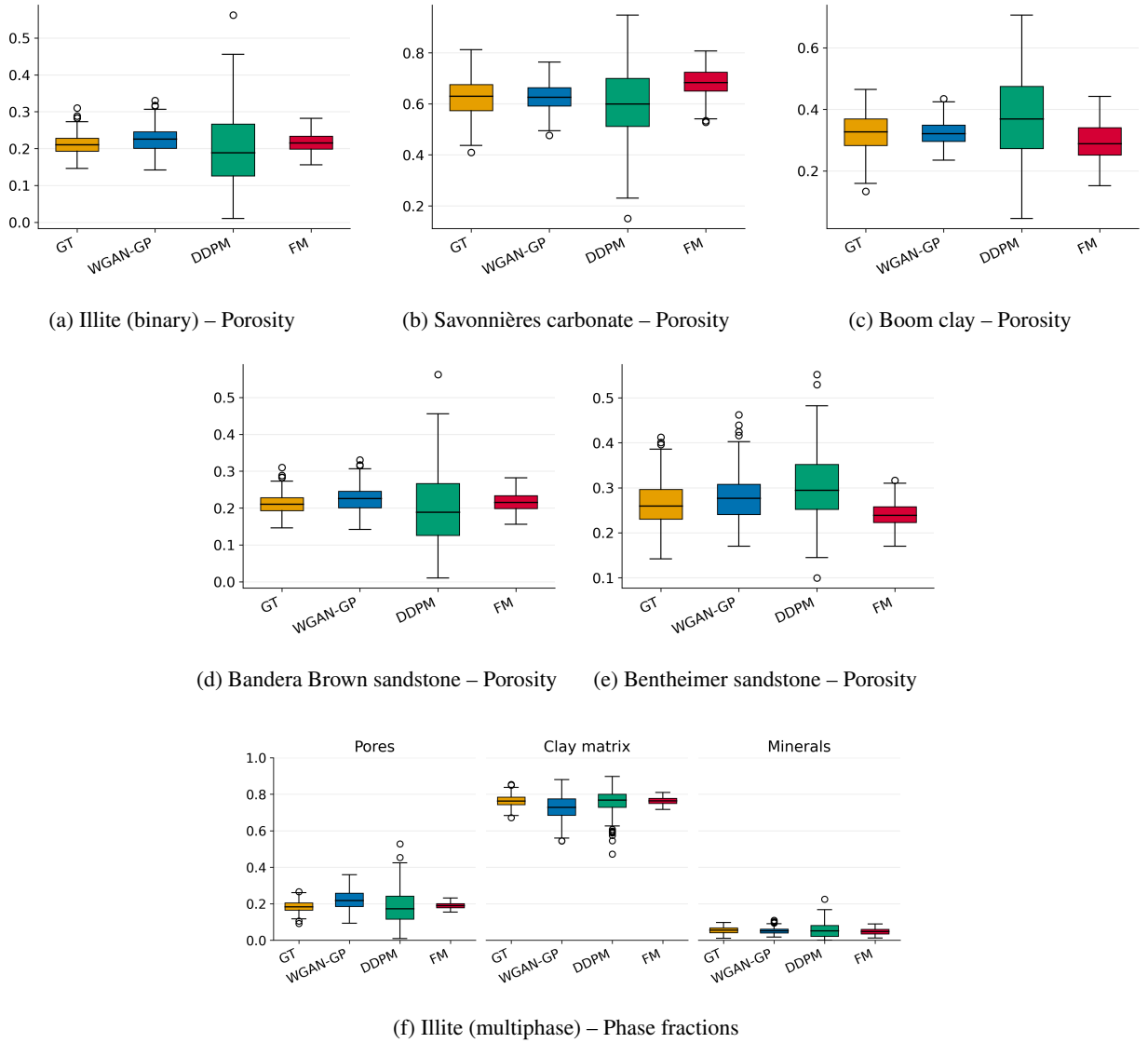


Figure 2: Distribution of porosity and phase fractions across datasets and generative models. For binary microstructures, boxplots represent the porosity (phase 0 fraction). For multiphase illite, the distribution of phase fractions (Phase 0, Phase 1, Phase 2) is shown instead. Boxes denote interquartile range, whiskers represent $1.5 \times \text{IQR}$, and markers indicate outliers.

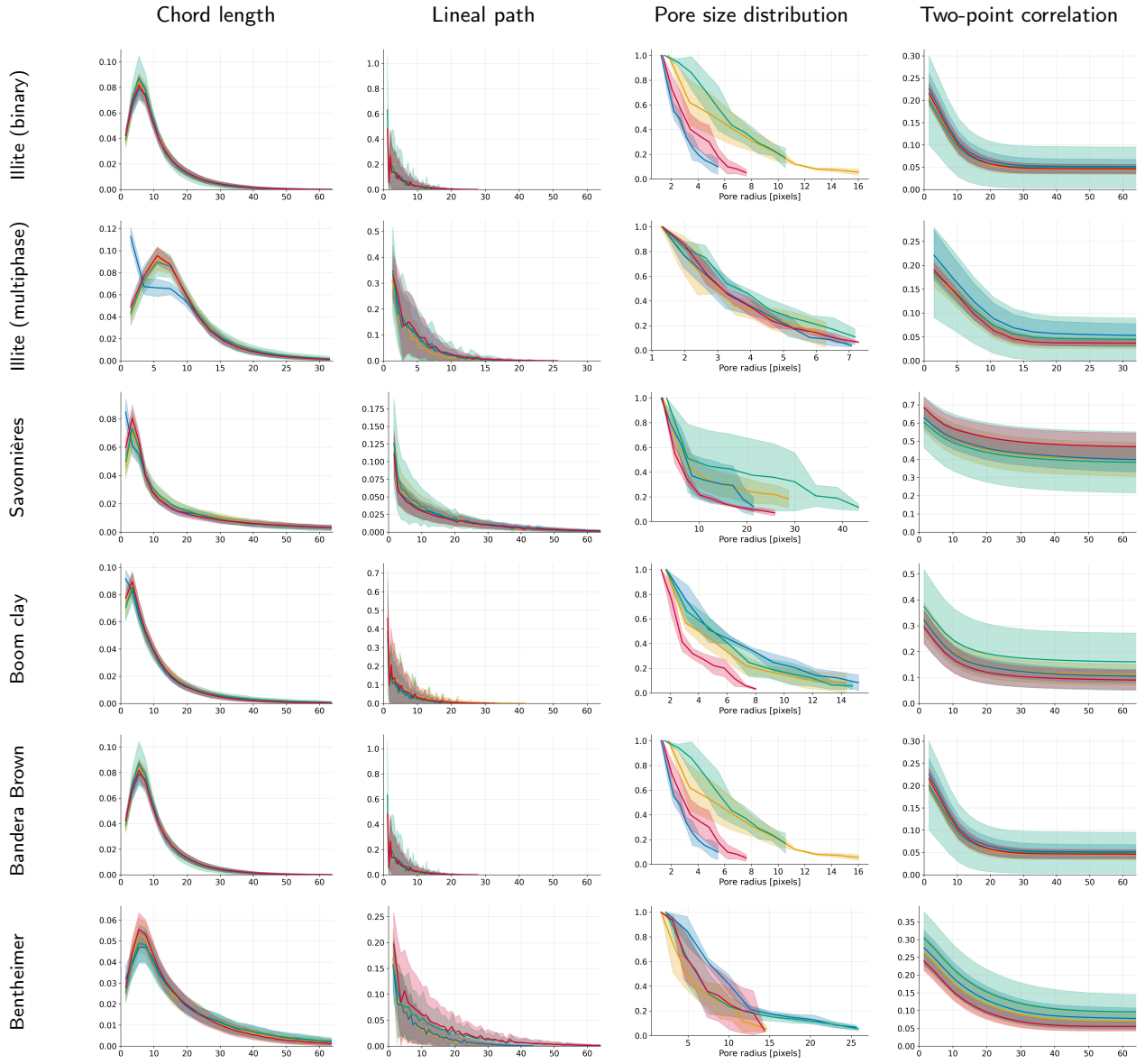


Figure 3: Comparison of structural descriptors for all investigated 2D microstructures. Columns correspond to chord length, lineal path function, pore size distribution, and two-point correlation function. Rows correspond to Illite (binary), Illite (multiphase), Savonnières carbonate, Boom clay, Bandera Brown sandstone, and Bentheimer sandstone. For the structural descriptor subplots, the x -axis represents the relevant length scale: chord length for the chord-length distribution, path length for the lineal-path function, pore radius for the pore-size distribution, and lag distance for the two-point correlation function. The y -axis shows the corresponding descriptor value: probability density for the chord-length distribution, probability for the lineal-path function, cumulative pore-volume fraction for the pore-size distribution, and two-point correlation probability for S_2 .

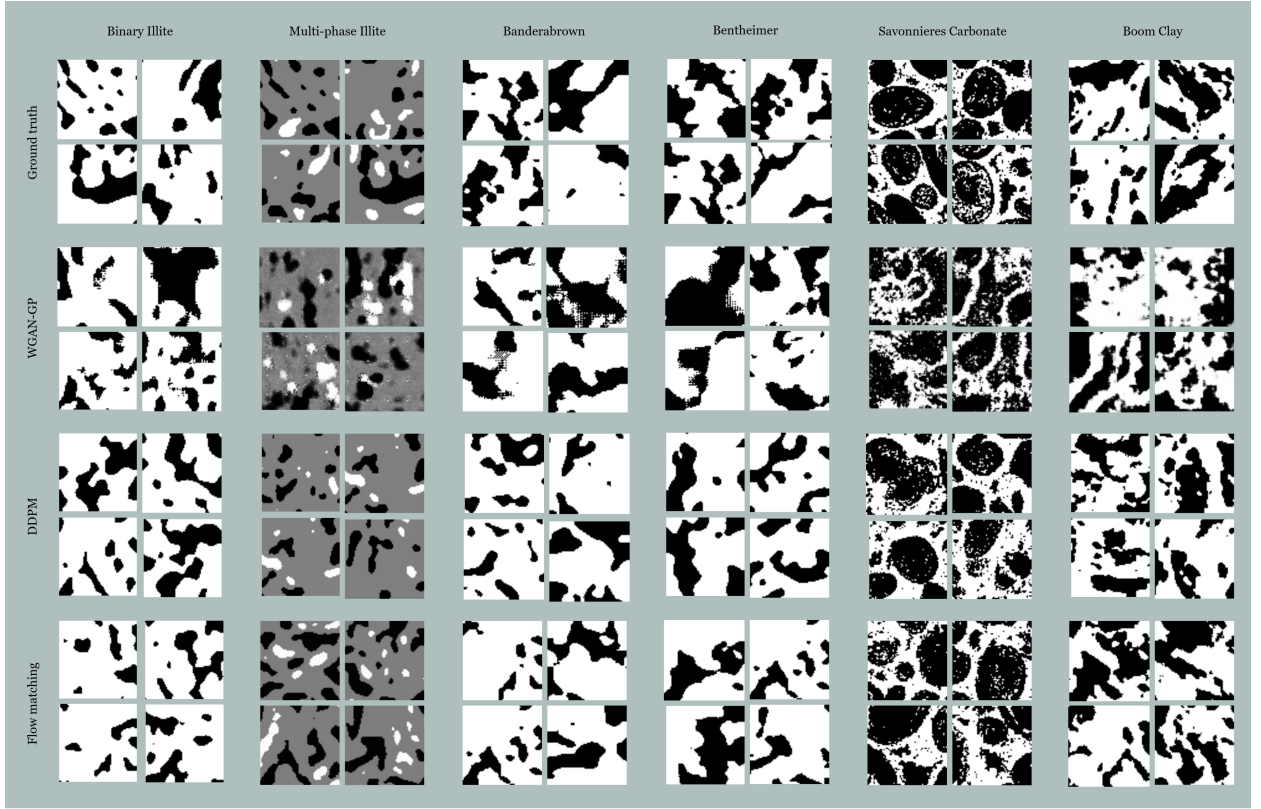


Figure 4: Qualitative comparison of views from the three-dimensional benchmark. Columns correspond to the six microstructure datasets and rows to the reference data and the three generative models (WGAN-GP, DDPM, and flow matching). Each cell contains representative views of the corresponding volumetric samples.

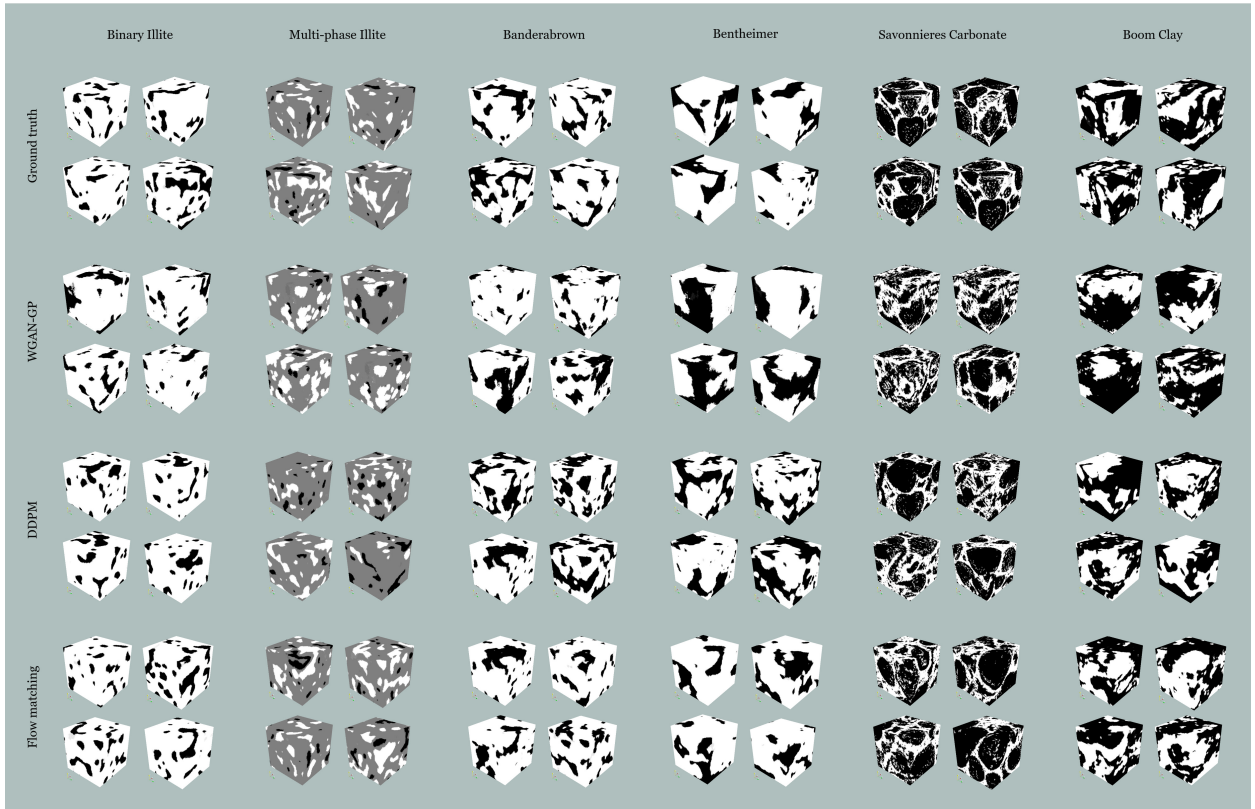


Figure 5: Additional qualitative comparison for the three-dimensional benchmark. Columns correspond to the six microstructure datasets and rows to the reference data and the three generative models (WGAN-GP, DDPM, and flow matching). The figure complements Figure 4 with additional views of the volumetric samples.

A Comparative Benchmark of GAN, DDPM, and Flow Matching for 2D and 3D Porous Microstructures Synthesis.

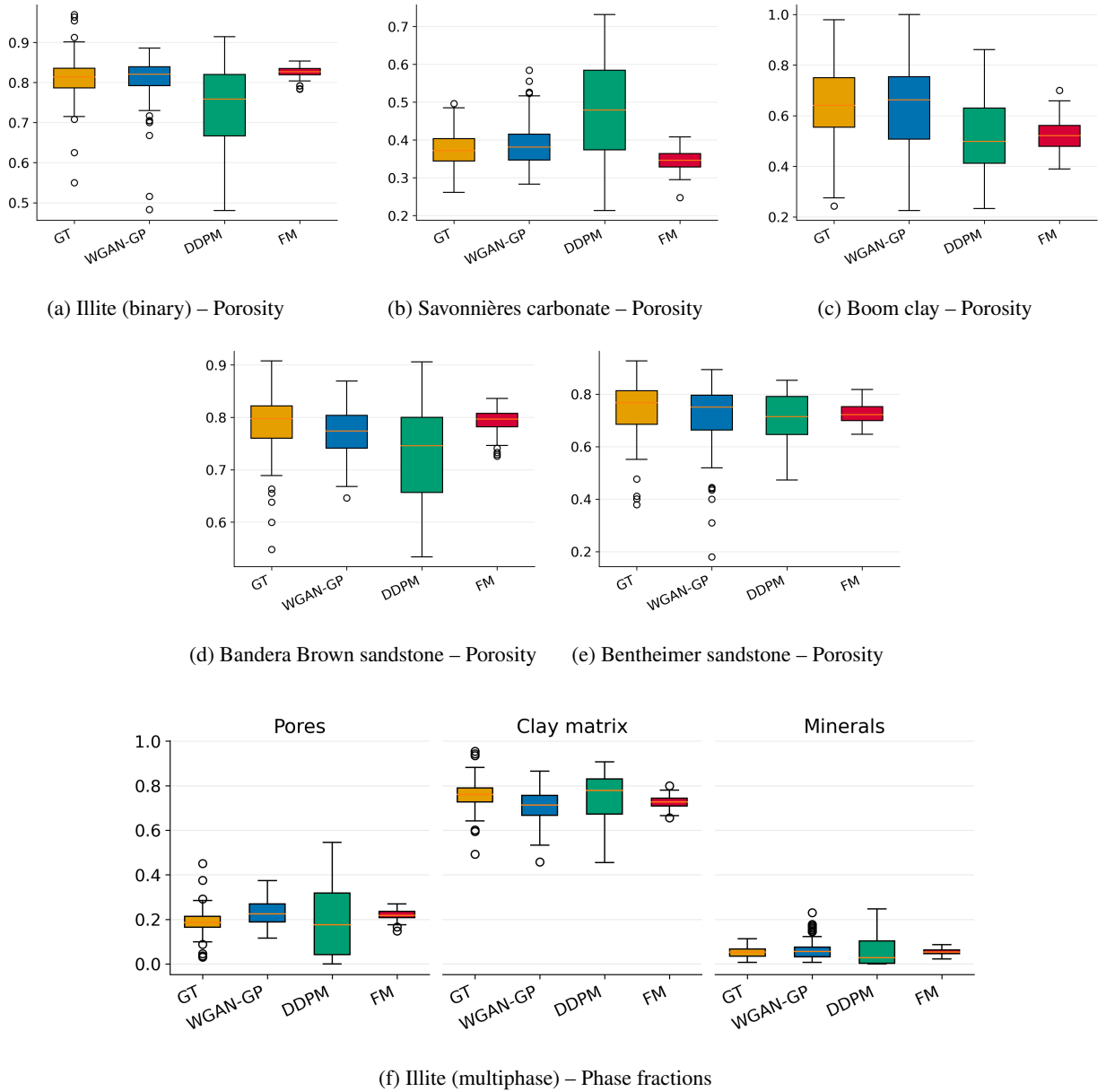


Figure 6: Distribution of porosity and phase fractions across datasets and generative models. For binary microstructures, boxplots represent the porosity. For multiphase illite, the distribution of phase fractions are shown instead. Boxes denote interquartile range, whiskers represent $1.5 \times \text{IQR}$, and markers indicate outliers.

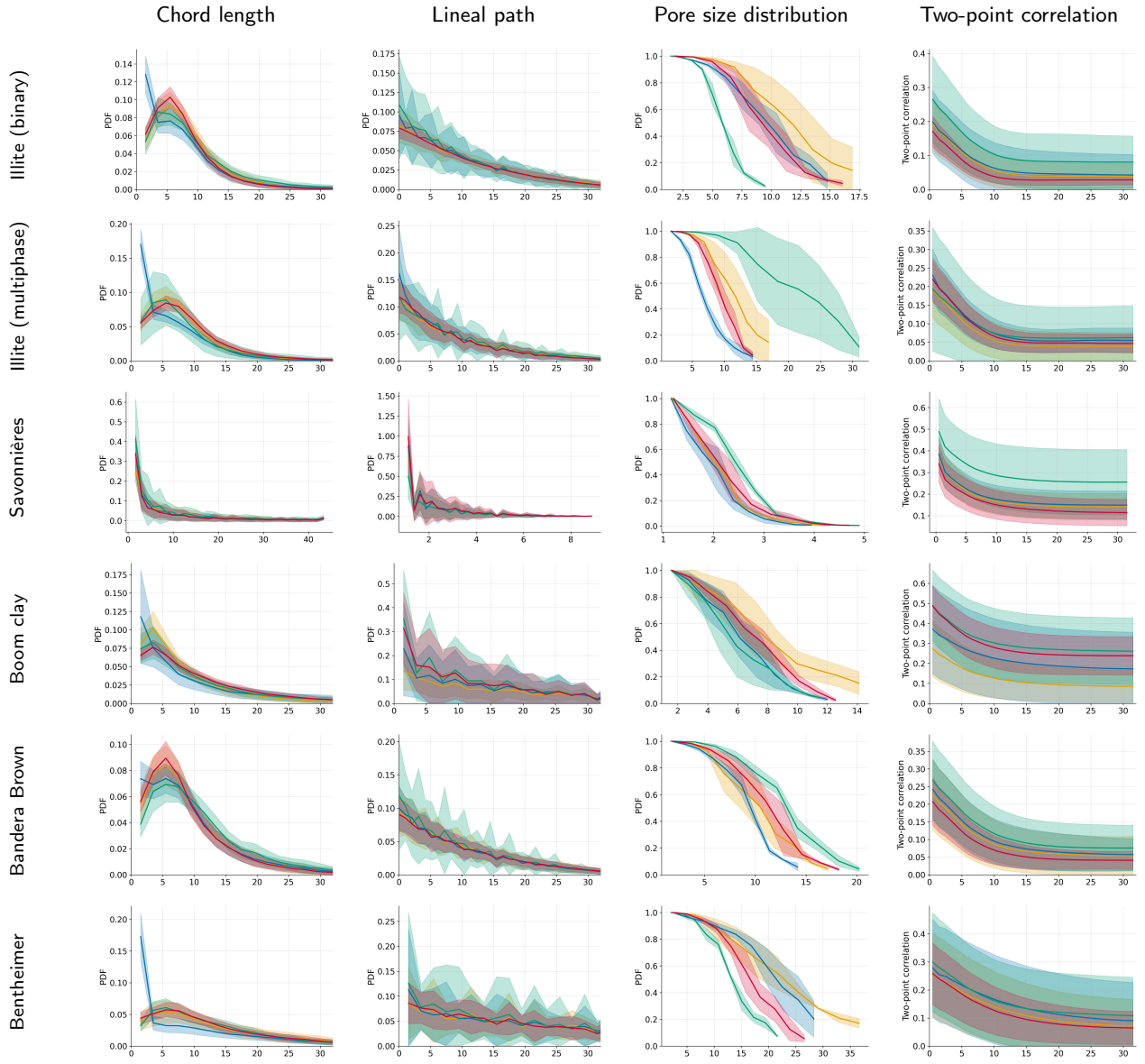
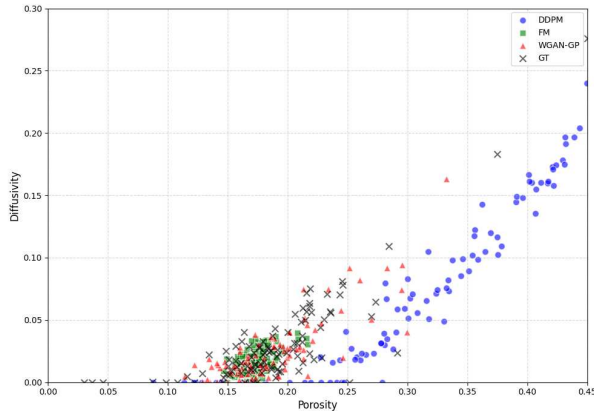
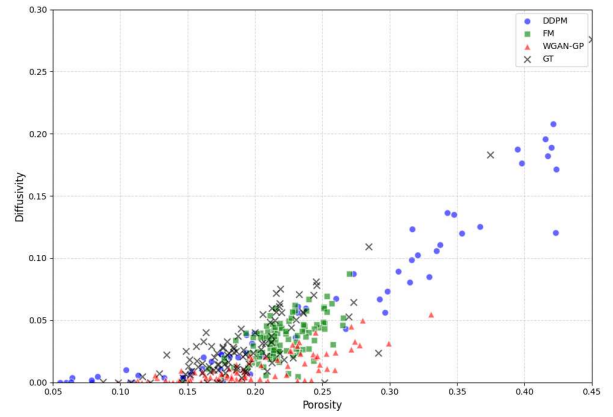


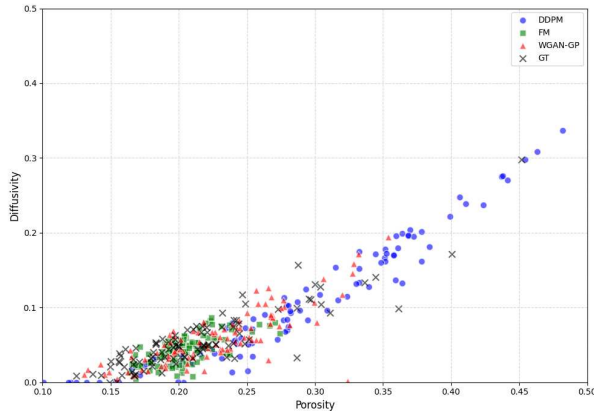
Figure 7: Comparison of structural descriptors for all investigated 3D microstructures. Columns correspond to chord length, lineal path function, pore size distribution, and two-point correlation function. Rows correspond to Illite (binary), Illite (multiphase), Savonnières carbonate, Boom clay, Bandera Brown sandstone, and Bentheimer sandstone. For the structural descriptor subplots, the x -axis represents the relevant length scale: chord length for the chord-length distribution, path length for the lineal-path function, pore radius for the pore-size distribution, and lag distance for the two-point correlation function. The y -axis shows the corresponding descriptor value: probability density for the chord-length distribution, probability for the lineal-path function, cumulative pore-volume fraction for the pore-size distribution, and two-point correlation probability for S_2 .



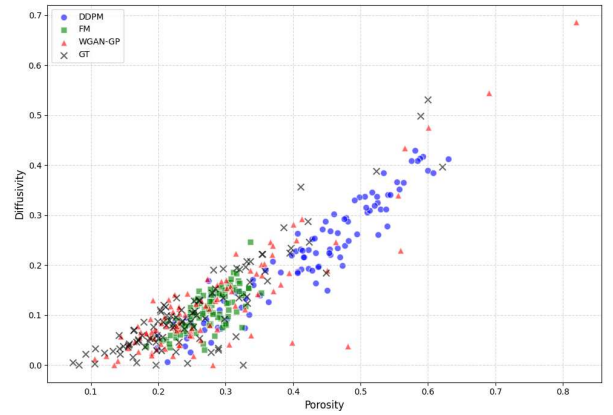
(a) Binary Illite



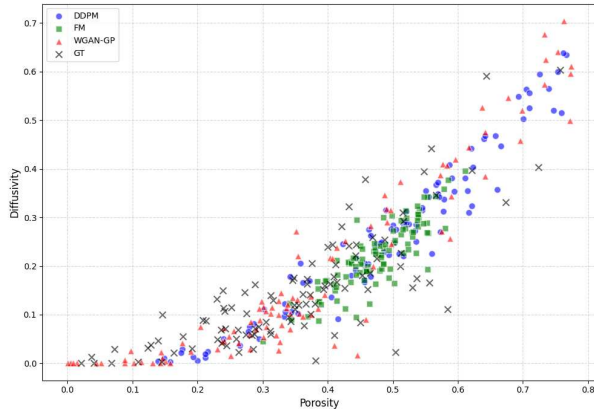
(b) Multi-phase Illite



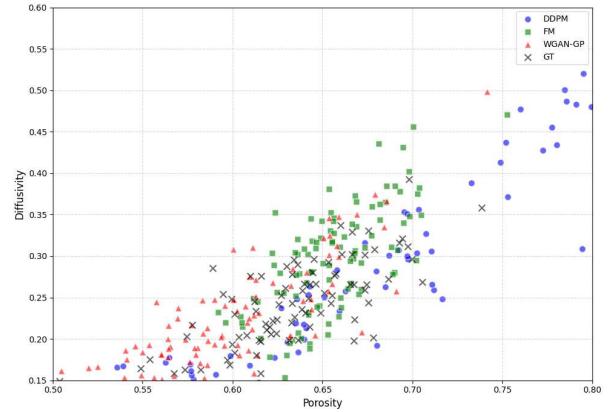
(c) Bandera Brown



(d) Bentheimer



(e) Boom Clay



(f) Savonnières Carbonate

Figure 8: Effective gas diffusivity as a function of porosity for 150 generated samples per model and dataset. Diffusivity is computed using lattice Boltzmann simulations on the processed binary volumes. Results are shown for six materials: (a) Binary Illite, (b) Multi-phase Illite, (c) Bandera Brown, (d) Bentheimer, (e) Boom Clay, and (f) Savonnières Carbonate. Markers correspond to different data sources: ground truth (black), WGAN-GP (red), DDPM (blue), and flow matching (green).