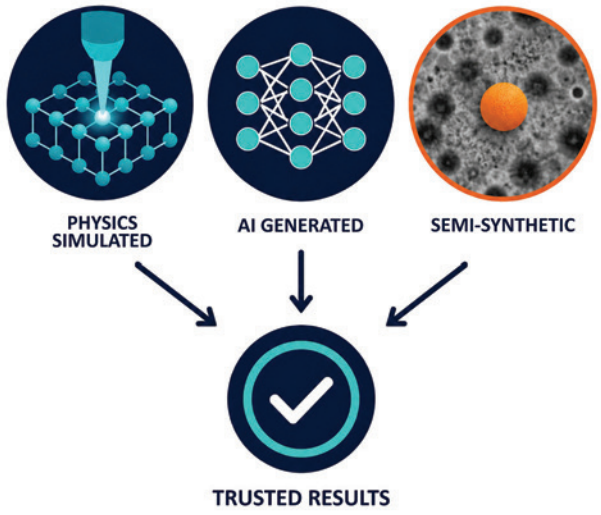
https://doi.org/10.59761/RCR5240

Synthetic data in electron microscopy. AI co-expert-assisted review

Daria M. Arkhipova,[a] Alexey S. Kashin,[a] Artem P. Vorozhtsov,[a] Valentine P. Ananikov[a,b]*

[a] *N.D.Zelinsky Institute of Organic Chemistry, Russian Academy of Sciences, 119991 Moscow, Russian Federation*
[b] *RUDN University, 117198 Moscow, Russian Federation*



Synthetic data are becoming an important methodological component of modern electron microscopy (EM), where quantitative analysis, machine learning, and automated workflows are advancing faster than the supply of well-annotated experimental data. Their value lies not simply in enlarging datasets, but in providing controlled comparisons, probing algorithmic failure modes, testing benchmark robustness, and linking image analysis to underlying physical and structural variables. At the same time, electron microscopy imposes unusually strict demands on synthetic-data design because image formation depends on complex electron–matter interactions, modality-specific contrast mechanisms, instrumental variability, and frequent mismatch between observable image features and the hidden properties of scientific interest. This review examines how synthetic data are generated, validated, and deployed across scanning electron microscopy (SEM), transmission/scanning transmission electron microscopy (TEM/STEM), cryo-electron microscopy/tomography (cryo-EM/ET), and volume electron microscopy. We discuss physics-based simulation, geometry- and microstructure-driven generation, generative artificial intelligence (AI), and hybrid semi-synthetic workflows, and analyze their use in segmentation, denoising, reconstruction, metrology, structure determination, and quantitative inference. Particular attention is given to quality, transferability, benchmark design, uncertainty, and the conditions under which synthetic data improve scientific reliability rather than merely visual plausibility. Taken together, the field is moving from synthetic images as auxiliary training material toward synthetic experiments as calibrated, task-aware companions to measurement, with growing implications for predictive and increasingly autonomous microscopy. In the preparation of this review AI co-expert technology was utilized (methodology details are described in the text). The bibliography includes 198 references.
Keywords: electron microscopy, synthetic data, machine learning, artificial intelligence, image simulation, generative models, domain adaptation, diffusion models.

Contents

1. Introduction 2
1.1. Why synthetic data have become important in electron microscopy 3
1.2. Data scarcity, annotation bottlenecks, and limits of experimental ground truth 4
1.3. Scope of this review: from SEM and TEM/STEM to cryo-EM and volume EM 4
1.4. What counts as ‘synthetic data’ in electron microscopy 5
2. Conceptual framework for synthetic data in electron microscopy 5
2.1. Fully simulated, semi-synthetic, augmented, and generative data 5
2.2. Physics-based ***vs.*** data-driven synthetic image generation 6
2.3. Synthetic data as a bridge between imaging physics and machine learning 7
2.4. Synthetic data for analysis, validation, inverse problems, and autonomous microscopy 8
3. Why electron microscopy is a particularly demanding domain for synthetic data 9
3.1. Electron-matter interactions and image formation complexity 9
3.2. Noise, drift, charging, contrast transfer, sampling, and detector effects 10
3.3. Projection, overlap, missing information, and dimensionality reduction 11
3.4. Why experimental labels are often incomplete, subjective, or impossible to obtain 12
4. Strategies for generating synthetic data 13
4.1. Physics-driven forward models 13
4.2. Geometry- and microstructure-driven synthetic data 13
4.3. Generative AI for synthetic microscopy data 14
4.4. Semi-synthetic and pseudo-labeling approaches 15
5. Synthetic data for scanning electron microscopy 17
5.1. Synthetic SEM for morphology quantification 17
5.2. Synthetic SEM for detection and segmentation 18
5.3. Synthetic SEM for inverse and predictive analysis 18
6. Synthetic data for TEM and STEM 20
6.1. High-resolution and atomic-resolution image synthesis 20
6.2. Synthetic data for denoising, restoration, and sparse acquisition 22
6.3. Synthetic data for segmentation, object detection, and metrology 23
6.4. Synthetic data for specific tasks in catalysis and functional materials analysis 23
7. Synthetic data for cryo-electron microscopy and cryo-electron tomography 25

Дата поступления 13 Мая 2026
© 2026 Uspekhi Khimii, ZIOC RAS, Russian Academy of Sciences

7.1. Why cryo-EM is a distinctive case 25
7.2. Synthetic data for particle picking and detection 25
7.3. Synthetic data for alignment, registration, and orientation estimation 26
7.4. Synthetic data for denoising and low-SNR image restoration 26
7.5. Synthetic data for *ab initio* reconstruction and heterogeneous reconstruction 27
7.6. Generative and physics-informed models in cryo-EM 28
7.7. Limits of realism in macromolecular and conformational heterogeneity 28
8. Synthetic data for FIB-SEM, volume EM, and 3D reconstruction 28
8.1. Synthetic data for serial section and volumetric segmentation 28
8.2. Synthetic datasets for neuronal and biological ultrastructure 29
8.3. Synthetic artifacts in FIB-SEM: curtaining, slicing, deformation, and contrast variation 29
8.4. Synthetic data for 3D reconstruction, registration, and benchmarking 29
8.5. Toward realistic digital volumes for large-scale EM 30
9. Major application tasks enabled by synthetic data 30
9.1. Classification 31
9.2. Detection and counting 32
9.3. Semantic and instance segmentation 33
9.4. Denoising, restoration, and super-resolution 33
9.5. Registration and alignment 33
9.6. 3D reconstruction and tomography 33
9.7. Quantitative metrology and uncertainty-aware measurement 34
9.8. Inverse problems and structure/property inference 35
10. Validation: when do synthetic data actually help? 35
10.1. Sim-to-real transfer: what should be validated 35
10.2. Visual realism ***vs.*** task realism 37
10.3. Pixel-level, object-level, and physics-level fidelity 37
10.4. Benchmark design, ablation studies, and volume effects 38
10.5. Failure modes: shortcut learning, domain gaps, and hallucinated realism 38
10.6. Confidence scoring, uncertainty, and out-of-distribution detection 38
11. Design principles for useful synthetic datasets in electron microscopy 39
11.1. Matching the data generator to the scientific question 39
11.2. Choosing the right balance between physical accuracy and throughput 40
11.3. Encoding morphology distributions, not only idealized objects 40
11.4. Modelling imaging artifacts explicitly 41
11.5. Dataset diversity, class balance, and rare-event representation 41
11.6. Annotation strategy and label ontology 41
11.7. Reproducibility, metadata, and open synthetic benchmarks 42
12. Emerging directions 43
12.1. Foundation models and synthetic pretraining for microscopy 43
12.2. Differentiable simulators and simulation-aware learning 44
12.3. Self-supervised and generative representation learning 44
12.4. Synthetic data for multimodal microscopy and correlative workflows 45
12.5. Human-in-the-loop and active learning pipelines 45
12.6. Autonomous experimentation and adaptive data generation 46
12.7. Digital twins of microscopes, samples, and experiments 46
13. Challenges and unresolved questions 47
13.1. How realistic is realistic enough? 47
13.2. Can synthetic data replace manual annotation, or only reduce it? 47
13.3. Transferability across instruments, laboratories, and sample classes 48
13.4. Generalization under changing imaging conditions 48
13.5. Standard benchmarks ***vs.*** application-specific workflows 48
13.6. Interpretability and scientific trust in generative pipelines 49
14. Outlook 49
14.1. From synthetic images to synthetic experiments 49
14.2. From passive analysis to predictive and autonomous microscopy 50
14.3. The future role of synthetic data in quantitative electron microscopy 51
15. Summary tables and practical reference guide 51
15.1. Synthetic-data categories 51
15.2. Electron-microscopy modalities *vs.* synthetic-data strategies 52
15.3. Downstream tasks *vs.* suitable synthetic-data paradigms 52
15.4. Validation criteria and characteristic failure modes 52
15.5. Open challenges and methodological directions 52
16. Conclusion 54
17. List of abbreviations 56
18. References 56

1. Introduction

Electron microscopy has entered a phase in which the bottleneck is no longer only the instrument, but the relationship between image formation, annotation, and inference. Modern scanning (SEM), transmission (TEM/STEM), cryogenic (cryo-EM/ET), and volume electron microscopy workflows generate data of extraordinary richness, yet the scientific value of those data increasingly depends on whether they can be interpreted reproducibly, quantitatively, and at a scale commensurate with contemporary experiments. This shift has elevated synthetic data from a peripheral computational convenience to a central methodological theme. The field is no longer asking only how to acquire better micrographs. It is asking how to construct controlled computational counterparts of experiments that can support learning, validation, benchmarking, and decision-

V.P.Ananikov. D.Sc., Academician of the RAS, Laboratory Head at ZIOC RAS.
E-mail: val@ioc.ac.ru
Current research interests: AI in science, digital technologies, microscopy, molecular complexity and transformations, metal complexes and nanoparticles, development of new generation of highly active nanosized and molecular catalysts, mechanistic studies of chemical reactions by experimental and theoretical methods.
D.M.Arkhipova. Ph.D., Senior Researcher at the same Institute.
E-mail: arkhipova_daria@ioc.ac.ru
Current research interests: synthetic data, electron microscopy, transition-metal catalysis, ionic liquids.
A.S.Kashin. D.Sc., Leading Researcher at the same Institute.
E-mail: a.kashin@ioc.ac.ru
Current research interests: electron microscopy, microscopy data processing, chemistry of disperse systems, transition-metal catalysis, X-ray spectroscopy.
A.P.Vorozhtsov. Research Engineer at the same Institute.
E-mail: artem2001@ioc.ac.ru
Current research interests: synthetic data generation, multimodal large language models, deep learning, AI in chemistry.

making when direct experimental truth is sparse, ambiguous, or fundamentally inaccessible.

This change did not arise in isolation. The broader microscopy and machine-learning literature has already shown that AI-driven analysis is reshaping electron microscopy from post hoc image interpretation toward automated, data-rich, and increasingly autonomous pipelines.[1–5] At the same time, recent developments have revealed an important tension. Powerful neural models can extract structure, classify morphology, detect defects, segment complex scenes, and even assist in experiment control, but their success is constrained by the rarity of expertly labeled datasets, by unstable transfer across instruments and specimen classes, and by the fact that many of the quantities scientists care about are latent rather than directly visible in any one image. Synthetic data have emerged precisely at this junction: as a way to expose hidden variables, stress-test interpretation pipelines, and formalize what the microscope is expected to reveal.

1.1. Why synthetic data have become important in electron microscopy

The immediate appeal of synthetic data in electron microscopy is pragmatic. Many downstream tasks now expected from EM images — particle detection, morphology segmentation, atom localization, denoising, tomographic reconstruction, defect metrology, conformational assignment, and property prediction — are supervised or benchmarked tasks, whereas the corresponding experimental ground truth is often expensive, incomplete, or nonexistent (Fig. 1). Synthetic data provide a route to paired inputs and outputs in situations where the experiment yields only the input. In this sense, they do not merely enlarge datasets; they change what kinds of computational questions can be asked.

Yet the deeper reason synthetic data have become important is conceptual rather than logistical. Electron microscopy is not simply a source of images. It is a family of measurement systems in which specimen structure, beam–matter interaction, detector

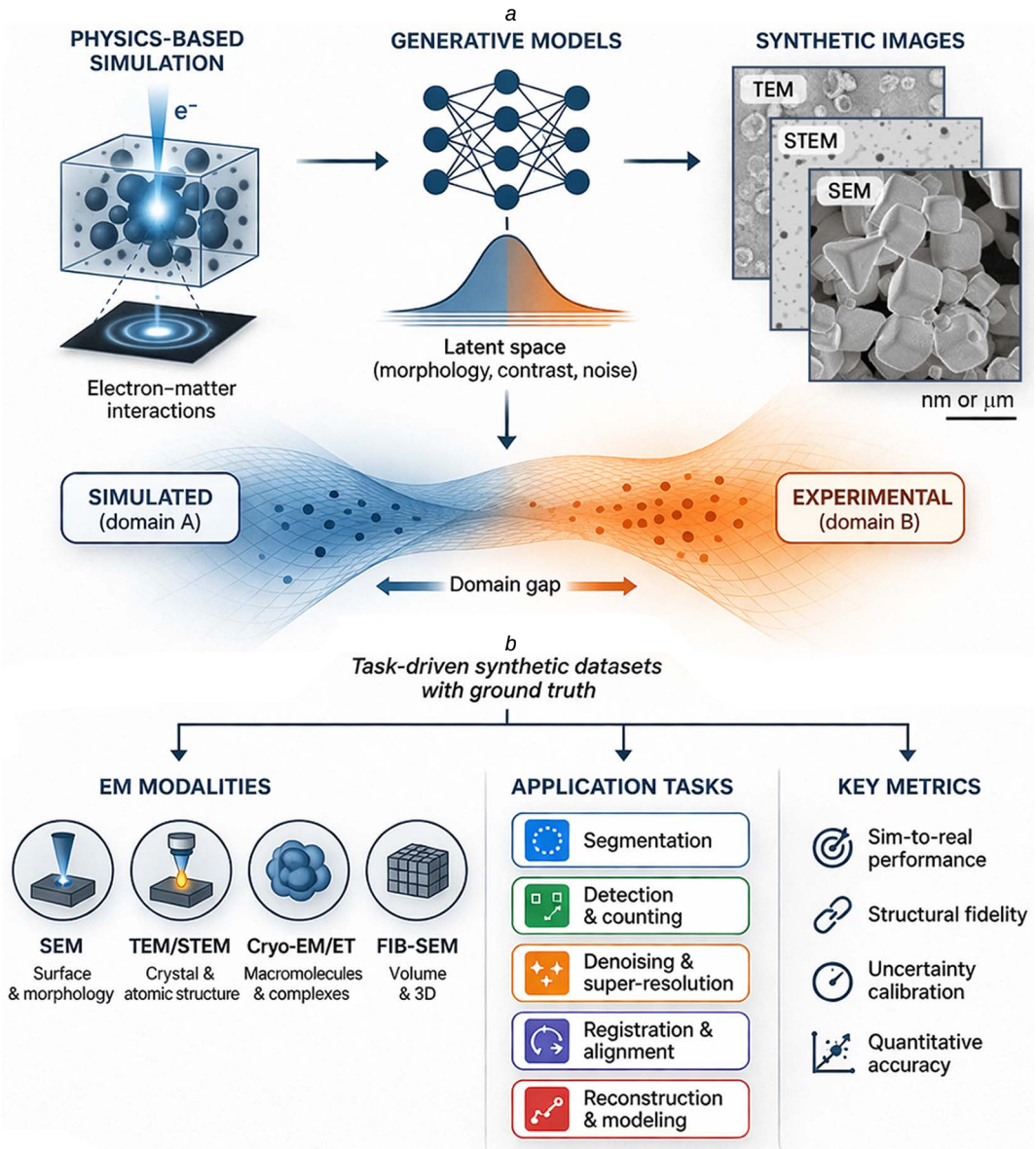


Figure 1. Synthetic data in electron microscopy (EM): AI bridges the gap between physics-based simulation and experimental images (*a*); modality-specific, task-oriented synthesis and validation for synthetic datasets (*b*).

physics, acquisition strategy, and reconstruction algorithm are tightly coupled. Once machine learning enters this picture, the problem becomes one of matching representations to measurement physics (see Fig. 1 *a*). Synthetic data are powerful because they allow one to construct that match deliberately. They can encode known symmetries, simulate nuisance factors, generate latent-variable labels, and define controlled perturbations that make algorithmic behavior scientifically interpretable rather than merely statistically successful (see Fig. 1 *b*).

Recent literature illustrates the flow of this transition. Review-level analyses now frame machine learning in EM as part of a larger methodological reorganization of microscopy itself.[1–3] At the application level, synthetic or simulation-linked learning has been used for cryo-EM map enhancement,[6,7] for particle-picking systems and curated benchmark resources,[8–10] for automated cryo-EM model building,[11] for multimodal integration of cryo-EM with structure-prediction priors,[12] and for domain-randomized cryo-ET analysis.[13] In materials-focused microscopy, synthetic and simulation-aware learning have supported quantitative TEM inference, atom finding, and defect analysis.[14–16] Across these otherwise diverse studies, the same idea recurs: synthetic data are becoming a means to turn underdetermined image-analysis problems into controlled inference problems (see Fig. 1 *a*).

A second, equally important development is the rise of automation and closed-loop experimentation. Once microscopy begins to move from passive acquisition toward active decision-making, synthetic data acquire a new role. They become training environments in which models learn what variability matters, what uncertainty looks like, and what measurement should come next. This emerging logic is visible in broader discussions of autonomous microscopy and self-driving experimentation.[4,5,17–20] Synthetic data are therefore important not only because labeled data are scarce, but because future microscopes will increasingly require model-based priors before exhaustive human annotation is even possible.

1.2. Data scarcity, annotation bottlenecks, and limits of experimental ground truth

If synthetic data are now central, it is because electron microscopy faces an unusually severe combination of data scarcity and truth scarcity. These are not the same problem. Data scarcity refers to the limited availability of sufficiently diverse, well-curated, and task-relevant datasets. Truth scarcity is more fundamental: even when many experimental images exist, the scientifically relevant answer may not be directly knowable from the experiment itself. A nanoparticle (NP) image may not reveal exact three-dimensional size and depth; a cryo-EM micrograph does not come with known particle poses and conformational coordinates; a serial-section volume does not provide its own deformation-free ground truth; an atomic-resolution STEM image does not arrive with exhaustive, pixel-perfect labels for every chemically distinct site. Human experts can annotate proxies, but proxies are not the same as truth. The importance of comparing experimental and synthetic data is already well recognized in fields such as catalysis.[21,22] The broader push toward data-driven materials science has made clear that new, large volumes of structure-related microscopy data are urgently needed.[23,24]

This distinction explains why annotation alone cannot solve the problem. Manual labeling in EM is slow, expensive, and often subjective, but the more serious limitation is that many labels are only partial abstractions of a richer hidden state. Even apparently simple labels such as boundaries, particle counts, or class assignments may depend on object appearance in electron microscopy images, which is significantly influenced by dose, detector response, reconstruction choices, or specimen preparation history. The challenge is particularly acute in cryogenic imaging, in tomography, and in low-dose or *in situ* workflows, where signal is weak, corruption is structured, and variability across experiments is large. In these regimes, experimental datasets can be substantial and still remain insufficient as training resources because they do not span the latent variables that dominate failure.

Recent studies underscore how the field is responding to this bottleneck. Semi-supervised and self-supervised strategies have reduced dependence on large labeled datasets, but they do not eliminate the need for well-structured priors.[25,26] Benchmarking efforts such as EM-stellar make clear that performance claims in electron microscopy depend strongly on the chosen annotation regime and task design.[27] Domain-specific datasets remain valuable, as shown by curated cryo-EM particle-picking resources,[8] but curated real data alone do not resolve the latent-variable problem. In digital biology, the combination of synthetic data and image analysis has become a major driving force.[28,29] What synthetic data contribute is not just more images, but more controllable conditions. They allow one to vary thickness, dose, noise statistics, orientation, object rarity, contamination, or deformation one factor at a time and to observe how those factors alter performance. This is why synthetic datasets often function simultaneously as training resources and as epistemic tools for understanding where a pipeline will fail.

A recent systematic review of AI-driven microstructure and surface characterization reaches the same practical conclusion, *viz*., limited labeled microscopy data, domain shifts between imaging conditions, and incomplete interpretability remain central constraints, while data augmentation and synthetic generation are increasingly used to improve robustness in data-limited regimes.[30]

The broader methodological implication is significant. In computer vision for standard optical photography, labels are often treated as external truth attached to an otherwise fixed image. In EM, the label frequently depends on the measurement model itself. Synthetic data become especially valuable, and in some settings indispensable, because they allow the investigator to define the relation between hidden structure and observable image in an explicit way. They do not abolish experimental uncertainty, but they make it analyzable. The problem of lost (or unrecoverable) information in electron microscopy[31] only reinforces this conclusion.

1.3. Scope of this review: from SEM and TEM/STEM to cryo-EM and volume EM

This review is written from the premise that 'synthetic data in electron microscopy' is not one technical niche but a cross-cutting methodological framework. The discussion therefore spans the major EM modalities and the major task classes (see Fig. 1 *b*), while remaining focused on one unifying question: how can deliberately generated data be used to support trustworthy inference from electron-microscopy measurements?

The scope is intentionally broad and covers a large set of electron microscopy techniques, which differ significantly in the accessible specimen region, image-formation mechanism, and physical principles that connect the specimen to its appearance in electron micrographs. Scanning electron microscopy is

included because it is a key electron microscopy technique designed to study the surface morphology of various objects. Given the focus of the topic under discussion, it is important to note that, in the formation of an image in SEM, contrast is strongly governed by beam–specimen interactions, signal generation, and detector geometry, although electron optics and probe formation remain integral to image formation. TEM and STEM occupy a complementary position in which high-resolution imaging of the sample's internal structure is available and the forward model is more tightly linked to scattering physics and the properties of electron optics, enabling synthetic data to support atomic-resolution interpretation, restoration, and inverse analysis. Cryo-EM and cryo-ET introduce a distinct regime of extreme low dose, projection ambiguity, conformational heterogeneity, and three-dimensional reconstruction, making synthetic data especially valuable for benchmarking and for latent-variable inference, and in some settings close to indispensable. Focused ion beam scanning electron microscopy and related volume EM approaches extend the discussion to serial imaging, anisotropy, sectioning artifacts, and the challenge of large-scale three-dimensional semantic reconstruction. Across all of these modalities, the review asks not simply how synthetic data can imitate images, but how they can support analysis, validation, reconstruction, and eventually autonomous experimentation.

The scope is also broad in task space. The relevant applications include classification, detection, segmentation, denoising, super-resolution, registration, metrology, tomographic inversion, uncertainty quantification, and the prediction of hidden structural or physical variables from image observables. This breadth is not an optional expansion. It is necessary because the value of synthetic data depends on the downstream scientific objective. A synthetic dataset that is sufficient for counting particles may be inadequate for recovering conformational landscapes; a simulator useful for benchmarking a denoiser may fail if used to train a generative inverse model for quantitative metrology. The review therefore follows a logic that moves from concepts and generation strategies to modality-specific case studies, then to task-oriented synthesis, validation, design principles, emerging directions, unresolved questions, and outlook (see Fig. 1 *b*).

We aimed to include the studies necessary to represent and analyze the field. We did not attempt to cite all possible literature references on this or related subjects; rather, a focused literature selection was made.

1.4. What counts as 'synthetic data' in electron microscopy

A recurring source of confusion in the literature is that the term 'synthetic data' is used too narrowly in some contexts and too loosely in others. For the purposes of this review, synthetic data are understood as deliberately constructed image-like, signal-like, or measurement-like data, together with associated labels or latent variables, generated through simulation, hybrid composition, controlled perturbation, or learned translation for the purpose of training, testing, validating, or operating an inference system. This definition is deliberately functional. It focuses not on whether every pixel is generated *ab initio*, but on whether the data have been constructed in a controlled way to preserve a meaningful relation between hidden variables and the downstream task.

Such a definition must be broad because the field already includes several distinct but overlapping regimes. Fully simulated data remain essential wherever physical image formation can be modeled with sufficient fidelity. Hybrid and semi-synthetic strategies combine real backgrounds, experimental textures, or partial measurements with synthetic objects, labels, or transformations. Generative methods, including learned translation and distribution sampling, can create realistic-looking observations even when the forward physics is only partially explicit. Augmentation occupies yet another space, in which new samples are not synthesized de novo but derived through perturbation of existing data. The increasing sophistication of these regimes means that the practical question is no longer whether a dataset is 'truly synthetic' in a binary sense. The real question is what kind of synthetic construction is appropriate for a given scientific claim.

This question has become even more pressing as predictive and autonomous microscopy advance. Learned systems can now identify atoms, improve cryo-EM maps, infer quantitative sample properties, guide autonomous acquisition, and generalize across changing imaging conditions.[14–16,32,33] Meanwhile, the broader literature on autonomous scientific discovery and AI-guided materials workflows suggests that microscopy will increasingly function as part of a model-driven experimental loop rather than as an isolated imaging endpoint.[4,5,34] Under these conditions, synthetic data should be viewed as computational interfaces between specimen, microscope, and model. They are useful when they preserve the causal and statistical structure relevant to the task at hand, and misleading when they merely reproduce image style without preserving that structure.

This perspective provides the rationale for the rest of the review. Before one can judge realism, transferability, benchmark design, or scientific trust, one must first establish a disciplined conceptual vocabulary. Section 2 therefore follows directly by distinguishing fully simulated, semi-synthetic, augmented, and generative data, by comparing physics-based and data-driven generation strategies, and by defining synthetic data as an operational bridge between image-formation physics and machine-learning inference rather than as a generic substitute for missing pictures.

2. Conceptual framework for synthetic data in electron microscopy

This section defines what should count as synthetic data in electron microscopy, distinguishes the main generation paradigms, and argues that synthetic data are not merely substitutes for missing images but operational interfaces connecting forward imaging physics, annotation strategy, and downstream inference. This conceptual framing also prepares the transition to the next section, where the central question becomes why electron microscopy is a particularly difficult domain in which to make synthetic data realistic and useful.

2.1. Fully simulated, semi-synthetic, augmented, and generative data

A useful conceptual framework begins by separating synthetic electron microscopy data into four broad classes: fully simulated data, semi-synthetic data, augmented data, and generative data. These categories are related, but they are not interchangeable, and much confusion in the literature arises when they are discussed as if they were equivalent. Fully simulated data are produced by specifying an underlying object, structure, or scene and then applying a forward model of image formation. In this

regime, both the object-space ground truth and the image-space output are known by construction, which makes fully simulated data especially valuable for controlled learning, benchmarking, and inverse problems.[35–37]

Semi-synthetic data occupy an intermediate position. They typically preserve some part of the real measurement, *e.g.*, morphology, background texture, detector statistics, or experimentally acquired base images, while injecting synthetic labels, translated contrast, simulated objects, or simulated perturbations. In practice, semi-synthetic pipelines are often among the most useful in microscopy because they exploit real image priors while still creating labels or conditions that would be difficult to obtain experimentally. Pseudo-labeling through cross-modality translation,[38,39] physics-guided insertion of structures into real backgrounds, and style-adapted cryo-EM projections[40] are all examples of this middle ground.

Augmented data are derived from existing images by operations such as rotation, rescaling, cropping, blur, noise injection, intensity perturbation, or more microscopy-specific modifications. Augmentation increases dataset diversity, but conceptually it should not be confused with *de novo* synthesis. Augmentation perturbs already available samples; it does not create new physical scenarios. For this reason, augmentation is often necessary but rarely sufficient when the original dataset is small, biased, or missing critical states of morphology or imaging conditions.[41,42]

Generative data form the fourth class. Here, a trained model learns to sample from a data distribution or to translate one domain into another. Techniques such as adversarial training, diffusion-based generation, and style transfer can therefore generate synthetic images without explicitly encoding the full microscope physics. This is powerful, especially when rapid data generation or domain adaptation is required, but it also shifts the central question from explicit physical correctness to distributional faithfulness and task relevance. In microscopy, generative methods are most compelling when they are used not as free-standing image synthesizers, but as controlled operators that bridge a gap between a labeled domain and an unlabeled one, or between a simplified simulator and the visual statistics of experimental images.[39,40,43]

The practical implication is that synthetic-data papers should always state which of these four regimes they are using, what parts of the pipeline remain grounded in real measurements, and which variables are genuinely under user control. Without this distinction, it becomes impossible to compare claims about realism, label quality, transferability, or computational cost across different studies.

The workflow in Fig. 2 is especially useful as a conceptual template because it makes explicit that synthetic data are not generated in a single step. Reality is first decomposed into partial models, those models are assembled into parametric scenarios, concrete instances are sampled, and the resulting synthetic data are then validated against downstream learning performance or experimental data. This framing is important for electron microscopy because it emphasizes that synthetic datasets must be designed around variables that matter scientifically, not merely around image appearance.[35,36]

2.2. Physics-based *vs.* data-driven synthetic image generation

A second key distinction concerns how synthetic images are generated. Physics-based approaches begin from a forward model of the microscope and of the specimen. They aim to encode, at least approximately, the mechanisms by which structure becomes signal: scattering, projection, contrast transfer, detector response, sampling, noise, and other instrument- or modality-specific effects. Their principal strength is interpretability. Because the parameters correspond to physical or geometric variables, one can ask how performance changes with dose, blur, particle density, thickness, drift, or other factors, and one can generate labels that remain consistent with the simulated world.[37,44,45]

The weakness of purely physics-based synthesis is that realism is expensive. Accurate simulators may demand substantial computation, extensive prior knowledge, and many nuisance parameters that are difficult to estimate from routine experimental metadata. Moreover, many practically important image features in electron microscopy arise from imperfect samples, laboratory-specific settings, undocumented preprocessing, or coupled artifacts that are difficult to model exhaustively. The more realistic the simulator becomes, the higher the dimensionality of the parameter space and the more difficult it is to sample it meaningfully.[37,40]

Data-driven generation attacks the same problem from the opposite direction. Instead of deriving image statistics from first principles, it learns them from data. Domain randomization, adversarial training, diffusion-based translation, and style

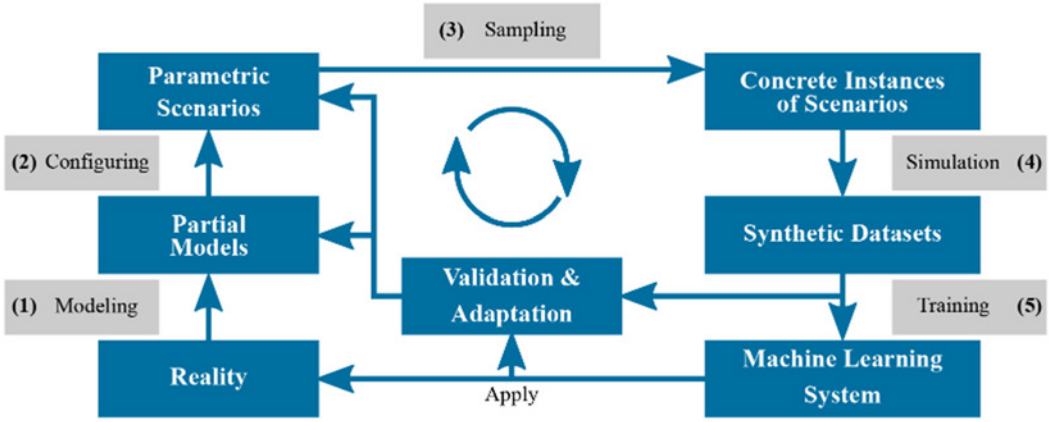


Figure 2. Parametric workflow for synthetic data generation and iterative validation. The figure is reproduced from Trampert *et al.*[36] according to Open Access Creative Commons Attribution (CC BY) license.

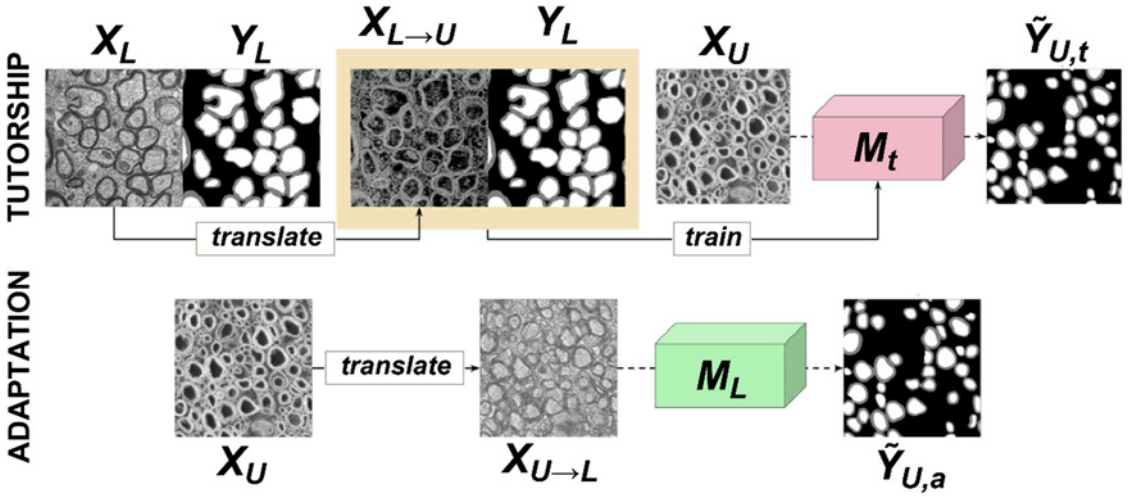


Figure 3. Cross-modality pseudo-labeling through tutoring and adaptation pathways between TEM and SEM domains. X_L is images from labeled (L) domain; X_U is images from unlabeled (U) domain; $X_{L\to U}$ and $X_{U\to L}$ are synthetic images resulting after translation between domains L and U; Y_L is segmentation labels; M_L is model pre-trained on [X_L, Y_L]; M_t is model trained on [$X_{L\to U}$, Y_L]; $\tilde{Y}_{U,t}$ and $\tilde{Y}_{U,a}$ are sets of pseudo labels. The figure is adapted from Boschet *et al*.[38] and published according to Open Access Creative Commons License CC BY 4.0.

transfer can rapidly create large volumes of images that visually resemble the target domain or that span a broad range of plausible conditions. In nanoscience image analysis, for example, configurable generators based on domain randomization can synthesize image-mask pairs at scale for particle segmentation and measurement tasks.[41] In cross-modality microscopy labeling, unpaired translation can create synthetic target-domain images from labeled source images, thereby replacing manual annotation by a learned transport between domains (Fig. 3).[39]

The limitation of data-driven generation is that it may reproduce appearance without reproducing causality. A generated image can look realistic while encoding shortcuts, omitting rare but important failure modes, or distorting the relation between morphology and signal. For scientific imaging, this matters because the objective is not simply to fool a discriminator or a human observer. The objective is to produce data that preserve the invariances, ambiguities, and uncertainties that a downstream model will face on real measurements. The central criterion is therefore not photorealism alone, but task realism. This is why several of the most convincing recent studies adopt hybrid strategies: physics-based structure generation followed by learned appearance adaptation; simulated labels combined with real backgrounds; or simulation-generated volumes refined through style transfer to match experimental statistics.[40,45,46]

For electron microscopy, the most useful conceptual conclusion is not that one paradigm must replace the other. Rather, physics-based and data-driven synthesis should be understood as complementary ends of a design spectrum. Physics-based methods give controllability and causal interpretability; data-driven methods give adaptability, efficiency, and domain matching. Productive pipelines increasingly combine them.

2.3. Synthetic data as a bridge between imaging physics and machine learning

The conceptual role of synthetic data in electron microscopy becomes clearest when they are viewed as bridges rather than substitutes. On one side lies imaging physics: objects, structures, contrast mechanisms, acquisition geometry, detector response, and noise. On the other side lies machine learning: classification, localization, segmentation, denoising, reconstruction, and representation learning. Synthetic data connect these domains by translating physically meaningful variation into trainable examples. In the best cases, a synthetic dataset does more than enlarge a corpus: it expresses a hypothesis about which hidden variables matter and how they shape the images used for inference.[35,36]

This bridging function operates at several levels. First, synthetic data make latent variables observable during training. In atomic-resolution STEM, simulated training libraries can provide exact atomic positions, column labels, blur states, and noise levels that are inaccessible as direct ground truth in routine experiments, enabling robust segmentation, localization, denoising, and deblurring models.[44] Second, synthetic data can align domains that are related in content but dissimilar in contrast or acquisition. The tutoring and adaptation routes in Fig. 3 show that labeled TEM images can be translated into a SEM-like domain, or unlabeled SEM images can be mapped toward a TEM-like domain, making the segmentation task transferable even when direct manual annotation of the target domain is scarce.[39]

Third, synthetic data can serve as priors for sim-to-real learning. In microscopy-based pose estimation, simulated data enlarge the available labeled set, while learned translation and matching reduce the domain gap between simulation and experiment.[46] In semiconductor SEM metrology, models trained on simulated data can be adapted in an unsupervised manner to experimental images for three-dimensional surface reconstruction and dimension prediction.[45] In such examples, synthetic data are not replacing real measurements; they provide the structured source domain from which transferable representations can be learned.

This is also why synthetic data are central to foundation-like pretraining in microscopy. A model pretrained on richly parameterized synthetic or microscopy-native data may learn features that are much closer to the target imaging domain than features inherited from natural-image datasets. Reviews of microstructural image analysis increasingly stress that microscopy-specific pretraining and synthetic data generation

are essential for improving generalization under limited labeling budgets.[42,47]

Seen from this perspective, the quality of a synthetic dataset is inseparable from the mapping it proposes between hidden physical variables and learned representations. The synthetic data problem is therefore also a representation-design problem. One does not merely ask whether the images look real, but whether they teach the model the right abstractions.

2.4. Synthetic data for analysis, validation, inverse problems, and autonomous microscopy

Once synthetic data are treated as operational bridges, their range of use becomes broader than standard supervised training. A first major role is analysis. Synthetic datasets enable classification, detection, segmentation, denoising, localization, and measurement tasks in regimes where direct annotation is slow, expensive, or ambiguous. This includes NP instance segmentation, atom finding in STEM, and cross-modality tissue segmentation.[39,41,44]

A second major role is validation and benchmarking. In many electron microscopy workflows, ground truth for the final scientific quantity of interest is only partially known or not knowable at all. Synthetic data can therefore function as controlled testbeds in which the true state of the object, structure, conformation, or geometry is available by construction. This is especially important in cryo-EM, where benchmarking heterogeneous reconstruction algorithms requires synthetic micrographs with known conformational trajectories and known sources of variability. Roodmus explicitly addresses this need by generating cryo-EM micrographs from molecular-dynamics-derived conformations so that reconstruction pipelines can be evaluated against ground truth rather than only against internal consistency.[37] Profet, EMPIARreader and CAKED (PERC) extends this logic at the workflow level by facilitating access to curated experimental and simulated cryo-EM data and machine-learning-ready augmentation and labeling utilities.[48]

A third role is in inverse problems. Many electron microscopy tasks do not end with image interpretation; they aim to infer latent structure, geometry, pose, or dynamics from incomplete and noisy measurements. Synthetic data are invaluable here because they define paired examples between image space and latent-variable space. The pose-estimation study of micro-objects using SEM and optical microscopy is emblematic: the network does not simply classify images but learns to infer three-dimensional pose from monocular observations using synthetic-to-real transfer.[46] The SEM surface-reconstruction work similarly uses simulated data and domain adaptation to recover three-dimensional topography from two-dimensional SEM imagery.[45]

A fourth role is in autonomous microscopy and closed-loop experimentation. When coupled with measurement pipelines, synthetic data can accelerate automated image analysis, support uncertainty-aware decision-making, and provide training environments for systems that must operate before exhaustive manual labels exist. The aUTonomous Image anaLysis to accelerate the discovery and integration of energy matErials (UTILE) synthetic dataset generator (UTILE-Gen) framework demonstrates this transition well: synthetic images generated by domain randomization support instance segmentation and downstream particle-size analysis, and the trained models are then deployed in an automated analytics platform.[41] In this sense, synthetic data are not only a data-engineering convenience; they are enabling infrastructure for autonomous, scalable microscopy workflows.

The cryo-EM case illustrates both the promise and the caution required. Neural style transfer in FakET markedly lowers the cost of generating labeled cryo-ET-like data and can approximate benchmark performance for localization and classification tasks, yet the authors explicitly note that their simulator does not embed the full underlying TEM physics and therefore must be interpreted according to the task for which it is used (Fig. 4).[40] This is a crucial lesson for the field: synthetic data can successfully support analysis and benchmarking even when they are not complete digital twins of the microscope, but the allowed

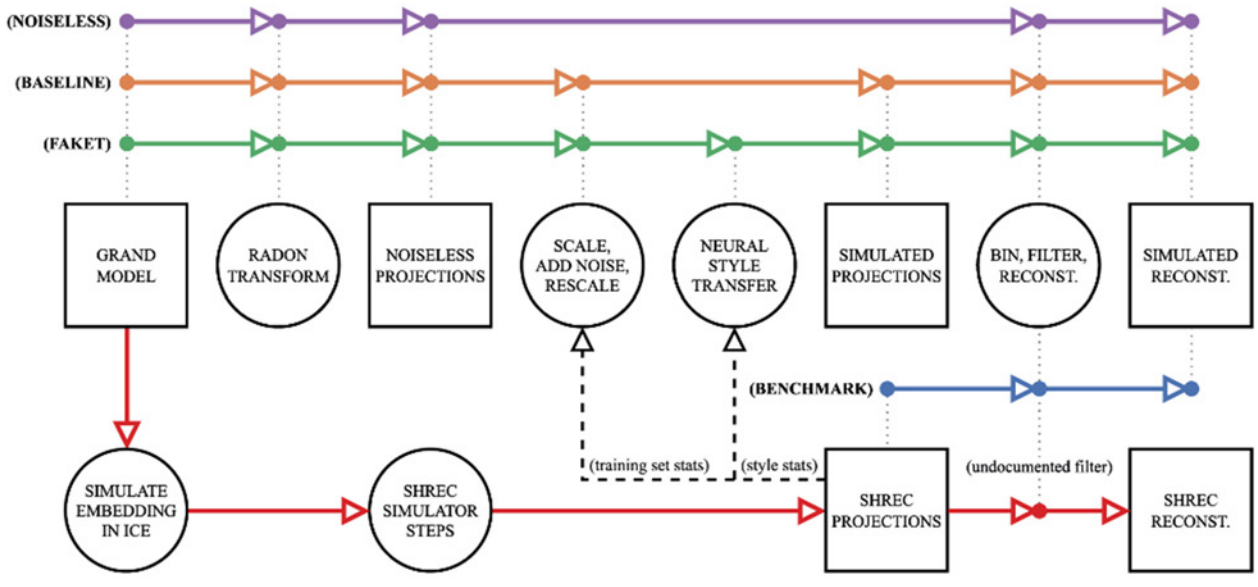


Figure 4. Comparison of benchmark, baseline, and neural-style-transfer-based cryo-ET simulation routes. The figure is published from Harar *et al.*[40] according to Open Access Creative Commons License CC BY.

degree of approximation depends on what scientific claim is being made.

Taken together, these examples suggest a broad but disciplined definition. Synthetic data in electron microscopy should be understood as deliberately constructed image-like or measurement-like data, together with associated labels or latent variables, generated through simulation, hybrid composition, or learned translation for the purpose of training, testing, or operating inference systems. The important point is not whether every pixel is born from first-principles simulation. The important point is whether the dataset faithfully preserves the relation between relevant hidden variables and the inferential task for which it is intended.

If synthetic data are meant to connect specimen physics, image formation, and machine learning, then the immediate question becomes why this connection is unusually difficult in electron microscopy. The answer lies in the domain itself: electron-matter interactions, low-dose constraints, projection effects, missing information, detector artifacts, and the frequent impossibility of obtaining definitive experimental ground truth. Section 3 therefore follows directly from the present discussion by examining why electron microscopy is a particularly demanding test case for synthetic-data methodology.

3. Why electron microscopy is a particularly demanding domain for synthetic data

Electron microscopy is not merely another imaging domain in which synthetic data can be used to enlarge a training set. It is a measurement regime in which the observed image is shaped by a long chain of coupled physical, instrumental, computational, and specimen-dependent transformations. This can be illustrated by a number of key examples that demonstrate the full complexity of the relationship between an object and its resulting image. In cryo-EM, the forward problem already involves projection of a three-dimensional object from an unknown orientation, convolution by a contrast transfer function (CTF), in-plane shifts, amplitude scaling, and severe noise; in heterogeneous systems, conformational state becomes an additional latent variable.[49] Even the idealized projection model is only an approximation, because the Fourier data are sampled on the Ewald sphere rather than exactly on a central plane. Thus, at sufficiently high resolution or for sufficiently large particles, Ewald-sphere curvature causes departures from the central-section (projection) approximation.[50] In SEM, STEM, four-dimensional scanning transmission electron microscopy (4D-STEM), and ptychography, the imaging chain is equally complicated: electron–matter interactions, secondary and backscattered electron generation, detector response, sampling trajectory, charging, drift, dose budget, and post-processing all influence what is finally recorded.[51–54]

For this reason, synthetic data in EM cannot be treated as a generic computer-vision convenience. They are often needed because the experimental image does not contain direct, exhaustive, or uniquely interpretable ground truth. A single image may represent a lossy two-dimensional projection of a three-dimensional, dynamically fluctuating object; contrast may oscillate with spatial frequency; some information is blurred or never measured; the specimen may change under observation; and the labels required by machine learning may be incomplete, subjective, biased by the acquisition protocol, or physically impossible to obtain experimentally.[37,52,55] These features make EM an unusually demanding but also unusually revealing testbed for synthetic data. The key challenge is not only realism, but deciding which parts of the forward process must be preserved so that the synthetic images remain useful for inference, validation, and algorithm development.

3.1. Electron-matter interactions and image formation complexity

The first source of difficulty is that, although the idea of using physical models of electron–matter interactions to establish links between a real object and its image is appealing, more and more factors need to be taken into account as the model is refined. Some examples are given below. In cryo-EM single-particle analysis, the object is represented as a three-dimensional scattering potential or density, yet each measurement is a noisy two-dimensional image formed after projection, CTF modulation, and translation by unknown shifts; the viewing direction itself must be inferred from the data.[49,56] This means that image formation is inseparable from inverse modeling. A network trained on images alone can easily learn shortcuts tied to microscope settings, defocus values, or preprocessing artifacts instead of structural content unless those factors are represented in the synthetic generator.

The physical chain becomes more intricate when the standard projection approximation breaks down. Wolf, DeRosier, and Grigorieff[50] showed that the Fourier transform of an image corresponds to the surface of the Ewald sphere, not exactly to a central section through the origin. For sufficiently large particles or sufficiently high resolution, this deviation produces non-negligible reconstruction error, which simulation can expose but simplistic image generators would miss. This example is important conceptually: even in a mature and highly formalized branch of EM, the apparent image is only an approximation to a richer scattering problem. Synthetic data are therefore most useful when they explicitly encode which approximation is being made and under what regime it remains valid.

A central point for the entire review: in EM, realism depends on whether the synthetic image generator respects the actual imaging operator rather than merely reproducing image appearance. This is equally evident in newer modalities such as electron ptychography. In the Correlated Atomic Vibration Imaging with sub-Ångstrom Resolution (CAVIAR) framework, the forward model for 4D-STEM data is extended from a simple two-dimensional projected-phase description to multislice propagation, finite source width, incoherent averaging, and thermal diffuse scattering arising from atomic motion (Fig. 5).[53] Here, the image is not just a picture of a structure but an integrated record of beam propagation through a specimen with thickness, multiple scattering, partial coherence, and configuration-dependent diffuse intensity.

The same message appears in SEM metrology and process inspection. Monte Carlo SEM simulations remain essential because sub-resolution roughness, sidewall rounding, haloing, pits, dishing, buried defects, or other process subtleties can alter the observed signal in ways that are not obvious from geometry alone.[51] Bunday *et al*.[51] emphasize that ideal smooth structures are often inadequate surrogates for real semiconductor samples, yet physically richer simulations are computationally expensive and historically limited by usability and model complexity. In other words, the mapping from structure to image is many-to-one unless the simulator knows enough about the beam, the material, and the detector.

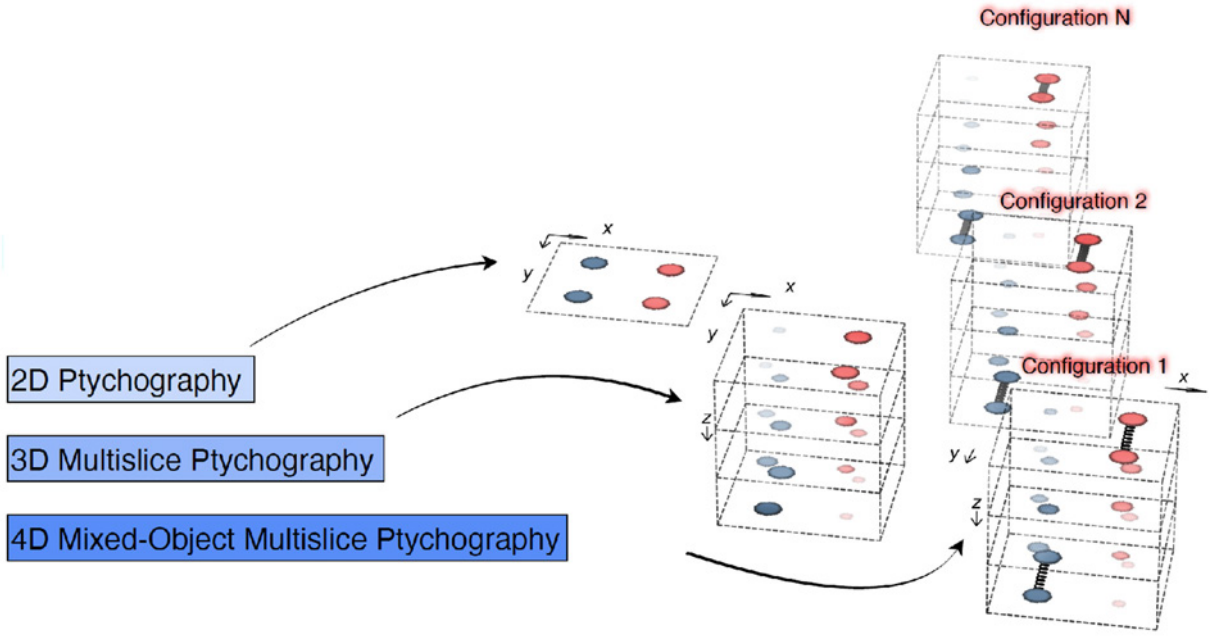


Figure 5. Schematic evolution of ptychographic reconstruction models: from 2D projected phase to 3D multislice to 4D mixed-object multislice ptychography, where multiple object states capture correlated atomic displacements. This progression illustrates that realistic forward models must go beyond simple projections to include multislice propagation, partial coherence, and thermal diffuse scattering. The figure is adapted from Gladyshev *et al*.[53] and published according to Open Access Creative Commons License CC BY 4.0.

3.2. Noise, drift, charging, contrast transfer, sampling, and detector effects

A second reason EM is demanding for synthetic data is that, in addition to the complexity of the physical model, there are factors relating to the imperfections of electron microscopy experiments. Moreover, often such deviations from ideal conditions are random by nature. For example, in porous-material SEM, common challenges explicitly include shot noise, charging artifacts, shadowing, segmentation ambiguity, and resolution constraints.[57] But in practice, the problem extends further: the noise statistics depend on frame integration, detector electronics, image encoding, beam current, defocus, drift correction, and even the image-processing pipeline built into a specific instrument.[52,58]

In cryo-EM, the same issue appears under a different vocabulary. Contrast is deliberately generated through defocus, but the CTF oscillates with spatial frequency and therefore both suppresses and inverts information in a resolution-dependent manner. Simulation studies of small proteins stress that detector performance, beam quality, phase-plate behavior, beam-induced motion, specimen overlap in ice, and ice thickness all contribute to the practical detectability of weak signals.[55] The optics model used in those simulations also includes temporal coherence, spatial coherence, detector conversion, modulation transfer function, and detective quantum efficiency, showing that synthetic data become realistic only when instrumental effects are propagated all the way to the recorded image.[55] Likewise, pixel-size calibration itself is not a cosmetic metadata issue: it changes the absolute scale of the reconstructed map and the quality of the fitted atomic model.[59]

SEM provides an especially clear example of how measurement conditions reshape both image quality and the specimen itself. Sezestre *et al*.[52] show that contour metrology in low-frame critical-dimension scanning electron microscopy (CD-SEM), limited by noise, can become less effective when noise is reduced by frame accumulation because stage drift, sample shrinkage, and charging accumulate concurrently. An important implication of this observation is that, in order to solve the problem of denoising or to develop noise-robust algorithms, it is necessary to use synthetic data generated for various noise levels, because several above-mentioned real sources of uncertainty are excluded from the simplified generator. The authors explicitly note a trade-off between reducing noise through more frame integration and increasing sample degradation through resist shrinkage and charge deposition. This is a critical point for synthetic data design: the target label may itself depend on dose history, so there is no single immutable ground truth image waiting to be measured.

In 2026, a study on detector processing in 4D-STEM adds another layer of instrument-related uncertainty.[54] Direct electron detectors have dramatically improved performance, yet accurate electron counting becomes difficult at high fluence because of Landau-distributed energy deposition, multi-pixel charge spreading, and detector noise. Back-Propagation Counting (BPC) was proposed precisely because standard counting breaks down when occupancy rises, producing distortions that propagate into diffraction contrast and real-space reconstructions (Fig. 6). Thus, even after the interaction of electrons with the sample, the detector remains an active component of the forward model rather than a passive recorder.

Scanning strategy also matters under real conditions. In STEM, raster scans are not neutral defaults: fly-back distortion and the time penalty of coil stabilization create practical limits for fast or low-dose acquisition. Alternative sparse trajectories such as spiral or Lissajous scans can reduce dose or increase speed, but then the inverse problem changes from image denoising to masked reconstruction from irregularly sampled measurements.[60] A synthetic dataset that ignores scan-path geometry may therefore teach an algorithm to solve a different problem from the one encountered experimentally.

The domain-gap problem appears most starkly in noise modeling. Rahman *et al*.[58] reported that classifiers trained on

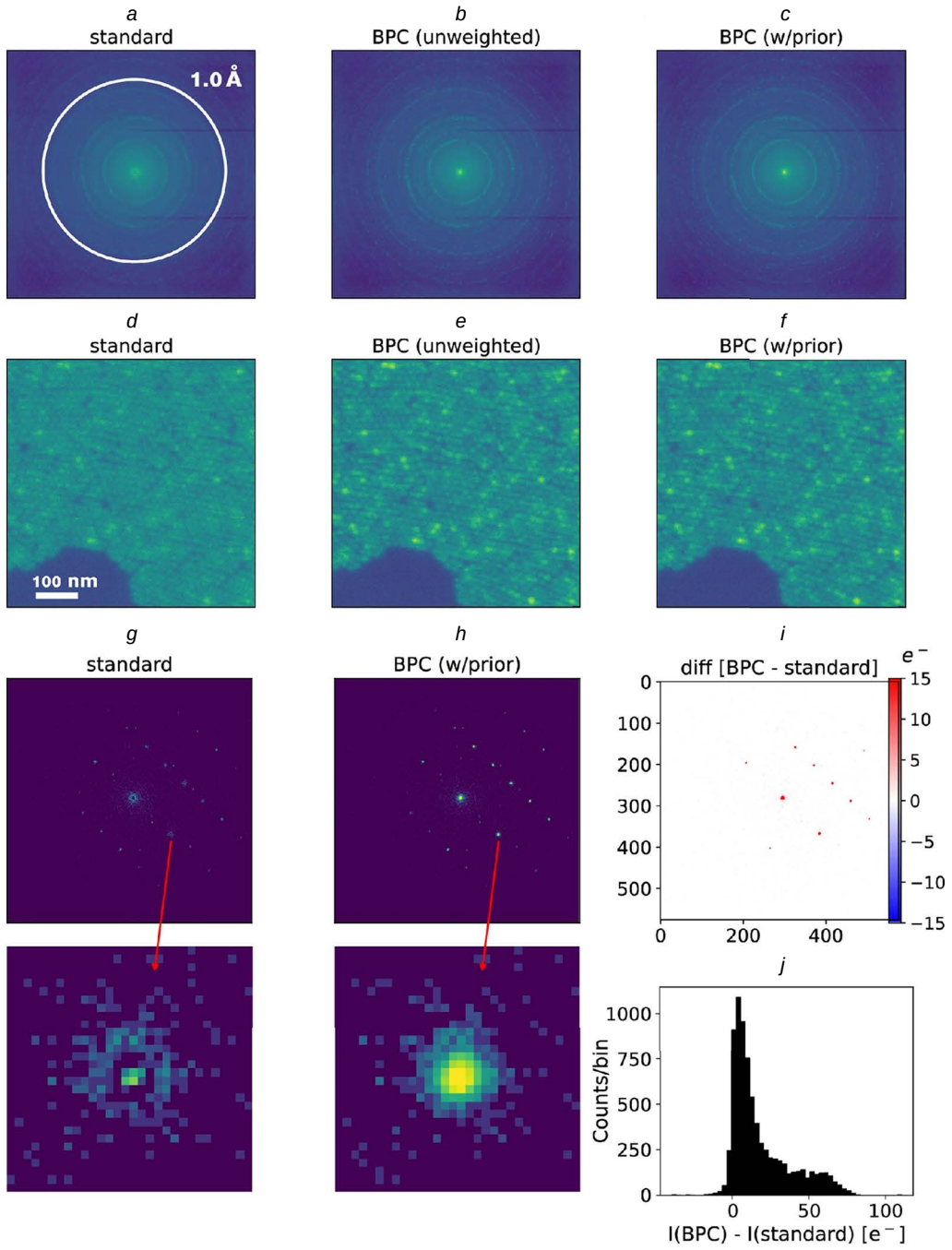


Figure 6. Summed diffraction patterns and corresponding real-space images reconstructed from 4D-STEM data using standard and BPC counting methods (*a*–*f*), artifacts in standard electron counting at high occupancy (halo around intense peaks, g) are removed by BPC (*h*), leading to improved diffraction peak intensities (*j*) and difference map $I_{BPC}-I_{standard}$ (*i*). The figure is reproduced from Renner *et al*.[54] according to Open Access CC BY-NC-ND License.

artificially added Gaussian and Poisson noise perform well on synthetic SEM data but transfer much less effectively to real Zeiss field-emission SEM images. This is an instructive warning: it is not enough for synthetic noise to be mathematically convenient; it must be instrument-faithful, context-faithful, and task-faithful.

3.3. Projection, overlap, missing information, and dimensionality reduction

A third source of difficulty is that EM images routinely discard information before any learning algorithm begins. For example, a single SEM micrograph is a two-dimensional representation of a three-dimensional object, which obviously does not allow the internal structure of sample to be seen in detail even if the subsurface layer is accessible for scanning. In particular, in the case of porous samples with an internal structure that can be examined, the two-dimensional nature of the SEM images leads to the fact that connectivity, tortuosity, and hidden overlap are only partially observable. A single SEM image is inherently limited for connectivity analysis; volumetric methods such as serial block-face SEM or FIB-SEM are needed when true three-dimensional pore architecture matters.[57] When direct 3D acquisition is infeasible, stochastic or AI-based reconstruction

from 2D data becomes attractive precisely because the experimental image is informationally incomplete.

In cryo-EM, the information loss is even more severe because each particle image is a noisy projection of one unknown conformation observed from one unknown orientation. For small proteins, the problem is exacerbated by weak scattering signal, overlap within the ice layer, air–water-interface effects, and beam-induced motion, so that alignment of the five pose parameters becomes fundamentally difficult (Fig. 7).[55] The projection image therefore conflates at least three latent factors: pose, conformation, and corruption by the microscope. Even when latent-variable methods are successful, they work on compressed low-dimensional embeddings whose interpretation is not automatic.

This is why recent cryo-EM literature repeatedly links heterogeneous reconstruction to manifold learning and latent-space analysis. Low-dimensional embeddings can be powerful because the underlying conformational landscape may indeed be intrinsically low-dimensional, but the same low-dimensional representation can also entangle conformation with pose uncertainty, orientation bias, or data quality effects.[37,61] Evans *et al*.[61] explicitly note that low signal-to-noise ratio (SNR) and uncertainty in pose assignment can be confused with conformational change in simulation-based inference pipelines. Joosten *et al*.[37] similarly emphasize that latent spaces from heterogeneous reconstruction algorithms require careful interpretation and benchmarking against known ground truth.

The broader lesson is that dimensionality reduction in EM is not merely a computational convenience but a response to irreversible information loss in the measurement itself. Hypermolecule formulations make this explicit by treating conformation as an additional hidden coordinate of the object, but they also reveal a combinatorial explosion: once state, viewing direction, shift, contrast, and other latent variables are all included, exhaustive comparison becomes computationally prohibitive.[49] Synthetic data are therefore needed not just to train discriminative models, but to probe identifiability: which latent variables can in principle be recovered from the measurements, and under what SNR, sampling, and model assumptions?

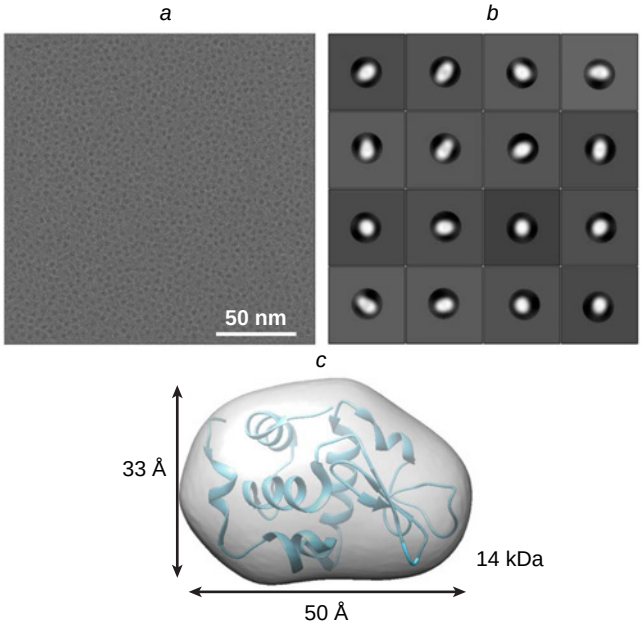


Figure 7. Cryo-EM reconstruction of 14 kDa lysozyme from simulated data without a phase plate: micrograph (*a*); blurred 2D class averages (*b*), a 16-Å resolution map, illustrating the failure of particle alignment due to insufficient signal (*c*). The figure is reproduced from Zhang *et al*.[55] according to Open Access CC BY-NC-ND License.

3.4. Why experimental labels are often incomplete, subjective, or impossible to obtain

The final reason EM is exceptionally demanding for synthetic data is that the labels required for supervised learning are often unavailable in a strict experimental sense. In many computer-vision tasks, a human can annotate what is visible. In EM, however, the scientifically relevant target may be the unobserved three-dimensional structure, the local electrostatic potential, the continuous conformational coordinate, the true particle pose, the undamaged specimen morphology, or the physically correct contour before beam-induced shrinkage. These are not labels that can simply be drawn on the image.

Sezestre *et al*.[52] make this problem concrete in CD-SEM. For synthetic images, the mask layout provides a reference contour, but for experimental images there is no equivalent ground truth; instead, the fully integrated image or an averaged contour must be used as a surrogate reference, even though integration itself changes noise and can intensify specimen modification. The paper explicitly states that the analysis becomes more complex in the absence of a ground-truth reference and that shrinkage must be treated as a source of variability. This is not a niche complication; it is a general EM phenomenon. The act of measuring can change the thing being measured.

In cryo-EM heterogeneity analysis, the problem is different but equally fundamental. Consensus reconstructions average over conformational states and blur flexible regions; discrete classification can recover a few states, but continuous motions are not naturally partitioned into a finite number of human-labeled classes.[37,49] As a result, labels such as ‘state A’ and ‘state B’ may be convenient abstractions rather than physically complete truth. Joosten and co-workers argue that heterogeneous reconstruction methods would benefit from synthetic datasets with concomitant ground truth precisely because nonlinear latent-space methods are otherwise difficult to validate, interpret, or compare quantitatively.[37]

The same argument extends to mechanistic and dynamical inference. In early work on stochastic macromolecular complexes, Wang and Doerschuk[56,62] proposed mapping cryo-EM-derived statistics to a mechanical model, but the only rigorous way to test recovery of such latent physical variables was to begin with synthetic data where the underlying system is known. Here, the desired label is not a visible structure in one image, but a latent dynamical law behind an ensemble of noisy projections.

Synthetic data are therefore indispensable in EM not because experiments are unimportant, but because experiments alone rarely expose complete truth. They provide the only practical route to benchmark particle picking under known positions, evaluate orientation bias, test recovery of continuous trajectories, quantify robustness to poor SNR or sparse sampling, and separate algorithmic success from accidental agreement with a particular microscope setting.[37,52,60] At the same time, the cited literature also warns that synthetic labels can be misleading if the generator omits exactly those nuisances that dominate real experiments, such as realistic ice, detector response, scan-path distortions, charging, or process subtleties.[51,55,58]

Taken together, these studies show that EM is a demanding domain for synthetic data because the challenge is structural, not cosmetic. The image is generated by coupled physics; corruption is acquisition-dependent; missing information is intrinsic; and ground truth is often latent, dose-dependent, or experimentally inaccessible. The next step is to distinguish the main generation strategies, which briefly discussed above: fully simulated, semi-synthetic, augmented, and generative and to ask which level of physical fidelity is necessary for each intended task. Only by making that distinction can synthetic data serve as a reliable bridge between EM image formation and downstream inference rather than as a source of visually plausible but scientifically fragile examples.

4. Strategies for generating synthetic data

The central problem in synthetic data generation for electron microscopy is not merely to create more images, but to create images whose variability is aligned with the downstream scientific task. In this sense, the literature surveyed here converges on a useful organizing principle: every synthetic-data pipeline makes an explicit choice about what should be modeled mechanistically, what can be approximated statistically, what may be learned by a generative model, and which parts of the annotation burden can be shifted from humans to algorithms. The strategies are therefore best viewed not as competing categories but as a spectrum, ranging from first-principles forward simulation to geometry-driven rendering, to generative realism enhancement, to semi-synthetic and pseudo-labeling workflows that exploit limited real data.[51,63−66]

4.1. Physics-driven forward models

Physics-driven forward models begin from the premise that the most reliable labels are those that are generated together with the image itself. If the specimen geometry, composition, thickness, beam conditions, optics, detector response, or acquisition sequence are known or controlled, then the image can be synthesized together with ground-truth masks, particle coordinates, thickness values, or structural parameters. This logic underlies classical simulation-based work on SEM imaging, where stochastic 3D models of fibrous media were coupled to SEM image formation to validate fiber-thickness measurement under conditions where experimental truth is fundamentally inaccessible.[63] The same philosophy is now reappearing in more modern forms: process-aware SEM simulation that explicitly targets subtle acquisition effects,[51] physics-based synthetic data models for automated segmentation in catalysis microscopy,[67] and robust high-resolution TEM synthetic-data frameworks in which atomic structure, imaging parameters, and target labels are generated in a single forward pipeline.[64]

A major strength of such forward models is controllability. One can sweep beam energy, defocus, specimen orientation, particle loading, phase contrast, detector conditions, or artifact severity over parameter ranges that would be impractical to obtain experimentally. In cryo-EM, this idea has been pushed particularly far. CryoGEM combines controllable physical simulation of single-particle cryo-EM with generative noise translation so that particle positions and poses remain physically meaningful while the final micrographs acquire realistic experimental noise characteristics.[65] CryoCCD extends this direction further by coupling biophysical modeling of heterogeneity with conditional cycle-consistent diffusion for cryo-EM synthesis, thereby treating realism not as an aesthetic objective but as a task-aware property needed for particle picking and pose estimation.[68] In cryo-electron tomography, FakET similarly starts from simulated structural content and then applies neural style transfer to transfer the appearance of experimental tomograms, thus preserving controllable geometry while moving synthetic images closer to the visual statistics of real data.[40]

The recent literature also shows that pure first-principles models are increasingly being augmented with learned surrogates. A good example is the multislice surrogate approach for NP size estimation from STEM, where a learned model accelerates the otherwise expensive image-formation calculation while retaining a strong link to the underlying physics.[69] Likewise, in focused ion beam (FIB) tomography of nanoporous gold, Monte Carlo-based synthetic data remain physically meaningful but do not fully reproduce the appearance of real images; domain-adaptation or generative refinement is therefore introduced as an additional stage after simulation.[70] This hybridization is important because it marks a shift in emphasis: the best physics-driven models are no longer judged only by their physical rigor, but by whether they provide the right invariants for machine learning. In other words, forward modeling now serves both as a scientific simulator and as an annotation engine.

Figure 8 exemplifies an important principle of physics-aware rendering workflows: one does not need to simulate every electron event from first principles if the rendering pipeline preserves the geometric and contrast relations that are decisive for the downstream task. This is why synthetic image rendering for NP segmentation,[71] artificial SEM image creation from simulated particle assemblies followed by generative adversarial network (GAN) refinement,[72] and SEM particle-detection pipelines based on 3D particle models and background extraction[73] all occupy an intermediate but strategically important space between explicit physics and purely data-driven synthesis.

4.2. Geometry- and microstructure-driven synthetic data

A second major strategy starts not from a full microscope simulator, but from the morphology of the specimen itself. Here, the key assumption is that for many tasks the dominant bottleneck is not the lack of exact beam physics, but the lack of representative structures, shapes, textures, spatial organizations, and masks. This idea is especially powerful in materials microscopy, where the scientific object of interest is often a microstructure rather than a single immutable image. Virtual microstructure design with GANs has already shown how microstructures can be sampled in image space as design variables rather than reduced to a few handcrafted descriptors.[74] Deep learning for synthetic microstructure generation in a materials-by-design framework makes a similar point at a broader scale, treating generative models as tools to explore structure space while preserving task-relevant statistics.[75]

Within electron microscopy, geometry-driven synthesis has evolved into several practical forms. One line of work relies on explicit rendering of particles, grains, or pores. Synthetic image rendering for NP segmentation uses limited real examples to estimate particle size, shape, and distribution, then creates large synthetic corpora with paired masks.[71] UTILE-Gen generalizes this philosophy into a configurable synthetic dataset generator for nanoscience image analysis and shows that such rendered

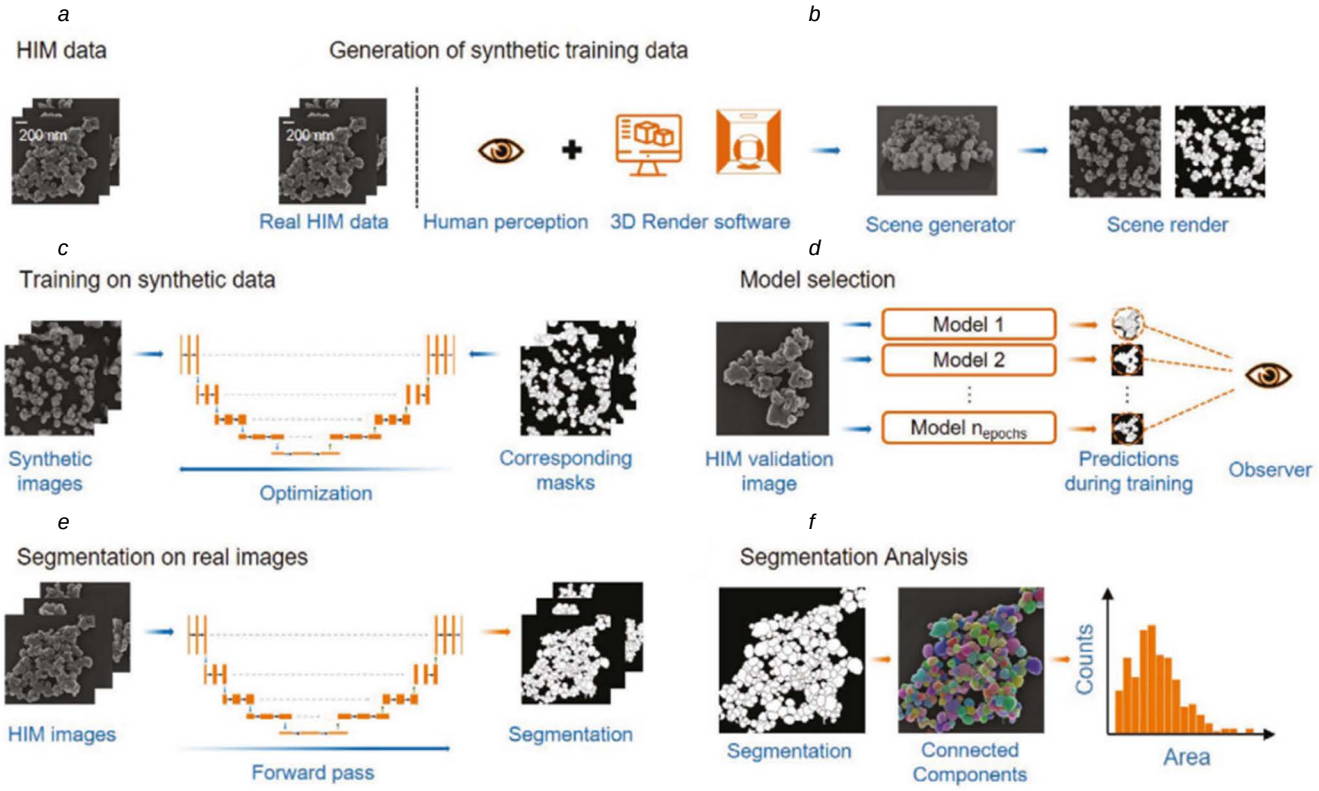


Figure 8. Synthetic image rendering workflow for NP segmentation, showing how limited real microscopy images are transformed into rendered synthetic images with automatically generated masks and then used for training and deployment on real data (*a*–*f*). HIM is helium ion microscope. The figure is reproduced from Mill *et al*.[71] according to Open Access Creative Commons Attribution License.

synthetic data can improve both classification and instance segmentation in low-data regimes.[41] Semi-synthetic SEM particle detection likewise relies on artificial placement of NP instances to train detection networks when annotated real images are scarce.[76] For fibrous materials, semi-supervised large-scale fiber detection has combined synthetic geometric modeling with domain adaptation so that the learned detector remains robust to degradation and crowding.[77]

A related line of work is microstructure-driven rather than object-driven. Crystal-grain analysis using diffusion models starts from concept-oriented synthetic generation of grain structures and uses that structured prior to support grain-boundary completion and analysis.[78] Micrograph-annotation pairs for steel microstructure recognition can be produced with hybrid deep generative models even in extremely small and imbalanced datasets.[79] Synthetic dual image generation with a vector-quantized variational autoencoder based (VQ-VAE) workflow shows that a few labeled micrographs can be reorganized into a generative source of both images and masks, explicitly targeting semantic segmentation under limited data and imperfect labels.[66] At the minimalist end of the spectrum, even self-generating simple figure patterns can serve as a surprisingly effective source of training data for binary microscopy segmentation, provided the target task depends more on topology and foreground-background separation than on fine texture realism.[80]

The conceptual value of geometry- and microstructure-driven synthesis is that it externalizes prior knowledge. Instead of asking a network to infer particle overlap, grain impingement, pore morphology, or microstructural connectivity from a handful of examples, the synthetic generator encodes those regularities explicitly. This can make training more stable, reveal which aspects of morphology are actually necessary for good generalization, and expose where geometry alone is insufficient because image realism remains too idealized. That final limitation, in turn, motivates the next family of strategies: generative AI methods designed specifically to close the simulation-to-reality gap.

4.3. Generative AI for synthetic microscopy data

Generative AI is increasingly used to improve the experimental plausibility of synthetic microscopy images. The literature shows three main uses. First, generative models can directly synthesize new microstructures or microscopy-like images from learned distributions. Second, they can translate simulated images into more realistic ones while retaining labels from the source domain. Third, they can act as fast surrogates for expensive forward models. Across SEM, STEM, FIB-SEM, and cryo-EM, GANs, cycle-consistent generative adversarial networks (CycleGANs), variational autoencoders (VAEs), diffusion models, and style-transfer methods are now being used not simply to increase visual realism, but to preserve task-relevant morphology while injecting appropriate texture, noise, and artifact statistics.[65,68,72,81]

For SEM and STEM, the generative-AI literature consistently points to the same challenge: clean simulations or rendered scenes usually contain the correct objects but the wrong image statistics. Artificial SEM images created from simulated particle assemblies become substantially more realistic after GAN-based conversion, making them more useful for training than raw rendered scenes alone.[72] GANs have also been used to create realistic scanning transmission electron microscopy images from atomistic or simulated inputs,[81] while CycleGAN-based

translation has been used to bridge experimental STEM images and simulated structural representations in supported catalyst NPs, enabling more precise extraction of structural information than direct image analysis.[82] A similar theme appears in FIB-SEM, where GAN-based artifact generation is used to reproduce curtaining, one of the most characteristic and algorithmically important degradations in synthetic FIB-SEM data.[83]

Diffusion models are now entering this space directly for TEM. A recent semiconductor-metrology study used progressive patch-to-image denoising diffusion probabilistic model training to generate high-resolution synthetic TEM images from as few as 15 real images, while also supporting domain transfer, guided generation, inpainting, and segmentation from learned diffusion features.[84] This study shows that generative synthesis can remain useful under extreme data scarcity when structural and noise statistics are evaluated explicitly rather than by visual appearance alone.

The cryo-EM literature is particularly instructive because realism cannot be reduced to style alone. CryoGEM explicitly combines physical cryo-EM simulation with unpaired noise translation so that annotations remain controllable while the output micrographs better match experimental noise and contamination patterns.[65] CryoCCD extends this idea with conditional cycle-consistent diffusion and richer biophysical modeling of heterogeneity, making the generator not only more realistic but also more controllable.[68] FakET, in turn, uses neural style transfer to adapt synthetic cryo-electron tomograms to the appearance of unlabeled reference data, thereby reducing the cost of generating benchmark tomograms for segmentation and analysis.[40] Together, these studies show that generative AI is most effective when it is anchored by a physically or geometrically meaningful latent source rather than asked to hallucinate scientific content *de novo*.

This point also explains why theory- and method-oriented contributions remain important. Manifold-regularized GAN argues that scientific data generation benefits from manifold-aware regularization because the generator must respect the geometry of the real scientific data distribution rather than merely deceive a discriminator.[43] Reviews and perspective articles in the materials-microscopy space further emphasize that generative models should be evaluated by distributional fidelity, morphological plausibility, and downstream utility rather than by generic visual metrics alone.[64,81,85] In practice, this means that the role of generative AI is not to replace simulation, but to transform simulation into task-effective data.

Figure 9 captures the broader methodological shift in the field: the most successful generative pipelines are not purely image-to-image translators, but composite systems in which a forward model specifies what should remain scientifically invariant and a generative model learns what should be adapted to match the empirical data distribution. This hybrid logic is now appearing far beyond cryo-EM, including domain-adapted FIB tomography,[70] realistic STEM generation,[86] and phase-field-to-SEM translation for labour-free segmentation.[87]

Beyond technical challenges, the use of generative AI for synthetic microscopy data also raises ethical and reproducibility concerns. As recently highlighted by Davydiuk *et al*.,[88] AI-generated microscopy images can now be indistinguishable from real ones even for expert observers, creating risks of data fabrication and undermining research integrity. While generative models remain valuable for controlled synthetic-data generation, the community must develop transparent practices − including clear provenance tracking, open benchmarks, and rigorous validation — to preserve the trustworthiness of EM data and analysis workflows.

4.4. Semi-synthetic and pseudo-labeling approaches

Semi-synthetic and pseudo-labeling strategies address a different but equally important problem: in many microscopy problems one has some real images, some weak labels, some unlabeled

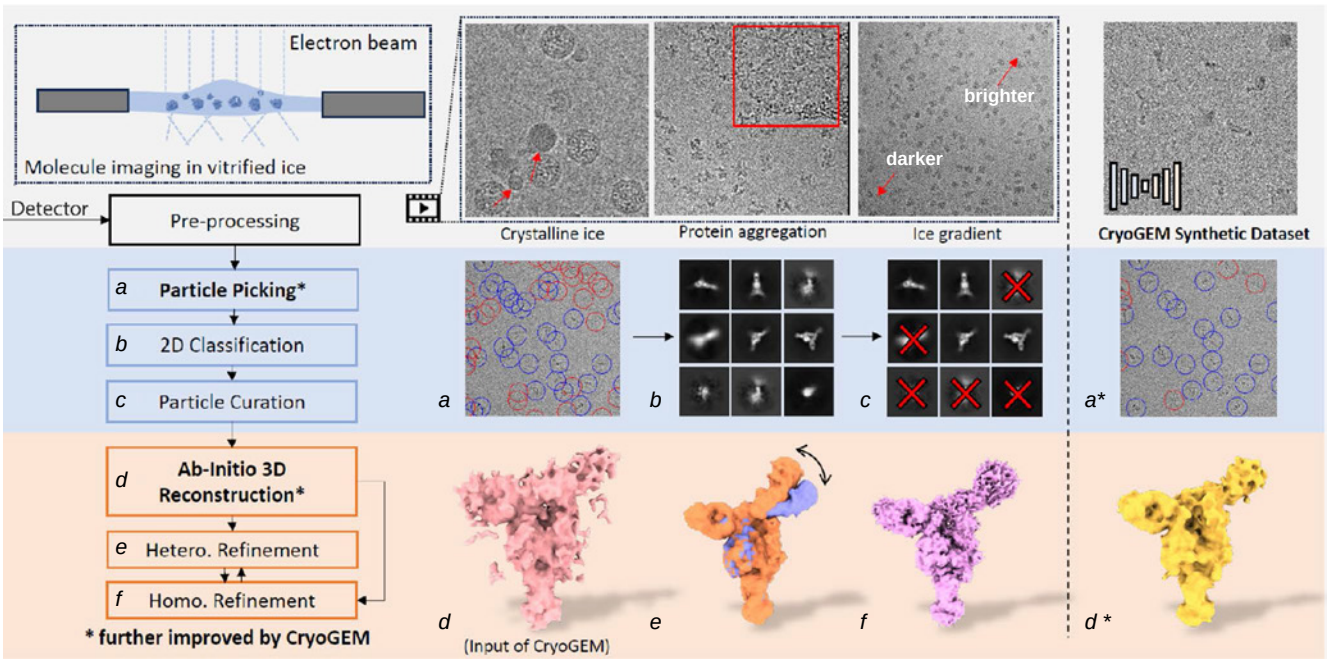


Figure 9. Physics-informed generative cryo-EM workflow in which controllable simulated cryo-EM data are combined with learned noise translation to generate realistic annotated micrographs for downstream cryo-EM analysis. The figure is adapted from Zhang *et al*.[65] and published according to Open Access Creative Commons License CC BY 4.0.

data, and perhaps a limited simulator, but not the complete paired dataset needed for conventional supervised learning. These methods therefore treat synthetic data not as a full replacement for experiments, but as an intermediary representation that transfers information from one domain to another. Weakly supervised instance segmentation of SEM images *via* synthetic data is an early and clear example of this logic, where synthetic images compensate for the lack of dense instance labels in real SEM data.[89] NP detection on SEM images using semi-synthetic training data takes a comparable approach, constructing training samples that are artificial but sufficiently structured to learn a useful detector.[76]

More recent work makes the transfer mechanism itself the main object of design. Synthetic dual image generation for semantic segmentation explicitly generates micrographs together with masks, thereby reducing labeling effort while keeping the segmentation target tied to the visual synthesis process.[66] Physics-informed generative AI for labour-free segmentation uses phase-field simulations to obtain intrinsically correct masks and then applies CycleGAN translation so that the images resemble real SEM data while the labels remain exact.[87] For high-resolution materials characterization, Mask Region-based Convolutional Neural Network (Mask-RCNN) training on synthetic TEM data,[85] robust high-resolution transmission electron microscopy (HRTEM) synthetic-data generation,[64] and physics-based catalysis-microscopy segmentation models[67] all exemplify the same principle: use synthetic labels where they are exact, and use real data where they are visually authoritative.

Semi-supervised extensions push this one step further by allowing unlabeled real data to shape the training process. Large-scale fiber detection in material images uses synthetic data together with domain adaptation to reduce annotation effort while improving robustness to real-image degradation.[77] EMDiffuse, although framed primarily as a restoration method, is also relevant here because it uses carefully aligned multi-dose EM data to learn realistic image enhancement under conditions where direct ground truth is expensive and imperfect.[90] In autonomous experimentation, active-learning surrogates that integrate EM with simulation data point toward a future in which synthetic and experimental data are updated iteratively rather than being prepared once and used statically.[91]

The practical lesson from this literature is that pseudo-labeling works best when domain shift is explicitly managed. If the shift is small, synthetic labels can directly bootstrap a useful model. If the shift is moderate or severe, one needs either domain adaptation, generative translation, or a task-specific coupling between simulation and experiment (Fig. 10). This is why the field is moving toward pipelines rather than single generators. The synthetic-data engine, the realism adapter, and the learner are increasingly co-designed. The result is not only better predictive performance, but a more transparent decomposition of what is known from physics, what is assumed from morphology, and what is inferred from data.

Taken together, the literature on synthetic-data generation in electron microscopy shows an emerging methodological convergence. Forward models provide control, interpretability, and exact labels; geometry- and microstructure-driven generators

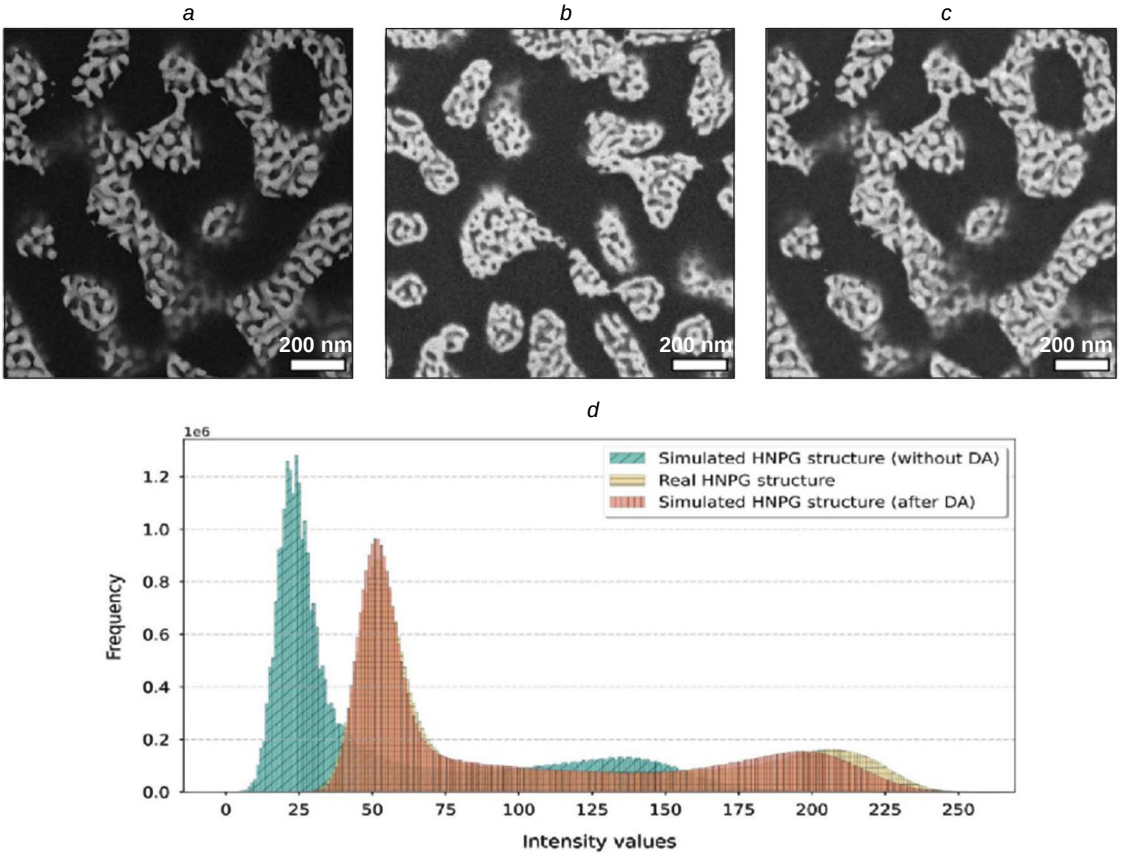


Figure 10. Example of domain adaptation for FIB tomography, comparing simulated data, real data, and domain-adapted synthetic data, with corresponding intensity histograms that show improved alignment to the experimental distribution. HNPG is hierarchical nanoporous gold; DA is domain adaptation (*a*–*d*). The figure is reproduced from Sardhara *et al*.[70] according to Open Access Creative Commons License CC BY 4.0.

provide efficient, scalable variation in morphology; generative AI supplies realism, domain adaptation, and artifact synthesis; and semi-synthetic or pseudo-labeling workflows exploit the limited experimental data that are actually available.[41,63,66,81] The most impactful studies do not choose one of these strategies in isolation, but combine them into layered pipelines. Once the generation strategies are understood in general terms, it becomes possible to ask how they perform in the most practically important modality-specific settings. Scanning electron microscopy is the logical starting point because it has been the earliest and broadest testbed for rendered, simulated, semi-synthetic, and domain-adapted data, especially for morphology quantification, particle detection, and segmentation.

5. Synthetic data for scanning electron microscopy

Scanning electron microscopy is the most natural proving ground for synthetic data in electron microscopy because many of the downstream questions are simultaneously practical, quantitative, and annotation-hungry. SEM is routinely used to count particles, measure diameters and aspect ratios, segment pores and fibers, detect process defects, reconstruct topography, and increasingly infer hidden process or materials variables from image appearance. Setting aside the qualitative description of individual images, and focusing instead on developing universal algorithms for processing electron microscope data for widespread use, it can be said that in those use cases the real bottlenecks are difficult to resolve using experimental data alone: true ground truth is often unavailable, manual annotation is slow and subjective, rare events are underrepresented, and imaging conditions vary strongly across materials, detectors, magnification regimes, charging states, and sample preparation protocols. As a result, the SEM literature shows a particularly rich spectrum of synthetic-data strategies, ranging from fully simulated fiber packs used to validate morphometry algorithms to semi-synthetic and photo-realistic rendering pipelines for NP segmentation, and from domain-adapted semiconductor metrology to direct structure-property prediction from synthetic microstructures.[45,63,71,73,92]

The development across this literature is not merely a gradual improvement in model accuracy. More fundamentally, synthetic SEM data have changed the role of the image itself. In early work, synthetic data were mainly a validation device: they made it possible to test whether a measurement algorithm could recover a known geometric truth under realistic overlap and projection artifacts. In newer studies, synthetic data became a training substrate for dense segmentation, instance detection, regression, and sim-to-real transfer. In the most recent studies, synthetic SEM-like images are no longer only substitutes for missing labels; they are becoming a compact representation of prior knowledge about morphology, process physics, and even the target variable to be predicted. Seen in this way, the SEM field already spans much of the trajectory from measurement validation to increasingly autonomous image understanding and predictive metrology.[46,93–95]

5.1. Synthetic SEM for morphology quantification

Synthetic data entered SEM analysis first through the problem of trustworthy measurement. A representative example is the automatic thickness measurement of fibers in fibrous filter media, where the decisive challenge is that the true fiber thickness needed as a reference cannot be extracted directly from SEM images using any other complementary methods because of projection issues, fibers overlapping, bundle formation and so on. Altendorf *et al*.[96] addressed this by building synthetic three-dimensional fiber systems with controllable bundle formation and realistic overlap, and then using these images to validate automated thickness measurement. Their analysis showed that the morphological measurement pipeline reproduced the peaks of true thickness distributions with approximately ±1 pixel accuracy for thin fibers, while systematic deviations emerged for thicker fibers because overlap effects are intrinsic to the SEM projection itself.[63] This is an important conceptual starting point for the whole field: synthetic SEM data do not only enlarge training sets, they provide the only route to measurement traceability when the experimental image does not contain a directly observable truth.

Once this principle was established, morphology quantification broadened from simple geometric measurement to richer descriptors of heterogeneous microstructure. In fibrous systems, synthetic or augmented SEM data are now used not only to isolate fibers, but also to quantify their lengths, widths, and orientation distributions. Instance-segmentation-based measurement of short carbon fibers was demonstrated with additional synthetic training in carbon-fiber-reinforced systems,[97] while a later pipeline combined Segment Anything Model (SAM) with DeepLabV3+ model to segment short carbon and glass fibers using very few real labels and additional synthetic data, increasing intersection over union (IoU) and pixel accuracy and yielding end-to-end statistics for geometric fiber characterization.[98] These studies are methodologically significant because SEM morphometry is rarely limited by classification alone. The real output of interest is typically a distribution: lengths, diameters, aspect ratios, area fractions, or orientation metrics. Synthetic data become valuable precisely when they stabilize these distributions against overlap, partial visibility, and background variability.

A parallel line of work concerns orientational order and anisotropy. In graphene oxide membranes, synthetic-data-trained U-shaped convolutional network (U-Net) segmentation was combined with the OrientLayer algorithm to quantify lamellar ordering directly from cross-sectional SEM micrographs, linking image analysis to materials-level descriptors of orientational order.[99] An even more explicit shift from descriptive to quantitative morphology appears in short-fiber composites produced by material extrusion large-format additive manufacturing. There, Python-generated synthetic SEM-like images with known second-order orientation tensors were used to train a convolutional neural network (CNN) that directly regressed tensor components from the micrographs; the reported cross-validation performance approached $R^2 \approx 0.989$ on a dataset of 40 000 synthetic images.[95] Instead of fitting individual fibers one by one, the image is treated as a statistical field whose global appearance encodes the latent orientation tensor. This is a key step in the evolution of synthetic SEM: morphology is no longer merely segmented and counted, but mapped to a physically meaningful state variable.

The NP literature adds another dimension to morphology quantification: the transition from conventional image processing to learning-based detection under severe crowding (see Fig. 8). Semi-synthetic YOLO-based (YOLO is you only look once) workflows were used to detect cubical and quasi-spherical NPs in SEM images,[76] while weakly supervised and instance-segmentation variants were developed for particle morphometry and counting using minimal manual labels.[89,100] A more recent study targeted topological NP structures and

showed that carefully designed combinations of synthetic data and augmentation can improve U-Net segmentation metrics relative to augmentation alone, highlighting that synthetic diversity is not interchangeable with generic image augmentation.[101] Across these studies the morphology problem is not simply 'find the object', but 'recover the correct ensemble statistics despite overlap, blur, and class imbalance'. Synthetic data are attractive in SEM exactly because they can tune these nuisance factors in a controlled way.

The recent review on machine learning for SEM-based pore analysis confirms that these observations generalize far beyond NPs and fibers. It emphasizes recurring difficulties in porous-material SEM analysis, *viz*., noise, shadowing, insufficient contrast, segmentation ambiguity, variable imaging conditions, and limited labeled datasets, and therefore frames synthetic or otherwise model-guided training data as part of a broader transition from manual, image-by-image interpretation to automated, reproducible quantification.[57] Taken together, the morphology-quantification literature shows that synthetic SEM data are most useful when the target is a robust quantitative descriptor of a heterogeneous ensemble rather than the classification of a single isolated object.

5.2. Synthetic SEM for detection and segmentation

If morphology quantification established why synthetic SEM data are valuable, detection and segmentation showed how they could be operationalized at scale. The central obstacle is the annotation problem. Real SEM images are expensive to acquire, difficult to annotate, and often contain dense overlap, ambiguous boundaries, and material-specific contrast behavior. A landmark study by Mill *et al*.[71] demonstrated that photo-realistic rendering can effectively solve this bottleneck for NP segmentation. Using a limited number of real helium ion microscope reference images, which, at first glance, resemble SEM microphotographs, together with prior information on size, shape, and distribution, they rendered large synthetic datasets in Blender and automatically generated pixel-perfect masks. A U-Net trained on these synthetic images transferred successfully to real images and enabled downstream particle-statistical analysis. The importance of this work lies not only in its strong performance, but in its workflow logic: the real image serves as a reference for scene design, the renderer injects controlled diversity, and the synthetic mask is obtained for free.

Closely related strategies appeared in SEM-specific NP detection. Kharin[73] used real SEM textures combined with rendering to create semi-synthetic data for NP detection, arguing that transfer learning plus synthetic imagery can sharply reduce the amount of manual labeling required. NP detection with semi-synthetic YOLO training further demonstrated that mixed real/synthetic datasets can generalize to images not seen during training, although close particle proximity remained a consistent challenge.[76] Weakly supervised instance segmentation reduced manual burden even further by requiring only one labeled instance per raw SEM image to construct synthesized training data, yielding about 75% recall and 74% average precision in a difficult particle-segmentation setting.[89] A subsequent Modified BlendMask pipeline addressed false negatives caused by overlap, blur, sparse distributions, and very small particles, improving automated NP counting and reaching an overall count accuracy of 70.2% across 38670 particles in 1141 test images.[100] These studies collectively show that for SEM detection tasks the decisive issue is not the absolute photorealism of any single synthetic image, but whether the generated population spans the failure modes of the detector.

The same principle now extends to more application-driven segmentation problems. In aluminide diffusion coatings on steel, synthetic SEM images created in Blender were combined with manually refined Weka labels to train a U-Net for multilayer segmentation and metallographic evaluation (Fig. 11). The model trained on synthetic plus real SEM data achieved mean Dice scores of 98.7%±0.2 for the Fe_2Al_5 layer, 82.6%±8.1 for pores, and 81.48%±3.6 for precipitates on manually labeled SEM images.[102] Here the impact of synthetic data is especially clear because the segmentation output immediately feeds quantitative metallurgy: layer thickness, pore-line position, pore fraction, and precipitate fraction. In other words, segmentation is not the end point but the front end of an automated measurement chain.

A particularly data-efficient SEM example is a recent morphology-resolved SEM-EDS (energy-dispersive X-ray spectroscopy) framework for MXene analysis, in which only five fully human-labeled SEM images were expanded by a canvas-based generator into approximately 12000 labeled objects. Adding the synthetic data increased mean IoU from 0.623 to 0.743, while the resulting masks were coupled directly to EDS maps for morphology-specific compositional analysis.[103]

Synthetic SEM detection is also expanding toward defect metrology in semiconductor manufacturing. In photolithographic line patterns, artificially generated SEM images with known defect distributions were used to train deep learning models for bridge and break defect detection. The best-performing YOLOv8 model achieved a mean average precision of 96% on synthetic evaluation and still transferred meaningfully to real SEM data, correctly detecting 84.6% of bridge defects and 78.3% of break defects across relevant instances.[104] At the same time, more model-based SEM metrology studies indicate the same broader tendency. Synthetic variability was injected into calibration data to understand how metrology uncertainty propagates into electron-beam proximity-effect correction models,[105] while contour extraction for very low-frame SEM images was framed explicitly as an enabler for metrology under acquisition constraints.[52] Even when the final method is not a pure deep-learning pipeline, the synthetic-data mentality remains the same: model the measurement process, expose the system to controlled variability, and quantify robustness before deployment.

One broad conclusion emerges from these detection and segmentation studies. In SEM, synthetic data rarely replace real data completely; instead, they shift scarce real data to a higher-value role. Real images define appearance priors, validate domain realism, calibrate noise and contrast, or serve as a small anchor set for fine-tuning and model selection. The heavy combinatorial lifting, namely, size distributions, overlap statistics, defect rarity, orientation spread, background diversity, and class balancing, are then delegated to the generator. This hybrid logic is visible across photo-realistic rendering,[71] semi-synthetic transfer,[73] weak supervision,[89] augmentation-plus-simulation,[101] and application-specific process metrology.[102,104]

5.3. Synthetic SEM for inverse and predictive analysis

The most mature direction of synthetic SEM goes beyond object recovery and asks a harder question: what hidden variable can be inferred from image appearance? This is where SEM synthetic data become a bridge between image analysis and inverse modeling. One instructive example is SEM analysis of air-

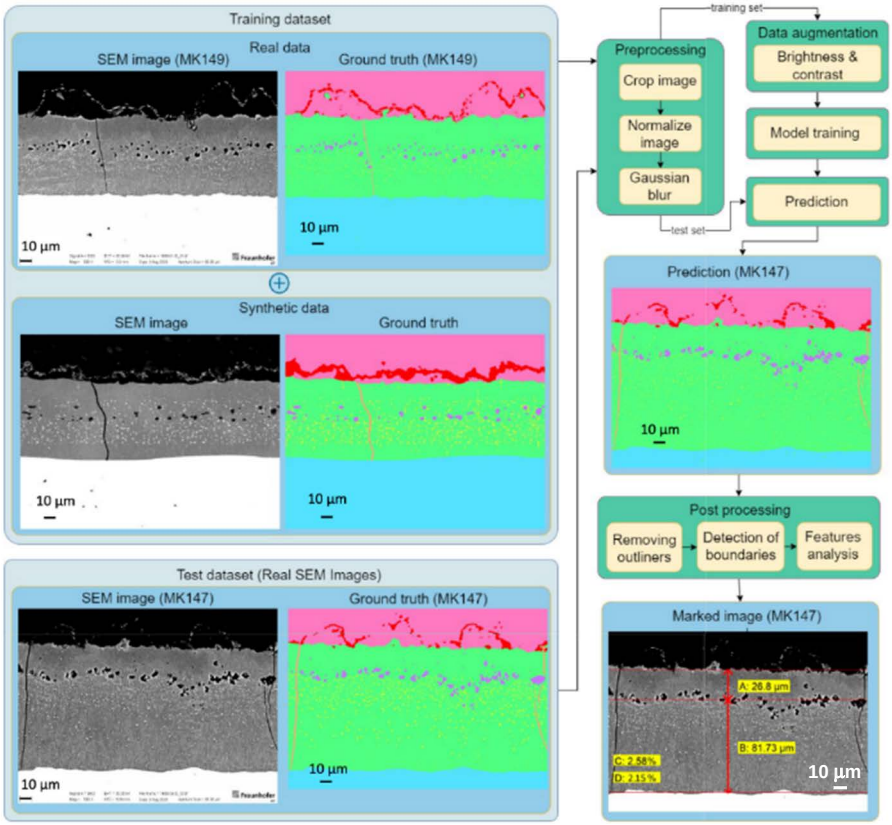


Figure 11. Example of a synthetic-data-assisted workflow for SEM segmentation and quantitative metallographic evaluation, combining real data, synthetic data, preprocessing, prediction, and post-processing into measurable structural outputs. MK147/149 are slurry composition deposited on steel. The figure is reproduced from Strakos and co-workers [102] according to Open Access Creative Commons License CC BY 4.0.

filtration media. Beckman *et al*.[92] created simulated nonwoven filter media whose grayscale renderings mimicked SEM images and used analytical single-fiber-efficiency calculations to provide target labels. A CNN trained entirely on this synthetic SEM-like corpus predicted filtration efficiency and airflow resistance from image appearance, demonstrating that synthetic data can directly encode a structure-property mapping rather than only a segmentation label. This is a conceptual milestone because the image is used here as a compressed descriptor of morphology from which functional performance is inferred.

A closely related inverse problem is the recovery of latent orientation or shape from two-dimensional SEM views. The additive-manufacturing study dedicated to the analysis of the orientation of short filler fibers in a polymer matrix based on electron microscopy data is emblematic. Determining the reference values of the orientation tensor components based solely on experimental SEM data is a laborious task; however, inverting the logic of obtaining reference points by generating synthetic microphotographs for predefined tensor components significantly simplifies the whole process. Synthetic SEM-like images with known tensor labels allow direct regression of the second-order fiber orientation tensor without explicit object-by-object fitting (Fig. 12).[95] Likewise, SEM image analysis of graphene oxide membranes used synthetic-data-trained segmentation to quantify orientational ordering from cross sections.[99] In robotics-oriented microscopy, sim-to-real transfer learning was used for micro-object pose estimation under the constraint that pose is fundamentally three-dimensional while imaging is predominantly two-dimensional.[46] These studies indicate that once a generator can span the relevant appearance manifold, synthetic SEM data become a practical way to supervise variables that are difficult or impossible to label exhaustively in experiment.

Inverse analysis becomes even more demanding when the target is three-dimensional topography or full surface geometry. Early work on multi-view SEM reconstruction already relied on synthetic and real data together to validate extrinsic calibration and three-dimensional surface recovery.[106] The broader methodological landscape was then surveyed in a review of 3D SEM surface reconstruction,[93] followed by 3DSEM++, which emphasized adaptive and intelligent reconstruction strategies.[94] High-resolution reconstruction from multiple views by feature-distance minimization[107] and more recent deep-learning approaches further demonstrate that SEM inverse problems are naturally coupled to simulation, because geometric ground truth is extremely costly to acquire directly.

This coupling is clearest in semiconductor metrology. Houben *et al*.[45] proposed a training procedure for SEM-based three-dimensional surface reconstruction that used simulated data together with unsupervised domain adaptation to bridge the gap between synthetic and experimental images. Their framework predicted height and critical-dimension information for fin-like structures and reported highly accurate interwafer classification with a root-mean-square error (RMSE) of 0.67 nm for depth-related prediction, showing that simulated SEM data can be transformed into viable z-metrology for manufacturing. On the related topic of extraction three-dimensional structural information from microscopy images, synthetic-data-driven deep center regression and geometric clustering for modeling and analysis of 3D fibrous objects similarly illustrate how geometry-aware priors and synthetic supervision can stabilize instance-level reconstruction in data-

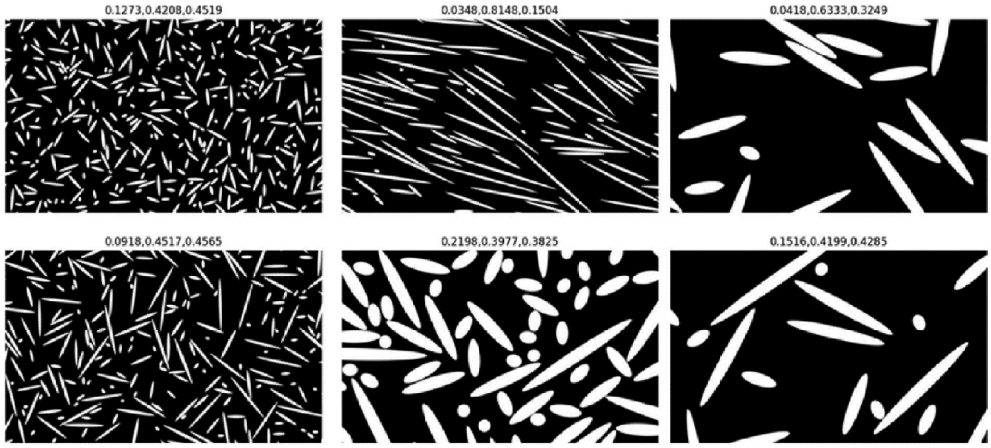


Figure 12. Examples of synthetic SEM-like micrographs with different fiber-orientation states used to supervise direct regression of orientation-tensor components from image appearance; the triplets above each panel are the corresponding second-order orientation tensor components (A_{xx}, A_{yy}, A_{zz}) for the fiber arrangement shown. The figure is reproduced from García-Gascón *et al*.[95] according to Open Access Creative Commons License CC BY 4.0.

rich microscopy settings.[108] These examples reinforce the same lesson: the harder the inverse problem, the more valuable the simulator becomes — not just for augmentation, but for defining the target variable itself.

From this perspective, Section 5 reveals an important progression. Synthetic SEM began as a way to validate whether a measurement was trustworthy; it has evolved into a framework for generating labels, balancing rare events, transferring across domains, and learning mappings from image morphology to latent structural or process variables. The most successful studies are those in which the synthetic generator is tightly aligned with the intended output: bundle-aware fiber models for diameter estimation,[63] scene rendering for NP masks,[71] process-aware coating textures for metallography,[102] controlled orientation ensembles for tensor regression,[95] and simulated media linked to analytical filtration theory for property prediction.[92] Conversely, when the generator is weakly tied to the real nuisance structure of SEM such as overlap, blur, charging, detector-specific contrast, or out-of-focus rare defects, performance gains become more limited or more task-specific.[76,109]

In SEM, synthetic data have already proved their value for mesoscopic surface morphology, defect metrology, and inverse inference from grayscale projections. However, many SEM tasks still rely on comparatively simplified image-formation assumptions, geometric priors, or empirical appearance matching. In transmission electron microscopy and scanning transmission electron microscopy, by contrast, the forward model becomes more tightly bound to electron scattering physics, thickness, diffraction, phase contrast, and often atomic-scale interpretation. Thus, the logic developed here, *viz*., ground-truth generation, domain adaptation, geometry-aware supervision, and structure-property inference, carries directly into TEM/STEM, but the next section will show that the simulation burden becomes substantially more physics-intensive and the meaning of realism correspondingly more stringent.

6. Synthetic data for TEM and STEM

Among all electron-microscopy modalities, TEM and STEM are among the most developed testbeds for physically grounded synthetic-data workflows because the forward physics is unusually rich but also unusually tractable. Image formation depends on crystal structure, sample thickness, orientation, probe conditions, detector geometry, dynamical scattering, dose, noise statistics, and, in analytical modes, spectroscopic mixing. This combination creates a paradox. On the one hand, manual annotation of experimental data is expensive and often ambiguous at atomic resolution. On the other hand, physically grounded simulations can provide labels that are either impossible or prohibitively expensive to obtain experimentally. As a result, synthetic data in TEM/STEM are no longer limited to simple augmentation. They increasingly function as a bridge between atomistic models, image formation physics, and quantitative inference from real experiments.[69,82,110,111]

Three major trajectories emerge from the relevant literature. First, synthetic data are used to generate realistic high-resolution and atomic-resolution images, often by combining multislice simulation with domain translation or surrogate models. Second, synthetic data support restoration tasks, including sparse acquisition, denoising, and super-resolution, where paired ground truth is otherwise unavailable. Third, synthetic data increasingly enable inverse and predictive analysis: particle sizing, defect recognition, phase separation, composition inference, and atomic-depth estimation. The most important conceptual shift is that TEM/STEM synthetic data are no longer merely ‘fake images for training’. They are becoming physically constrained latent representations of structure, chemistry, and measurement conditions.

6.1. High-resolution and atomic-resolution image synthesis

At high resolution, the central challenge is not simply to generate visually plausible images, but to preserve the mapping between image contrast and the underlying atomic structure, given that the small changes in orientation, thickness, or local composition can alter image appearance dramatically. A clear example is the CycleGAN-based framework reported for supported Pt/CeO_2 NPs, where atomistic models were converted to multislice STEM images and then translated toward the experimental domain to produce ‘experimental-looking’ synthetic data with known structural ground truth.[82] This study is important because it moves beyond standard data augmentation. The synthetic images are not generic variations of real data; they are physically rooted images that preserve NP nuclearity and morphology

while compensating for the domain gap between simulation and experiment.

Figure 13 illustrates this logic. The pipeline starts from atomistic particle models, proceeds through image simulation, and then uses CycleGAN translation to produce synthetic experimental images that can train a downstream size estimator. This type of workflow is especially valuable in TEM/STEM because direct labels such as particle nuclearity, depth, or exact 3D geometry are rarely accessible from experimental images alone. Similar logic underlies the recent multislice surrogate model for NP size estimation, where a 3D convolutional neural network was trained to emulate expensive multislice calculations and then used to create a large synthetic dataset for robust estimator training.[69] The key contribution here is computational scalability: once the surrogate reproduces multislice contrast with sufficient fidelity, the bottleneck shifts from simulation throughput to model design and uncertainty management.

A related direction uses generative translation not only to bridge domains but also to democratize data volume at the million-atom scale. The recent CycleGAN-based strategy for realistic STEM-image generation targets large-scale atom-by-atom analysis by constructing synthetic labeled STEM data that better match experimental appearance while retaining atomic labels from theory.[86] In parallel, rotationally invariant semi-supervised representation learning has been proposed to assemble NP libraries under strong orientational disorder, a regime common in electron-microscopy image streams.[112] These studies emphasize that high-resolution synthesis must solve two problems at once: realism and invariance. Realism is needed for transfer to experiment; invariance is needed because NPs, defects, and local motifs appear under broad distributions of pose, translation, and background.

Physics-guided generative inversion has also been demonstrated for HRTEM exit-wave recovery. Wave-Reconstruction GAN (WRGAN) was trained solely on simulated HRTEM data to predict wave-function amplitude and phase from a single image, and was then applied directly to experimental $Nb_8W_9O_{47}$ and single-wall carbon-nanotube images, recovering defect-sensitive structural information and enabling near-atomic-resolution 3D reconstruction.[113] The authors also delimit the present applicability to restricted imaging conditions and a constrained phase representation, underscoring the need to keep learned inverse outputs within their physical validity range.

Synthetic data have also become central for defect-centric atomic imaging. A particularly instructive example is the recent work on crystalline defects in TEM, which addresses the 'never enough training data' problem by explicitly synthesizing defect images together with realistic backgrounds, noise, and domain variability.[114] Rather than relying on a small set of manually labeled experimental defects, the workflow generates structurally controlled examples that span dislocations, contrast conditions, and realistic background textures. This is a crucial advance for atomic-resolution learning because defects are both physically important and statistically rare. In a similar spirit, physics-based deep data analysis of atomically resolved images has shown that local Fourier descriptors and blind unmixing can identify phases, interfaces, and order-parameter fields directly from real-space atomic images.[115] Although this work is not framed primarily as synthetic-data generation, it is conceptually aligned with the synthetic-data agenda: robust atomic interpretation depends on embeddings and descriptors that respect the underlying image-formation physics.

The newest inverse formulations push this idea further by using synthetic data to recover latent 3D information from 2D projections. Atomic depth estimation from noisy TEM data has recently been formulated as a deep semantic-segmentation problem trained on simulated images corrupted with synthetic noise.[116] This is a noteworthy development because it reframes image synthesis as a route to inaccessible labels rather than merely to larger training sets. The label is not a class or a mask, but a physically meaningful hidden variable. This direction is

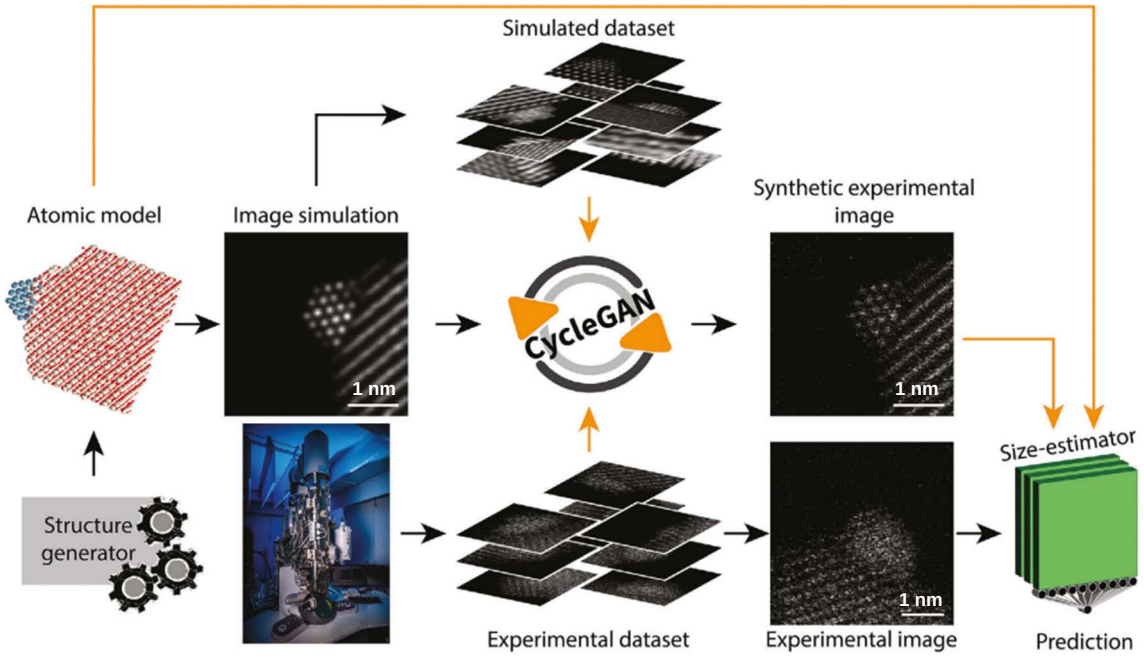


Figure 13. Synthetic-data workflow for catalyst NP analysis in STEM, linking atomistic models, image simulation, CycleGAN domain translation, and downstream size estimation. The figure is reproduced from Eliasson *et al.*[82] according to Open Access Creative Commons Attribution-NonCommercial License.

likely to become increasingly important as TEM/STEM moves from descriptive imaging toward quantitative structural inference.

Large-scale synthetic 4D-STEM training is also enabling near-live ptychographic reconstruction. Ptychography with Robust Artificial Neural Networks Optimized for Sub-Angstrom Accuracy and Ultrafast Reconstruction (PtyRANNOSAUR) trains convolutional autoencoders on 64 million simulated input-target pairs derived from approximately 16 000 crystal structures and multislice abTEM calculations; on experimental data it reconstructs atomic-resolution phase images in seconds, including separations below 0.5 Å, while accounting for variations in thickness, defocus, partial coherence, multiple scattering, and scan-position error.[117] This example shows how physics-based synthetic corpora can shift an expensive inverse problem from iterative optimization toward rapid inference.

6.2. Synthetic data for denoising, restoration, and sparse acquisition

TEM and STEM operate under a persistent dose-efficiency problem. Beam-sensitive materials, dynamic experiments, and spectroscopic mapping all require strategies that extract more information from fewer electrons. Synthetic data are exceptionally powerful in this regime because they provide paired corrupted/clean examples, controlled sampling masks, and known spectral or structural ground truth. One of the clearest cases is sparse scanning transmission electron microscopy with electron energy-loss spectroscopy (STEM-EELS) acquisition, where partially sampled spectrum-images were reconstructed using models that exploit transform sparsity and spatial-spectral structure.[118] The study shows that synthetic undersampling is not merely a benchmark convenience; it becomes a design tool for comparing reconstruction families and identifying which priors remain stable at atomic resolution. The important point for the present review is that synthetic data enable fair benchmarking under known acquisition conditions. In real low-dose experiments, the true unobserved signal is inaccessible, so method development would otherwise rely on visually persuasive but weakly validated outputs. Synthetic sparse acquisition avoids this trap.

The same motivation underlies compressed-sensing STEM on non-rectangular scans, where unconventional scan trajectories were investigated to increase imaging flexibility under dose constraints.[60] Here, synthetic or model-based sampling strategies are inseparable from instrument-aware restoration, because scan geometry itself becomes part of the data-generation process. This line of work broadens the meaning of synthetic data in TEM/STEM: not only the image, but also the acquisition pattern can be simulated and optimized.

Denoising has likewise moved toward physically informed or adaptively tuned learning. Frequency-domain denoising for atomic-resolution electron microscopy has been introduced as a deep-learning strategy operating in reciprocal space, explicitly aiming to suppress noise while preserving lattice information and high-frequency structural detail (Fig. 14).[119] In parallel, adaptive denoising *via* GainTuning shows how a pre-trained CNN can be selectively adjusted to a specific noisy image by modifying only low-dimensional gain parameters, improving robustness under distribution shift.[120] Although GainTuning is not exclusive to electron microscopy, its logic is highly relevant to TEM/STEM, where instrument conditions, detector response, and specimen class often differ from the distributions used for pre-training. The broader lesson is that synthetic data do not

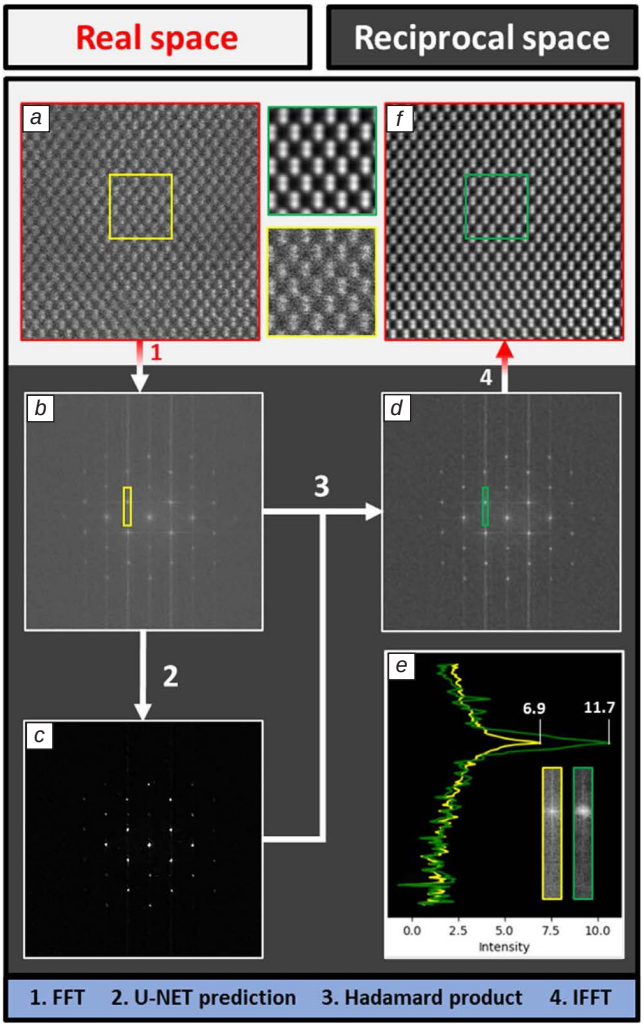


Figure 14. Pipeline of frequency-domain denoising using a U-Net trained on simulated FFT patterns: raw real-space image and its original FFT (*a*, *b*); U-Net-predicted frequency mask and the resulting enhanced FFT (*c*, *d*); peak intensity enhancement (*e*); denoised real-space image after inverse FFT (*f*). The figure is reproduced from Pinto-Huguet *et al.*[119] according to Open Access Creative Commons Attribution License.

eliminate the need for adaptation; rather, they define an initial prior that can later be tuned to the local experiment.

Super-resolution enters the same methodological space. PSSR2, while designed as a general microscopy package, is highly relevant here because it explicitly improves semi-synthetic data generation through controlled synthetic degradation and demonstrates performance on paired electron-microscopy images.[121] Its significance is practical as much as conceptual. The synthetic-data ecosystem around TEM/STEM is maturing from individual proof-of-principle studies toward reusable software infrastructure. This transition matters because restoration methods become scientifically useful only when they can be trained, audited, and transferred by non-specialist users.

A useful boundary case is Noise2Void denoising of atomic-resolution STEM, which avoids simulated clean/noisy training pairs by learning self-supervised from noisy experimental frames. Simulated data instead provide quantitative ground truth for benchmarking; experimentally, the workflow improved Au-adatom visibility and enabled substantially more features to be identified while remaining compatible with real-time-rate acquisition.[122] Thus, synthetic data can be scientifically

important as a validation reference even when they are not the training substrate.

A related strategy has been developed for noisy 4D-STEM orientation mapping. Kamijo *et al*.[123] combined self-supervised denoising that exploits neighboring diffraction patterns with rotation-equivariant unsupervised crystallographic-component analysis, using synthetic 4D-STEM datasets across controlled Poisson-noise levels for quantitative validation before application to experiment. The experimental analysis agreed with prior expert interpretation and additionally recovered minor crystallographic components that had been overlooked.

Taken together, the restoration literature shows that synthetic data serve three different but connected roles. They provide paired supervision for denoising and super-resolution, controlled masks for sparse acquisition, and formal testbeds for comparing priors and reconstruction algorithms. In TEM/STEM, this is especially powerful because image degradation is not a generic nuisance; it is entangled with the physics of dose, scan strategy, and spectroscopy. Therefore, good synthetic restoration data must encode not only visual realism but also acquisition realism.

6.3. Synthetic data for segmentation, object detection, and metrology

Segmentation remains one of the most persistent bottlenecks in electron-microscopy analysis. Manual tracing is slow, subjective, and poorly scalable, especially for NPs, overlapping clusters, and dense microstructural scenes. Synthetic data offer a direct route around this bottleneck by providing pixel-level labels, instance masks, and object counts. Importantly, the literature spans both classical generative segmentation and modern deep learning, showing that the value of synthetic data predates current neural architectures.

Earlier work already exploited explicit generative structure. Graph-cut segmentation with estimated symmetric 3D shape priors demonstrated how segmentation can be stabilized by injecting morphological expectations into nanoscale image volumes.[124] Template Rotation Expectation Maximization (TREM) similarly treated clustered-object segmentation as a generative-model problem, relevant to electron-microscopy scenes containing multiple nearly identical overlapping particles.[125] These approaches are important historically because they reveal a principle that remains valid in modern deep learning: synthetic data are most useful when they encode the geometry of the target objects rather than merely their appearance.

Recent deep-learning studies make this principle more explicit. Synthetic-data generation combined with Mask-RCNN has been proposed for TEM particle segmentation, motivated by the lack of large labeled microscopy datasets.[85] A related study on NP topological structures showed that rendering-based synthetic data, even when mixed with only a limited fraction of authentic data, can substantially improve U-Net segmentation performance.[101] Although the exact specimen classes differ across these works, both demonstrate the same pattern: carefully designed synthetic morphology can compensate, at least partially, for the scarcity of expert annotations.

For heavily overlapping soot agglomerates in TEM, the recent SnakesAgent framework paired a high-fidelity synthetic-data generator with deep-reinforcement-learning active contours. Synthetic scenes were assembled from randomized particle instances and chain-like aggregation patterns with exact overlap-aware instance masks, and the method achieved an IoU of 0.77 and a pixel-level F1 score of 0.87 on real soot data.[126] This is a useful example of synthetic labels being designed around intrinsically ambiguous overlap rather than around isolated objects.

Metrology-oriented segmentation is especially important for catalysis and NP science. Synthetic data and edge-detection strategies have been used to derive particle size distributions from TEM more reproducibly.[127] This direction is significant because the endpoint is not the mask itself, but a quantitative descriptor such as diameter, count, or shape factor. The model is therefore judged not only by segmentation overlap but also by downstream measurement fidelity.

An especially influential development is nNPipe, a neural-network pipeline for morphologically diverse catalyst systems trained exclusively on multislice simulations.[110] The workflow combines object detection and segmentation to analyze large TEM/STEM images of catalysts with very different particle shapes and contrast conditions. Its importance lies in showing that simulated data can support analysis across heterogeneous catalyst morphologies, not merely within a narrow benchmark class. This broadens the scope of synthetic data from single-system optimization toward reusable microscopy pipelines.

Related studies in adjacent microscopy domains reinforce the same lesson. A deep nanowire segmentation model trained with synthetic augmentation demonstrated strong transfer even across imaging modalities, suggesting that geometry-aware synthetic priors can sometimes outweigh exact modality matching.[128] Physics-based synthetic segmentation in catalysis microscopy further showed that introducing physically judged constraints reduces false positives relative to random synthetic generation.[67] Thus, the modeling of cracks formed on the surface of the cobalt oxide catalyst during the isopropanol oxidation reaction was based on the microstructure of the initial catalyst, as determined experimentally by electron microscopy. Although this latter work is centered on environmental SEM rather than TEM/STEM, it is conceptually valuable for the present section because it highlights that synthetic data quality depends strongly on whether the latent process being simulated is physically credible.

Overall, the segmentation literature reveals a hierarchy of synthetic-data sophistication. At the simplest level, one renders particles and masks. At the next level, one encodes realistic variability in overlap, background, orientation, and imaging artifacts. At the most mature level, one synthesizes not just appearance but the underlying physical process that gives rise to the morphology being segmented. TEM/STEM analysis is clearly moving toward this third level.

6.4. Synthetic data for specific tasks in catalysis and functional materials analysis

The deepest impact of synthetic data in TEM/STEM is arguably in catalysis and functional materials, where images are used not simply to classify objects but to infer structure-property relationships. In this context, synthetic data provide a route from atomistic or compositional hypotheses to measurable image features and back again.

Catalyst NPs are a particularly strong example. As discussed above, the CycleGAN-assisted framework for Pt/CeO_2 NPs demonstrates how simulated and translated STEM images can be used to estimate NP nuclearity from experimental observations.[82] The later multislice-surrogate study extends this idea by drastically accelerating image generation while preserving sufficient fidelity for particle-size estimation.[69] Together with nNPipe [110] and synthetic-data-based TEM particle metrology,[127] these works demonstrate how the specific tasks in

catalyst microscopy can transform the synthetic data generation and usage strategies: synthetic data are evolving from a support tool for segmentation into a quantitative engine for extracting particle populations, size distributions, and possibly structure-sensitive descriptors that are more relevant to catalysis than simple projected diameters.

The analytical microscopy literature extends this synthetic-data logic to composition and phase separation. For mixed materials, CNN-based composition estimation has been demonstrated using synthetic images generated from pure-material images rather than experimentally assembled mixtures.[129] This is conceptually elegant because it shows that combinatorial state spaces can sometimes be learned without experimentally collecting all combinations. In scanning transmission electron microscopy-energy-dispersive X-ray spectroscopy (STEM-EDS), physics-guided non-negative matrix factorization (NMF) has been developed to separate phases and quantify composition while respecting Poisson statistics, fluorescence effects, simplex constraints, and sparsity priors; importantly, the method is extensively validated on simulated data before application to experiments.[130] A related pan-sharpening NMF approach likewise uses synthetic noisy datasets to benchmark chemical unmixing and improve quantitative analysis of supported catalytic NPs and other complex nanomaterials.[131] These works collectively show that synthetic data are crucial not only for image interpretation but also for reliable decomposition of hyperspectral signals.

Functional materials at atomic resolution provide even more demanding test cases. In high-entropy oxides, a fully convolutional network has been trained on simulated high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) images to predict column heights and elemental distributions from single experimental images.[111] Figure 15 highlights this progression from atomistic model generation to simulated STEM images and finally to element-resolved prediction. The significance of this work is twofold. First, the labels are again latent physical variables, not directly observable ground truths from routine experiments. Second, the model must bridge a high-dimensional configuration space involving chemistry, structure, projection effects, and imaging conditions. This is exactly the type of problem for which synthetic data are indispensable.

More broadly, atomically resolved image analytics are becoming increasingly physics-aware. Deep-data analysis of phases and interfaces from real-space atomic images[115] and synthetic-data-driven learning of defect signatures[114] both show that interpretation at the atomic scale is moving toward hybrid workflows where simulation, feature learning, and physical constraints are tightly coupled. The recent atomic-depth estimation study[116] extends this trajectory to hidden 3D structure, and the CycleGAN-based STEM synthesis study aimed at million-atom-scale analysis suggests how such pipelines may eventually operate at very large throughput.[86]

A notable commonality across catalysis and functional-materials applications is that the most successful synthetic-data pipelines are not purely data-driven. They are hybrid constructions that begin with explicit physical models, then use machine learning to accelerate, translate, or invert them. This is visible in multislice surrogates for NPs,[69] in simulation-trained detector/segmenter pipelines for catalysts,[110] in physically constrained synthetic crack generation for catalytic microscopy,[67] and in physics-guided factorization of STEM-EDX(S) data.[130] The field is therefore converging on a practical principle: the closer the synthetic pipeline is to the actual material physics and microscope physics, the more useful it becomes for quantitative interpretation.

The TEM/STEM literature makes a broader point about synthetic data in electron microscopy. At these length scales, success is determined less by raw visual realism than by whether the synthetic pipeline preserves the causal chain linking structure, imaging conditions, and measurable contrast. This is why the strongest studies are those that combine atomistic or compositional models with explicit forward simulation, then use machine learning either to bridge the simulation-to-experiment gap, to accelerate otherwise intractable calculations, or to invert the image back to hidden structural variables.[69,82,111] Synthetic data are thus serving three intertwined functions in TEM/STEM: they expand training corpora, provide labels that experiments cannot directly supply, and create controlled environments for validating quantitative inference.

This perspective also clarifies the current limitations. Domain shift remains a central problem: highly realistic simulations can still fail when background statistics, contamination, detector artifacts, or sample preparation effects are not represented. Some studies address this through CycleGAN domain

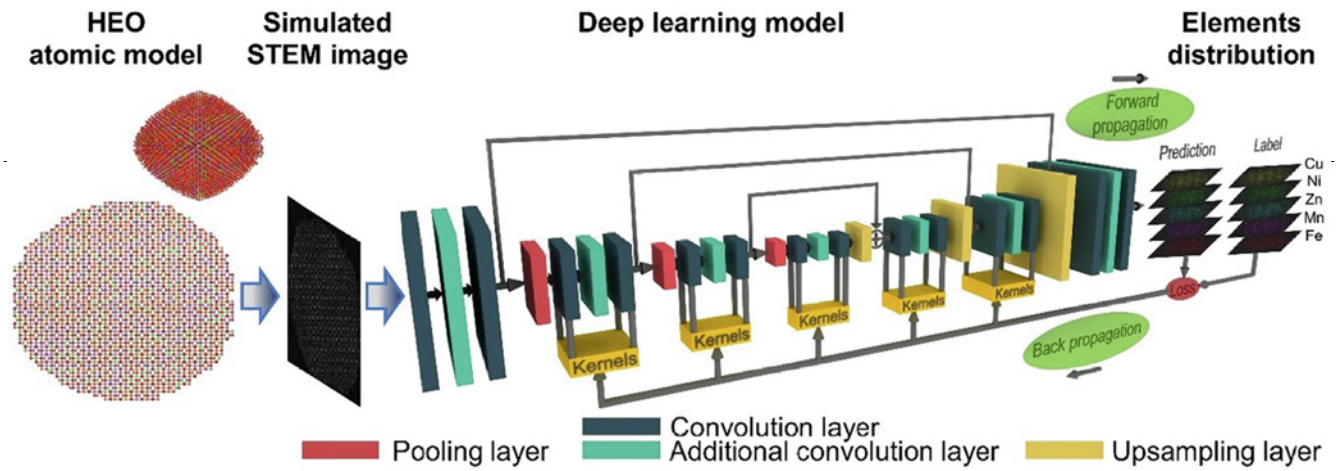


Figure 15. Physics-to-prediction pipeline for high-entropy oxides, where atomistic models generate simulated HAADF-STEM images that train a network to infer element-resolved distributions from experimental data. HEO is high entropy oxide. The figure is reproduced from Borshon *et al.*[111] according to Open Access Creative Commons License CC BY 4.0.

adaptation,[82,86] others through adaptive denoising,[120] and others through physically constrained synthetic generation.[67] Yet the fundamental lesson is that synthetic data do not replace careful experimental understanding; they amplify it when the relevant priors are encoded correctly.

This section naturally leads to the next one on cryo-EM and cryo-ET. TEM/STEM already show how synthetic data can link forward physics to inverse analysis in high-resolution imaging. Cryogenic modalities inherit the same need for physically faithful simulation, but under even harsher conditions: extreme low dose, stronger noise, conformational heterogeneity, projection ambiguity, and a much more explicit 3D reconstruction problem. In that sense, TEM/STEM provide the methodological bridge. They demonstrate that synthetic data become most powerful when they are not generic augmentations, but structured computational surrogates of the microscope-sample system itself.

7. Synthetic data for cryo-electron microscopy and cryo-electron tomography

Cryo-electron microscopy and cryo-electron tomography occupy a distinctive position within the synthetic-data landscape of electron microscopy because both the signal formation model and the biological meaning of the data are unusually difficult. Unlike many conventional image-analysis problems, cryo-EM begins from extremely low-dose images acquired at very low SNR, with strong contrast-transfer-function effects, beam-induced motion, preferred orientation, particle overlap, ice contamination, heterogeneity of conformation and composition, and imperfect knowledge of viewing parameters. In cryo-ET these difficulties are compounded by the missing wedge, depth-dependent defocus, tilt-series dose accumulation, and the need to reconstruct three-dimensional cellular or macromolecular context from anisotropic measurements. For this reason, synthetic data in cryo-EM are not merely a convenience for data augmentation; they are an essential instrument for benchmarking inverse algorithms, supervising denoisers and pickers, probing identifiability, stress-testing reconstruction pipelines, and quantifying how assumptions about heterogeneity shape the final map.[37,132–135]

The literature covered in this section shows a clear evolution in the role of synthetic cryo-EM data. Early studies used controlled simulations primarily to validate geometry, orientation assignment, and model-based reconstruction under known ground truth.[136–139] Later work increasingly used synthetic data to train task-specific discriminative models for particle picking, denoising, and segmentation.[140–143] The most recent generation of studies goes further still: synthetic data are now used to benchmark heterogeneous reconstruction methods, to amortize conformational inference, to drive physics-informed generative models, and to construct full end-to-end evaluation ecosystems for cryo-EM and cryo-ET pipelines.[37,40,48,65,68,144–146]

7.1. Why cryo-EM is a distinctive case

Cryo-EM is a distinctive case because the mismatch between what is measured and what must be inferred is exceptionally severe. The experimental image is a noisy two-dimensional projection or projection-like contrast-modulated view of an unknown three-dimensional object, often embedded within a heterogeneous ensemble. In single-particle analysis, each particle image is unique in orientation and often unique in conformation. In tomography, each tilt-series image contains depth variation, cumulative radiation damage, and projection geometry that make the forward model explicitly three-dimensional. Consequently, the value of synthetic data in cryo-EM is not limited to 'looking realistic'; the synthetic pipeline must also preserve the latent variables that matter scientifically, such as orientation, pose, defocus, CTF, particle occupancy, conformational coordinates, and the relation between 2D evidence and 3D structural truth.[135,144,147,148]

This is why cryo-EM synthetic data repeatedly appear in two complementary forms. One form is mechanistic and physics-driven, where synthetic images are generated from known structures, tilt geometry, and microscope optics to test inverse algorithms under controlled perturbations. The other is statistical or generative, where simulated or semi-simulated data are used to approximate the appearance distribution of real micrographs or tomograms so that neural networks can be trained robustly. The field increasingly recognizes that neither form alone is sufficient. Mechanistic realism without appearance realism may mislead downstream learning, whereas appearance realism without traceable latent truth may be inadequate for evaluating whether an algorithm recovers the right structure for the right reason. The most influential recent studies therefore combine controllable ground truth, forward-model consistency, and some strategy for narrowing the domain gap to experimental data.[37,48,65,68,140]

7.2. Synthetic data for particle picking and detection

Particle picking was one of the earliest cryo-EM tasks to benefit dramatically from synthetic data because labeled particle coordinates are laborious to produce and often subjective in low-contrast micrographs. A representative milestone is PARticle SEgmentation Detector (PARSED), where synthetic particle projections generated from known macromolecular structures were combined with noise to create supervised training data for a fully convolutional particle-segmentation network.[140] The conceptual importance of this work is not only that synthetic data enabled automated picking, but that it reframed particle picking as a segmentation problem with pixel-level ground truth rather than as a patch-classification problem driven by imperfect hand labels. The logic is following: known structures are projected, corrupted to different SNR levels, assembled into simulated micrographs, and then used to learn detection maps from which particle centers can be extracted.[140]

Subsequent developments broadened the detection problem from simple particle centers to more structured foreground-background separation and mask generation. CRyo-EM image Segmentation and Processing (CRISP) extends this direction by combining modular segmentation strategies with conditional random fields, emphasizing that cryo-EM image processing benefits from synthetic or semi-synthetic labels that regularize boundary definition and suppress background ambiguities.[143] In parallel, toolkit-oriented efforts such as PERC underscore that reliable curation, simulation, and annotation management are becoming part of the same ecosystem: synthetic data are useful only if they remain traceable, reproducible, and interoperable with downstream machine-learning pipelines.[48]

A central lesson from the particle-picking literature is that realism must be task-aware. For picking, the crucial requirement is not perfect biochemical fidelity of every density fluctuation, but preservation of the detection statistics that govern false positives and false negatives: background texture, ice-like contamination, particle crowding, size variability, CTF blur, and

the relation between SNR and apparent particle compactness. Synthetic datasets that vary these factors systematically allow one to map the operating envelope of pickers in a way that real data rarely permit. This ability to stress-test models under controlled conditions has become especially important as cryo-EM moves toward more heterogeneous, flexible, and compositionally mixed samples.[48,140,143]

7.3. Synthetic data for alignment, registration, and orientation estimation

Orientation assignment is the classical inverse problem at the heart of single-particle cryo-EM, and synthetic data have long been indispensable for validating new registration and orientation-estimation strategies. Earlier algorithmic work established the importance of simulation for testing common-lines methods, model-based reconstruction, and symmetry-aware estimation under known viewing directions.[136–138,147] The study of self-adapting common lines demonstrated that even subtle refinements in spectral treatment can improve high-resolution orientation refinement, but such claims are convincing only because they can be benchmarked against synthetic data with known poses.[147]

Registration-oriented work based on Fourier-Bessel and fast Laguerre projection methods continues this line of research, targeting efficient alignment of large projection sets while retaining robustness under cryo-EM noise.[149,150] These studies highlight a recurring advantage of synthetic data in cryo-EM: they permit separate examination of translational, rotational, and appearance-related errors, rather than collapsing all failure modes into a single empirical success metric. Likewise, end-to-end orientation estimation from 2D cryo-EM images shows how synthetic data enable direct supervision of pose prediction, making it possible to compare learned estimators with more classical optimization-based pipelines (Fig. 16).[148]

At a deeper mathematical level, inverse formulations involving the spherical Radon and cosine transforms provide a reminder that synthetic data are essential not only for training neural models but also for validating stable inversion algorithms in continuous geometric settings.[139] In cryo-EM, such mathematically grounded synthetic benchmarks remain important because they separate questions of identifiability from questions of implementation. Synthetic experiments can reveal whether an apparent reconstruction limit arises from noise, from sampling geometry, or from the intrinsic ill-posedness of the transform itself.

Overall, the alignment literature shows that cryo-EM synthetic data are uniquely valuable when the benchmark is defined in latent-variable space. For particle picking, image realism is the dominant requirement. For alignment and orientation estimation, by contrast, exact pose labels and controlled geometric perturbations are the critical assets. This difference explains why apparently simpler simulations may still be highly informative for one task while being insufficient for another.[139,148–150]

7.4. Synthetic data for denoising and low-SNR image restoration

Denoising is especially important in cryo-EM because low-dose imaging is not a nuisance but a physical necessity. Synthetic data therefore play a double role: they provide paired clean/noisy data for supervised training, and they enable principled tests of whether denoising preserves structural information needed for downstream tasks. SaID is emblematic of this direction, using simulation-aware training to improve denoising of cryo-EM micrographs by exploiting synthetic clean targets linked to the cryo-EM forward model.[142] The practical motivation is straightforward: purely supervised denoisers trained on limited or stylistically narrow synthetic data often fail to generalize, yet purely self-supervised approaches may not fully exploit known structure formation. Simulation-aware strategies attempt to reconcile these regimes.

The denoising problem also intersects with representation learning for classification and clustering. Cascaded denoising autoencoder shows that cascades of denoising autoencoders (CDAE) can improve feature quality for clustering single-particle cryo-EM images, illustrating that restoration should not

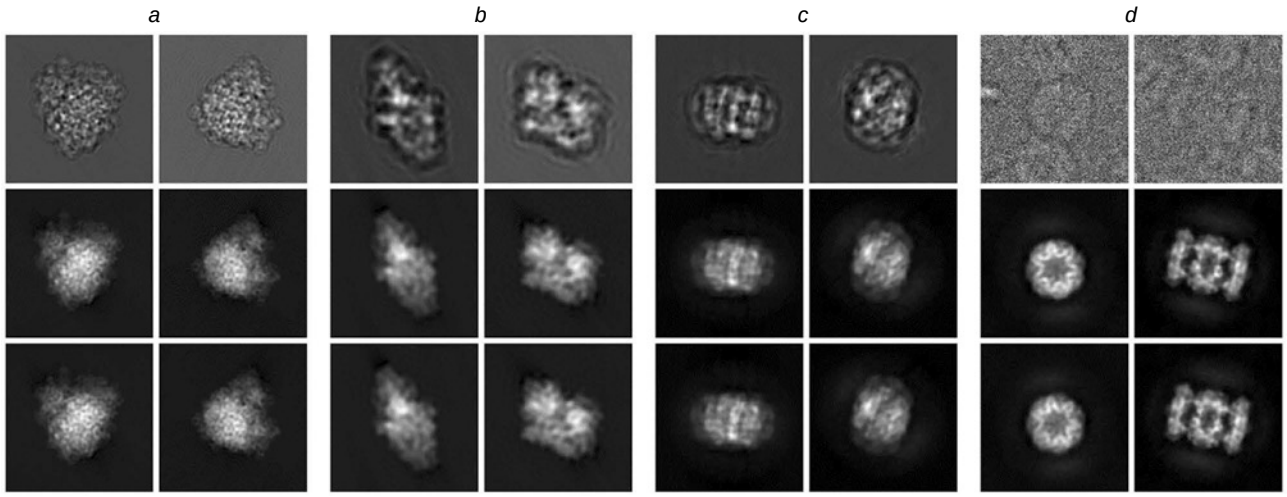


Figure 16. Orientation estimation from 2D cryo-EM images using a neural network trained on synthetic data. Test-set results for four datasets: EMPIAR-10028-simu (asymmetric 80S ribosome) (*a*), EMPIAR-10061-simu (D_2-symmetric β-galactosidase) (*b*), EMPIAR-10025-simu (D_7-symmetric 20S proteasome) (*c*), and EMPIAR-10025-real (*D*7-symmetric 20S proteasome) (*d*). In each panel, the top row shows network inputs, the middle row shows simplified projections along ground-truth orientations, and the bottom row shows simplified projections along network-predicted orientations. The figure is reproduced from Lian *et al.*[148] according to Open Access Creative Commons Attribution License.

be evaluated solely by visual smoothness but by the extent to which it stabilizes downstream structural grouping.[141] In cryo-ET, restoration is inseparable from corrections for the three-dimensional optical transfer. NovaCTF demonstrates that accurate 3D-CTF correction improves subtomogram averaging resolution, again relying on synthetic reasoning about how depth-dependent defocus propagates through the tomographic workflow.[135]

What makes denoising in cryo-EM particularly delicate is that the desired signal often consists of weak, orientation-dependent, and heterogeneity-sensitive features that can be erased or hallucinated by aggressive restoration. Synthetic data are therefore most valuable when they permit task-coupled evaluation. A denoiser should not merely increase apparent contrast; it should improve picking, alignment, classification, or reconstruction without collapsing conformational variability. This requirement is one reason why modern cryo-EM benchmarking increasingly couples denoising with downstream tasks rather than treating it as an isolated image-enhancement problem.[135,141,142]

7.5. Synthetic data for *ab initio* reconstruction and heterogeneous reconstruction

Ab initio reconstruction is one of the areas where synthetic data have had the deepest conceptual impact. Early orthogonal-tilt reconstruction studies addressed a fundamental challenge in generating reliable initial models for asymmetric particles while avoiding the missing-cone limitations of random conical tilt approaches.[132] The subsequent biological validation study confirmed that synthetic and semi-synthetic tests can reveal not only whether an algorithm converges, but whether the reconstructed class volumes remain faithful enough to serve as practical initial references.[133] Related model-based approaches for virus particles with incompatible symmetries used synthetic data to test reconstruction when symmetry is only partially valid, thereby exposing a regime where idealized assumptions about global symmetry break down.[136,137]

Likelihood-based subtomogram and random-conical-tilt averaging introduced another important principle: synthetic data enable validation of probabilistic optimization in the presence of missing information, uncertain alignment, and heterogeneous signal quality.[134] More recent work extends these ideas from discrete class averaging toward explicit continuous or high-dimensional heterogeneity. CryoDRGN-AI is particularly important here because it uses neural *ab initio* reconstruction to couple pose estimation and heterogeneous structure inference in challenging cryo-EM and cryo-ET datasets, reducing dependence on pre-existing initial models and making the reconstruction pipeline more robust to strong structural variability.[145]

The move from discrete heterogeneity to conformational continua is also evident in simulation-based inference and shape-based recovery frameworks. Amortized template matching uses simulation-based inference to recover molecular conformations directly from cryo-EM images, showing that synthetic training distributions can be used not only to benchmark reconstruction but to infer conformational state itself (Fig. 17).[144] Geometric shape matching approaches similarly treat cryo-EM as an indirect shape-recovery problem, emphasizing the recovery of protein backbone conformation from noisy projection data.[151] CryoChains extends the reconstruction discussion to semi-flexible assemblies with large shape deformations, indicating that synthetic priors over biologically plausible motion are becoming central to next-generation heterogeneous reconstruction.[146]

The main insight from this section is that synthetic data have changed from being a passive validation resource to being an active component of the reconstruction model. They now define priors over conformations, provide training targets for amortized inference, and determine which aspects of heterogeneity are even visible to an algorithm. This creates both opportunity and risk: sophisticated synthetic generators can expose failure modes of conventional reconstruction, but they can also bias a pipeline toward the family of heterogeneity that the simulator assumes. Benchmarking frameworks are therefore essential.[37,144–146]

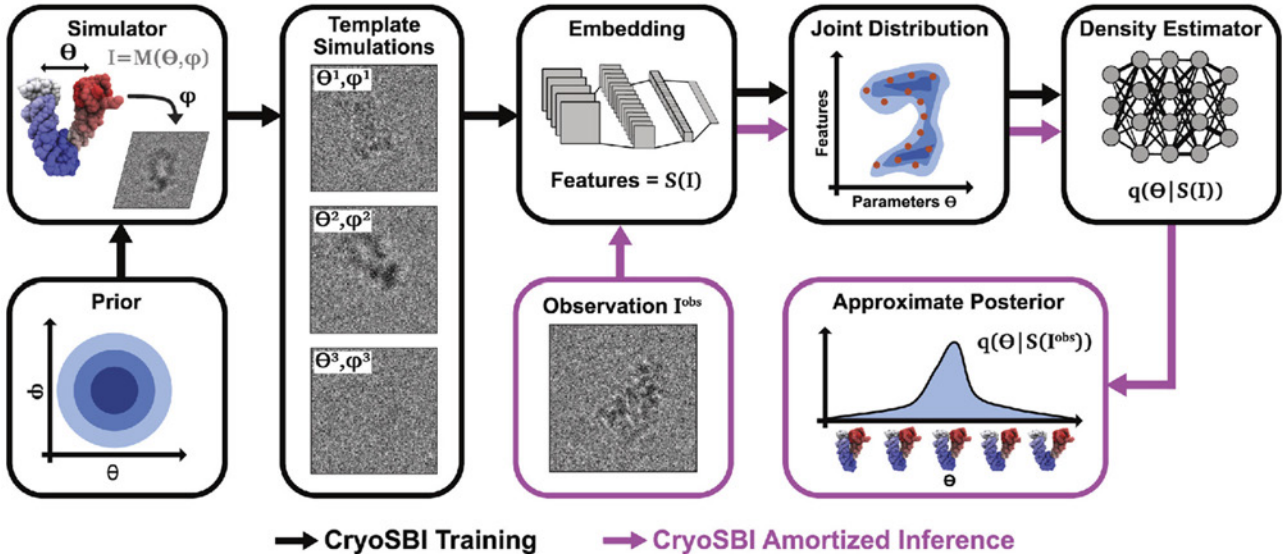


Figure 17. CryoSBI uses simulation-based inference to learn an amortized posterior over molecular conformations from synthetic images. Once trained, the model can rapidly infer conformations from experimental cryo-EM particles, showing that synthetic data serve not merely for validation but as an integral part of the inference pipeline. The figure is reproduced from Dingeldein *et al*.[144] according to Open Access CC BY-NC-ND license.

7.6. Generative and physics-informed models in cryo-EM

Generative modeling is emerging as one of the most consequential new directions in cryo-EM synthetic data because it promises to narrow the gap between mechanistic simulators and the visual statistics of real micrographs. CryoGEM is representative of physics-informed generative cryo-EM, explicitly combining cryo-EM image-formation knowledge with generative learning to synthesize annotated data of higher realism than naive projection-plus-noise pipelines.[65] CryoCCD pushes this trend further by introducing conditional cycle-consistent diffusion together with biophysical modeling of heterogeneity and cryo-EM noise, directly targeting the deficiencies of earlier synthetic generators that lacked realistic conformational variation or realistic background statistics.[68]

These models are important for two reasons. First, they make synthetic data more useful for supervision of data-hungry neural methods in detection, restoration, and reconstruction. Second, they expose a major conceptual shift: the field no longer treats realism as a single scalar property. Instead, one may separately seek realism of particle appearance, background noise, CTF behavior, conformational diversity, and metadata consistency. The most successful synthetic frameworks will likely be those that allow these realism axes to be tuned independently, so that one can generate datasets optimized for training, for benchmarking, or for stress-testing under distribution shift.[65,68]

7.7. Limits of realism in macromolecular and conformational heterogeneity

The central unresolved issue in synthetic cryo-EM is not whether realistic images can be generated, but what kind of realism is required for a given scientific claim. If the goal is particle picking, appearance realism may dominate. If the goal is orientation recovery, exact latent-variable labels matter more than textural fidelity. If the goal is heterogeneous reconstruction, however, both are insufficient unless the synthetic data also capture plausible conformational manifolds, occupancy distributions, compositional mixtures, and the coupling between structure and imaging parameters. This is why recent methods increasingly incorporate molecular dynamics, simulation-based inference, or explicit generative priors over conformational space.[37,68,144,146]

There are also fundamental limits. Biological macromolecules can exhibit hierarchies of heterogeneity spanning rigid-body motions, loop flexibility, ligand occupancy, assembly-state variation, and biochemical damage. No finite synthetic benchmark can capture this full diversity. Moreover, a simulator may reproduce average-looking images while still imposing an incorrect structural prior, leading a reconstruction method to perform well on the benchmark yet fail on real data. This is especially important when synthetic data are used to train generative or amortized models whose outputs may appear plausible even when the inferred conformational landscape is distorted. Consequently, the most trustworthy role of synthetic data in cryo-EM is not to replace experimental validation, but to provide controlled counterfactuals: known truths against which identifiability, uncertainty, and failure modes can be probed systematically.[65,68,144,145,151]

Taken as a whole, these advancements position synthetic data as a unifying medium between physics-based inverse problems and modern machine learning. Synthetic data enable detector training, denoiser supervision, pose estimation, tomography correction, *ab initio* reconstruction, heterogeneity benchmarking, and generative modeling. Yet the strongest message of this section is qualitative rather than merely technical: in cryo-EM, synthetic data are valuable precisely because they make the hidden variables of the experiment explicit. They convert an underdetermined imaging problem into a controlled computational laboratory. Cryo-EM and cryo-ET focus on reconstructing macromolecular structure from sparse and noisy projections, whereas the next domain shifts toward larger-volume electron microscopy settings in which synthetic data must address dense three-dimensional context, anisotropy, sectioning artifacts, and large-scale semantic labeling. In that sense, the logic remains continuous, but the dominant bottleneck moves from latent structural inference to volumetric representation and annotation at scale.

8. Synthetic data for FIB-SEM, volume EM, and 3D reconstruction

Synthetic data play a particularly important role in FIB-SEM, serial-section volume EM, and related 3D reconstruction workflows because the ground truth is intrinsically difficult to obtain after destructive acquisition, while the downstream tasks (segmentation, registration, artifact correction, and volumetric reconstruction) depend on information that is fundamentally three-dimensional. In this area, synthetic data are not used only to enlarge a training set. They are used to encode morphology, sectioning geometry, imaging artifacts, anisotropy, and reconstruction uncertainty in a controlled way, so that algorithms can be stress-tested under conditions that are difficult to isolate experimentally. The field has undergone a clear evolution, moving from synthetic phantoms for validating segmentation and tomographic reconstruction toward digital volumes, generative augmentation, and uncertainty-aware restoration pipelines that begin to approximate realistic large-scale EM experiments.[70,83,90,124,152–154]

8.1. Synthetic data for serial section and volumetric segmentation

One of the earliest and most persistent uses of synthetic data in volume EM has been the validation of 3D segmentation strategies for densely packed ultrastructure. In serial-section and block-face imaging, the computational goal is not merely to label a 2D image, but to preserve topological continuity across slices. This requirement immediately makes synthetic data valuable, because it permits control over object shape, section thickness, overlap, and noise while preserving known 3D identity. Early graph-cut and shape-prior approaches for serial block-face SEM and related nanoscale datasets explicitly relied on synthetic volumes to test how well 3D extraction could be recovered from ambiguous image stacks.[124,154] These studies established an important principle: realistic evaluation in volume EM must account for stack-wise consistency, not only per-slice visual agreement.

For porous and multiphase materials imaged by FIB-SEM, segmentation is complicated further by shine-through, depth-of-field effects, and ambiguous grayscale mappings. Classical mathematical-morphology pipelines already used synthetic FIB-SEM simulations to quantify reconstruction error against known 3D structure, demonstrating that simulated image formation can reveal where thresholding and watershed-based methods fail.[152] Later work replaced handcrafted segmentation with 3D convolutional networks trained on simulated stacks of stochastic

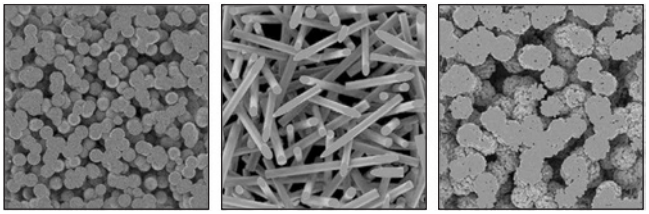


Figure 18. Examples of synthetic FIB-SEM image data generated from stochastic geometry models used to span diverse porous morphologies. The figure is reproduced from Fend *et al*.[153] according to Open Access Creative Commons Attribution-NonCommercial-NoDerivs License.

geometry models, including Boolean models, Cox-Boolean models, random packings, and fiber systems.[153] The central insight of this line of work is that for volumetric segmentation, the crucial realism is not photorealism alone, but the structural coherence of the image obtained; in other words whether the synthetic generator spans the relevant family of local textures, pore networks, and connectivity patterns (Fig. 18).

The same logic has increasingly appeared in biological volume EM, where annotation is expensive and class imbalance is severe. Synthetic brain EM datasets were generated to supply labeled organelles and underrepresented classes such as axon sheaths and postsynaptic densities, showing that mixed real-plus-synthetic training can markedly improve multiclass segmentation when the real dataset is small or morphologically biased.[155–157] At the same time, these studies also showed that not every synthetic addition helps automatically: diffusion-generated mitochondrial images with good visual similarity did not necessarily improve Dice score, reminding us that label-consistent diversity is more important than superficial realism.[158] Even seemingly simple perturbations such as explicit noise injection can alter the transfer value of synthetic microscopy images, again emphasizing that training utility depends on matching the nuisance distribution seen in real data.[159]

8.2. Synthetic datasets for neuronal and biological ultrastructure

In biological volume EM, synthetic data increasingly serve as proxies for neuronal ultrastructure and organelle-rich tissue volumes. Here, the challenge is not only scarcity of labels, but the combinatorial complexity of membranes, vesicles, mitochondria, synaptic densities, and long neurites across very large volumes. The reviewed studies pursue two complementary strategies. The first is geometry-driven synthesis, where biologically inspired primitives are assembled into labeled scenes for downstream segmentation and counting.[155–157] The second is generative augmentation, where learned models produce new image-label pairs consistent with known masks or domain-specific structural cues.[101,146,160]

The emergence of diffusion and ControlNet-style generators is especially significant for FIB-SEM because these datasets are typically grayscale, strongly correlated from slice to slice, and much smaller than natural-image corpora. Fine-tuning generative priors originally trained on natural images can nevertheless yield label-consistent FIB-SEM augmentations and measurable gains in downstream segmentation performance when conditioned properly on masks or structural controls.[160] This is an important shift in the field. Synthetic volume EM is no longer limited to handcrafted phantoms or forward Monte Carlo image formation; it now includes controllable generative models that can preserve semantics while enriching rare morphologies and mitigating low-variance training sets.[146]

A broader message emerges from these biological datasets. Useful synthetic volumes must encode not just isolated organelles but the relational organization of ultrastructure: adjacency, enclosure, membrane continuity, section-to-section evolution, and frequency of rare events. That is why mixed datasets so often outperform purely synthetic ones. The synthetic component is most powerful when it regularizes the representation space, expands rare structural classes, and makes the learning problem better conditioned, while real data anchor the model to laboratory-specific appearance.[101,155,156]

8.3. Synthetic artifacts in FIB-SEM: curtaining, slicing, deformation, and contrast variation

Perhaps nowhere is the value of synthetic data more obvious than in explicit artifact modelling. FIB-SEM data are affected by shine-through, curtaining, charging, blur, contrast drift, local damage, and slicing-induced distortions. These artifacts do not simply degrade image quality; they alter the statistical structure that segmentation and reconstruction algorithms interpret as morphology. Therefore, a synthetic benchmark that omits artifacts may be visually clean yet scientifically misleading. There is an increasing shift toward treating artifacts as first-class components of the generator rather than post hoc nuisance variables.[70,83,152,153]

In porous materials, synthetic forward models based on Monte Carlo electron transport established a route from known 3D microstructure to realistic FIB-SEM contrast, thereby exposing the role of depth ambiguity and shine-through in segmentation failure.[152,153] More recent work has moved beyond pure simulation by coupling physics-derived synthetic stacks with generative adversarial networks. Sardhara *et al*.[70] showed that GAN-based translation can reduce the distribution gap between Monte Carlo FIB tomography and real hierarchical nanoporous gold data while retaining structural supervision, enabling substantial gains in 3D semantic reconstruction accuracy. In parallel, Roldan and Barbosa-Torres[83] explicitly synthesized curtaining artifacts with controllable depth and width by superimposing them onto classically modeled secondary electron and backscattered electron SEM data, making it possible to probe the limits of decurtaining algorithms under known damage conditions (Fig. 19).

Biological serial-section datasets bring a related but distinct artifact problem: deformation between sections. Here, synthetic models are invaluable because they let one vary object size, section thickness, and deformation independently. Registration analysis based on simulated neuronal structures showed that adjacent-section correlation falls as section thickness increases and that registration accuracy degrades with neuronal structure size, clarifying why simple image-noise models are insufficient for serial EM alignment studies.[161] This is a highly relevant lesson for volume EM: many apparent reconstruction errors originate not from segmentation itself, but from the compounded effect of acquisition anisotropy, morphology-dependent section variation, and registration uncertainty.

8.4. Synthetic data for 3D reconstruction, registration, and benchmarking

Beyond segmentation, synthetic data underpin benchmarking for 3D reconstruction algorithms themselves. This includes

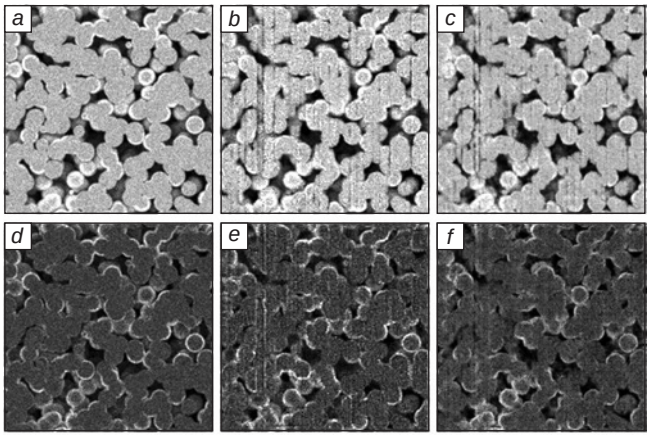


Figure 19. Synthetic FIB-SEM slices illustrating generated curtaining artifacts in BSE and SE images. BSE images: simulated without artifacts (*a*), Pix2Pix-generated with artifacts (*b*), CycleGAN-generated with artifacts (*c*). SE images: simulated without artifacts (*d*), Pix2Pix-generated with artifacts (*e*), CycleGAN-generated with artifacts (*f*). The artifacts are more visible in the BSE images because of their higher contrast. The figure is reproduced from Roldán and Barbosa-Torres [83] according to Open Access CC BY-NC-ND license.

tomographic reconstruction, subtomogram averaging logic, registration, parameter tuning, and evaluation of inverse solvers. Foundational work on electron microscopy tomography used synthetic data to validate model generation and structural averaging procedures, demonstrating early on that reconstruction methods must be tested against controlled missing-information scenarios before being trusted on experimental data.[134,162] Related studies in tomographic reconstruction and regularization also used synthetic data to compare optimization schemes under realistic noise models before applying them to EM measurements.[163]

This benchmark-oriented use of synthetic data persists in modern autonomous tomography. Machine-learning-guided electron tomography reconstruction has been evaluated on simulated datasets to optimize reconstruction hyperparameters and reduce computational cost, illustrating that synthetic data are essential not only for training recognition models but also for calibrating the reconstruction engine itself.[164] Similarly, efficient correction and reconstruction strategies in electron tomography have long depended on simulation to estimate the benefits of algorithmic choices under known imaging geometry.[135]

For volume EM, the benchmarking target is increasingly end-to-end: not one isolated algorithm, but the whole pipeline from anisotropic acquisition to isotropic volume, organelle reconstruction, and downstream measurement. This is precisely the setting in which synthetic or semi-synthetic data are most informative, because they enable paired comparisons between degraded inputs, restored volumes, uncertainty maps, and object-level reconstructions. Thus, synthetic data provide the missing calibration layer between acquisition conditions and the final biological or materials-science interpretation.[90,161,164]

8.5. Toward realistic digital volumes for large-scale EM

The most forward-looking trend in this area is the move from synthetic slices toward realistic digital volumes. Such volumes are expected to preserve morphology across 3D context, reproduce anisotropy and artifacts, and remain useful for both learning and benchmarking at larger spatial scales. A strong example in the present literature is EMDiffuse algorithm, which extends diffusion-based enhancement from denoising and super-resolution to isotropic reconstruction of volume EM data.[90] Its importance for this review is conceptual as much as technical. The method treats anisotropic volume EM not as a static deficit but as a generative inference problem, reconstructing intermediate layers and providing self-assessment of reliability (Fig. 20). In doing so, it points toward a future where synthetic data are no longer external training supplements but embedded components of the reconstruction pipeline.

This shift also redefines what realism means. For large-scale EM, a useful digital volume must represent multiscale continuity, class balance, uncertainty, and acquisition physics simultaneously. Purely handcrafted phantoms often fail to capture this complexity, whereas purely generative models risk hallucinated plausibility. The most promising approaches therefore combine explicit structural priors, physically grounded simulators, and learned domain adaptation or diffusion-based refinement.[70,83,90,153,160]

Across the literature, several design principles recur. First, synthetic datasets are most valuable when they target the bottleneck of the workflow: rare morphologies, unknown ground truth, section misregistration, or acquisition artifacts. Second, 3D context matters more than isolated slice realism. Third, dataset diversity must include nuisance variation as well as object variation. Fourth, the best-performing workflows increasingly use synthetic data not to replace experiments, but to structure and constrain learning in regimes where experiments cannot provide dense truth. Taken together, the field is moving toward digital twins of volume EM pipelines: controllable virtual volumes that support training, calibration, error analysis, and uncertainty estimation at scale.[70,90,152,153]

Once synthetic data are understood as tools for preserving volumetric identity, modelling artifact realism, and benchmarking 3D inference, the discussion can expand from modality-specific case studies to the major application tasks that synthetic data enable across EM as a whole. Classification, detection, segmentation, denoising, registration, reconstruction, metrology, and inverse analysis all appear in the present section as interconnected functions of one broader idea: synthetic data make otherwise unobservable error sources experimentally tractable. Next section can therefore build directly on this foundation by reorganizing the discussion around those cross-cutting tasks rather than around instrument class.

9. Major application tasks enabled by synthetic data

Synthetic data do not enable one single application in electron microscopy; rather, they reshape the entire task landscape. Once realistic or task-faithful synthetic observations become available, problems that were formerly limited by sparse labels, weak contrast, destructive preparation, low signal-to-noise ratio, or the lack of ground truth can be reformulated as trainable, testable, and benchmarkable workflows. Recent advances show that the same general idea recurs across very different settings: synthetic data can define class boundaries for classification, provide dense annotations for detection and segmentation, create paired clean/noisy observations for restoration, supply controlled deformations for registration, furnish volumes or projections for reconstruction, and connect image space to physical or functional variables in inverse problems and structure/property inference.[165–169]

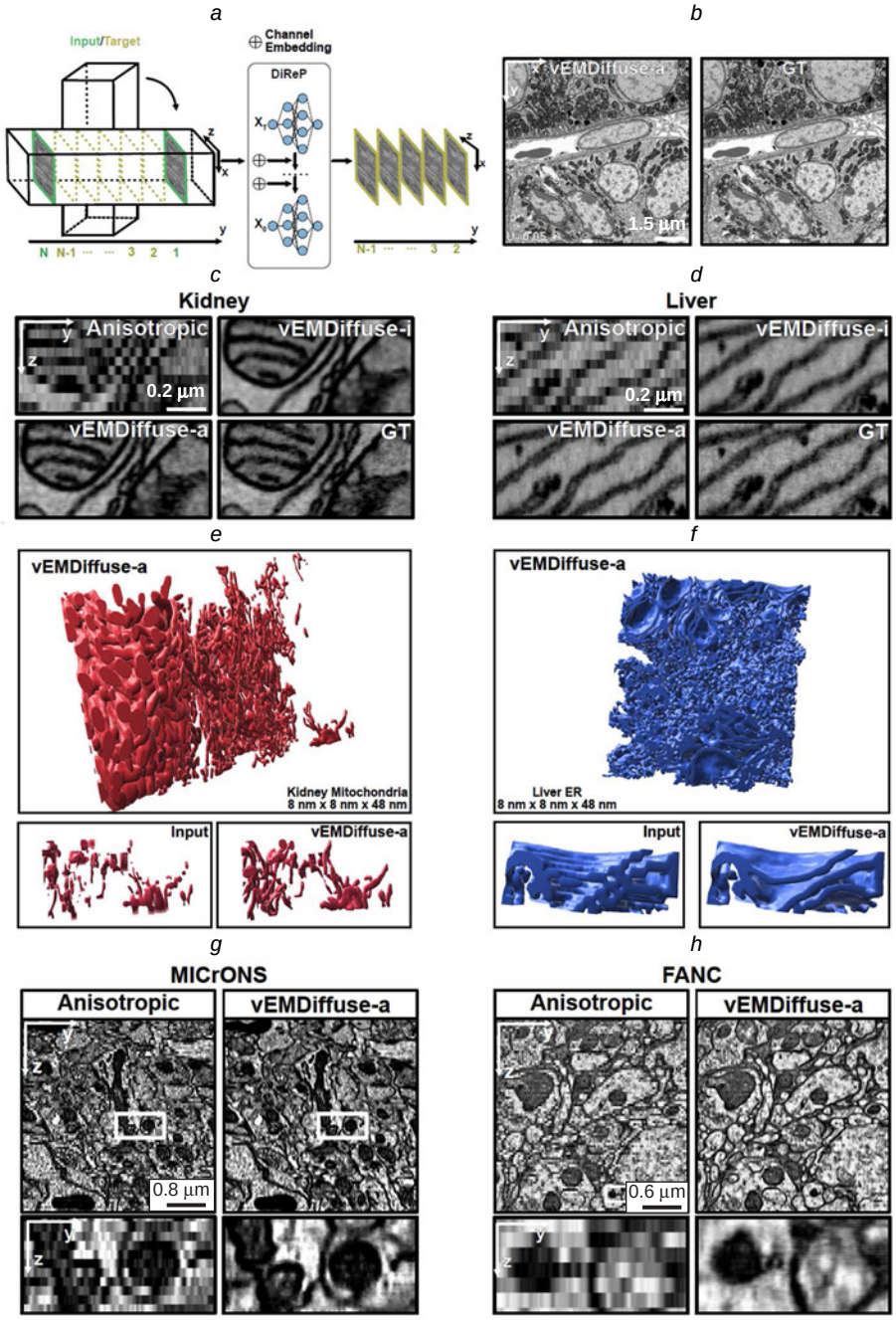


Figure 20. Isotropic volume reconstruction from anisotropic volume EM data using vEMDiffuse-a. Schematic of vEMDiffuse-a (*a*). Representative generated XY view (mouse kidney) (*b*). The method generates intermediate layers, restoring 3D continuity of organelles (*c*–*f*) and scales to large public datasets (*g*, *h*), illustrating a generative approach to reconstruction with built-in reliability assessment. DiReP is a diffusion model that preserves resolution; GT is ground truth image; ER is endoplasmic reticulum; MICrONS and FANC are datasets (*g*, *h*). The figure is reproduced from Lu *et al*.[90] according to Open Access CC BY-NC-ND 4.0 License.

A useful way to read the field is therefore task-first rather than modality-first. Earlier sections of this review discussed synthetic data in SEM, TEM/STEM, cryo-EM, and volume EM separately. Here the emphasis shifts to what synthetic data actually allow the investigator to do across modalities. Since the application of synthetic data to specific tasks have already been described in detail in the sections devoted to different electron microscopy techniques, this chapter will be general and concise, offering the reader an alternative perspective on the classification of the studies under consideration. This task-oriented view is especially important because visual realism alone is not sufficient: a synthetic dataset may look convincing yet still fail to improve counting, alignment, reconstruction, or quantitative measurement. Conversely, relatively simple simulations may be highly effective when they capture the nuisance factors that dominate a particular downstream objective. In Section 10, the central question becomes not whether synthetic data can be generated, but when they genuinely help and how that help should be validated.

9.1. Classification

Classification was among the earliest application domains in which synthetic EM-like data proved useful. In heterogeneous

3D-EM projection analysis, synthetic datasets were used to study how noisy projection images can be partitioned into homogeneous groups through graph construction and approximate Max k-Cut optimization, establishing a controlled setting for evaluating computationally demanding classification strategies.[165] Although this work predates modern deep learning, it already captured a durable lesson: synthetic data are particularly valuable when the true class structure is physically meaningful but difficult to verify in experimental images.

In materials and compositional analysis, classification tasks increasingly rely on synthetic data to bridge the gap between scarce labels and high-dimensional measurement space. For mineral-phase recognition in unpolished samples by SEM – EDS combined with advanced neural networks, generation of synthetic EDS data were used on the training dataset preparation step. Neural networks trained on synthetic EDS maps demonstrated high accuracy on the real data, with the Transformer model outperforming a U-Net baseline (97.0% *vs.* 63.7% accuracy) and a 3D ResNet (92.3%).[170] Another example of the synthetic data application for the solution of chemically-related image analysis task is a classification of cementitious materials on the basis of electron microscopy data. In a relevant study, semiautomatic analysis combined denoising, hierarchical segmentation, user interaction, and support vector machine-based class assignment to distinguish materials in SEM images, illustrating how synthetic or semi-synthetic task structure can support class decisions even when the imaging scene is chemically and morphologically heterogeneous.[171]

Recent work expands classification beyond predefined labels toward representation learning. A quaternion-based vision transformer masked autoencoder, pre-trained on large-scale synthetic electron backscatter diffraction (EBSD) orientation maps, learned a latent space suitable for downstream anomaly detection in polycrystalline materials.[172] Similarly, unsupervised representation learning in massive EM volumes used synthetic augmentations that preserve semantic meaning to support query-by-example and pattern discovery without requiring exhaustive manual annotation.[173] Taken together, these studies show that synthetic data aid classification not only by enlarging training sets, but also by shaping the geometry of the feature space in which classes, clusters, and anomalies become separable.

4D-STEM polarization mapping provides a particularly stringent classification example. Models trained on idealized synthetic diffraction patterns achieved high accuracy within the synthetic domain, but transfer to experimental data exposed a substantial domain gap; prototype-representation and PCA-based approaches, together with augmentation and filtering, were more promising than straightforward classification, although practical reliability remained limited.[174] The study therefore illustrates why synthetic-domain accuracy must be separated from experimental transfer performance.

9.2. Detection and counting

Detection and counting are among the most direct beneficiaries of synthetic data because the required labels are explicit, localized, and expensive to acquire at scale. In NP analysis, a Modified BlendMask framework combined synthetic images, augmented images, and instance segmentation to automate particle counting in SEM images. The approach tackled overlapping, blurring, sparsity, and the occurrence of minuscule particles, reaching an overall counting accuracy of approximately 70%.[100] The key point is not merely that synthetic images increased the training set, but that they allowed systematic control over nuisance variables that are otherwise hard to sample evenly from real experiments.

A similar pattern appears in biological object detection. For automatic recognition of herpesvirus secondary envelopment stages in TEM images, synthetic labeled images generated by a GAN expanded the training set from 315 ground-truth images containing 2610 capsids to a combined set with 9192 labeled capsids. This augmentation raised average precision from 53.81% to 76.48% for Faster Region-based Convolutional Neural Network (Faster R-CNN) detection (Fig. 21).[175] Such results are particularly significant because the downstream goal is not generic object recognition but biologically specific discrimination among subtle ultrastructural states.

Broader reviews in materials microscopy confirm that this pattern extends far beyond a few case studies. Object detection models now serve defect analysis, NP quantification, *in situ* tracking, and atom-scale localization, but their success remains strongly dependent on whether the training distribution reflects

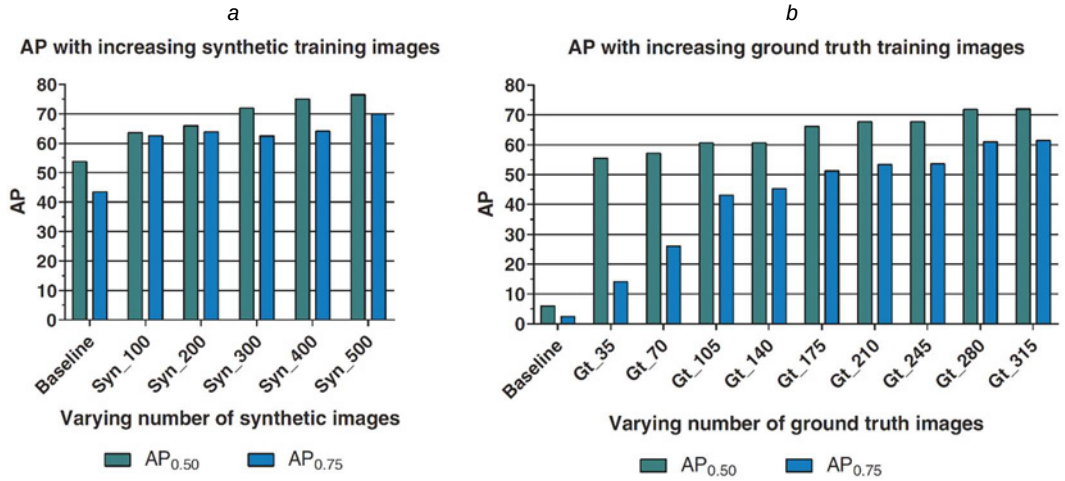


Figure 21. Quantitative evaluation of GAN-based synthetic image augmentation for detecting herpesvirus capsid stages (*a*, *b*). Training with 315 real images ($AP_{0.50}$ = 53.81%) plus 500 synthetic images achieves $AP_{0.50}$ = 76.48%, demonstrating that synthetic labeled data can substantially improve detection performance when real labeled data are scarce. AP is average precision. The figure is reproduced from Devan *et al*.[175] according to Open Access Creative Commons Attribution License.

the true domain of application.[169] Synthetic data are therefore valuable not simply as a larger pool of images, but as a way to enforce task-relevant balance among rare events, overlapping objects, weak-contrast cases, and physically plausible backgrounds.

9.3. Semantic and instance segmentation

Segmentation extends detection by requiring every pixel, region, or object instance to be assigned in a structurally meaningful way. This is precisely where synthetic data often become indispensable, because dense ground truth is tedious to create and may be inconsistent between annotators. The NP-counting study above already shows how instance segmentation can be built around synthetic masks.[100] In SEM analysis of cementitious materials, which was discussed earlier in the context of solving the classification task, segmentation was coupled with classification through block-matching denoising, binary partition trees, and supervised classification, demonstrating that segmentation quality directly conditions the reliability of subsequent material quantification.[171]

In filamentous-structure analysis, algorithm DSeg illustrates another use of synthetic data: not to maximize visual realism, but to verify algorithmic behavior under controlled shape and temporal conditions. The method was validated on synthetic data, differential-interference-contrast images, and TEM images of bacterial adhesion fimbriae, emphasizing the value of synthetic ground truth for testing backbone extraction, continuity preservation, and time-series robustness.[176] In a very different SEM context, a deep-learning framework for hardware assurance of digital integrated circuits relied on synthetic data and semi-automated label generation to support semantic segmentation, stitching, movement regression, and feature extraction in extremely large image collections.[177]

These examples underline that segmentation should not be evaluated only by pixel overlap. For EM, topological continuity, boundary fidelity, small-object recall, and compatibility with downstream counting or metrology are often more consequential than a single mean IoU value. Synthetic datasets are especially powerful here because they allow the same underlying object to be rendered under many nuisance conditions while preserving exact masks, instances, or centerlines.

9.4. Denoising, restoration, and super-resolution

Synthetic data are equally important for restoration tasks, where paired observations are rarely available experimentally. In secondary-ion mass spectrometry imaging, synthetic data were used to test pan-sharpening strategies that fuse chemically resolved secondary ion mass spectrometry images with higher-resolution SEM structure, thereby improving spatial detail without sacrificing chemical interpretability.[167] The significance of this study for the present review lies in its general lesson: synthetic data provide the only practical way to assess whether restoration improves a scientifically relevant signal rather than merely producing visually sharper images.

A more ambitious example is the recovery of clean 3D structure from noisy 2D STEM observations. By combining a differentiable image-formation model with an implicit 3D representation and an explicit noise model, the method of Kniesel *et al*.[178] disentangled signal and noise without supervised clean/noisy pairs, and was validated on both synthetic and real data. Here synthetic data are not just an auxiliary benchmark; they are central to defining the forward model and to testing whether learned restoration remains consistent with the underlying electron-optical process.

At the limit of low SNR, restoration and detection can no longer be separated cleanly. The multi-target detection framework motivated by cryo-EM showed that when direct detection is impossible, autocorrelation analysis and approximate expectation maximization can still recover the underlying signal from long noisy measurements without localizing each occurrence explicitly.[179] This is conceptually important for EM because many restoration problems are ill-posed in precisely this way: the question is not whether one can see an object directly, but whether enough statistical structure remains to reconstruct it indirectly.

9.5. Registration and alignment

Registration and alignment are ideal examples of tasks for which synthetic data can isolate the true source of algorithmic error. In serial-section EM of neuronal tissue, registration performance was analyzed with a spherical deformation model that simulates how neuronal structures change between adjacent sections. This analysis demonstrated that registration accuracy deteriorates as section thickness and characteristic structure size increase, thereby turning synthetic data into a tool for deriving error trends rather than merely reporting benchmark scores.[161]

The value of such work is methodological. In experimental section series, misregistration reflects many intertwined factors: structural change, cutting artifacts, intensity drift, staining differences, and algorithmic bias. Synthetic deformation models allow these contributions to be separated and quantified. This is crucial because poor alignment propagates directly into downstream 3D reconstruction, segmentation continuity, and morphometric measurement. Thus, synthetic datasets for registration should be judged by whether they reproduce the deformations that matter most for later biological or materials interpretation, not merely by whether adjacent images appear similar.

9.6. 3D reconstruction and tomography

Many electron-microscopy workflows are fundamentally inverse problems in which 3D information must be inferred from 2D observations. Synthetic data have long supported this class of task. In 3D SEM microscopy, multiple-view geometry combined with differential-evolution optimization was validated on synthetic and real data to reconstruct surface shape and calibrate SEM extrinsic parameters.[106] Although this work predates current data-centric deep learning, it shows that synthetic observations are indispensable whenever true 3D shape is unavailable but reconstruction fidelity must still be assessed quantitatively.

Modern methods push this logic further by embedding the reconstruction problem directly into a learnable forward model. The implicit-3D STEM framework discussed above[178] is one such case. Another, conceptually related example is the cryo-EM-inspired multi-target detection model, where reconstruction from noisy measurements is achieved without explicit detection of each target.[179] Across these studies, synthetic data are most powerful when they do not merely resemble the final image, but encode the generative pathway from latent structure to measured signal.

9.7. Quantitative metrology and uncertainty-aware measurement

Synthetic data are especially valuable for metrology because the central challenge is not only detection but calibrated measurement. In diffraction-pattern analysis from TEM images, synthetic tests were used to verify methods for accurate intensity recovery and localization of faint reciprocal-space features *via* expectation maximization, enabling statistically meaningful extraction of positions and intensities that would be difficult to validate otherwise.[166] In plant-cell biology, a broad set of quantitative immunogold-localization methods was illustrated with empirical and synthetic data, emphasizing that reliable measurement depends on random sampling, event counting, expected-*vs.*-observed comparisons, and explicit statistical testing rather than subjective visual impression.[180]

Recent spectroscopy workflows extend this task from geometry to chemistry. In STEM-EDS analysis, pan-sharpening non-negative matrix factorization (PSNMF) was validated on synthetic datasets with controlled noise levels and overlapping phases before being applied to real nanomineralogical and catalytic systems, showing how synthetic data can benchmark chemical unmixing, denoising, and phase quantification simultaneously (Fig. 22).[131] Reviews of object detection in materials microscopy make the same point from another angle: algorithmic metrics such as precision or recall are necessary, but materials-centric outputs such as defect density, size distributions, swelling, hardening, or composition are the quantities that actually determine scientific utility.[169]

For this reason, uncertainty-aware measurement in EM should be interpreted broadly. Synthetic data rarely provide uncertainty in a formal Bayesian sense, but they do provide

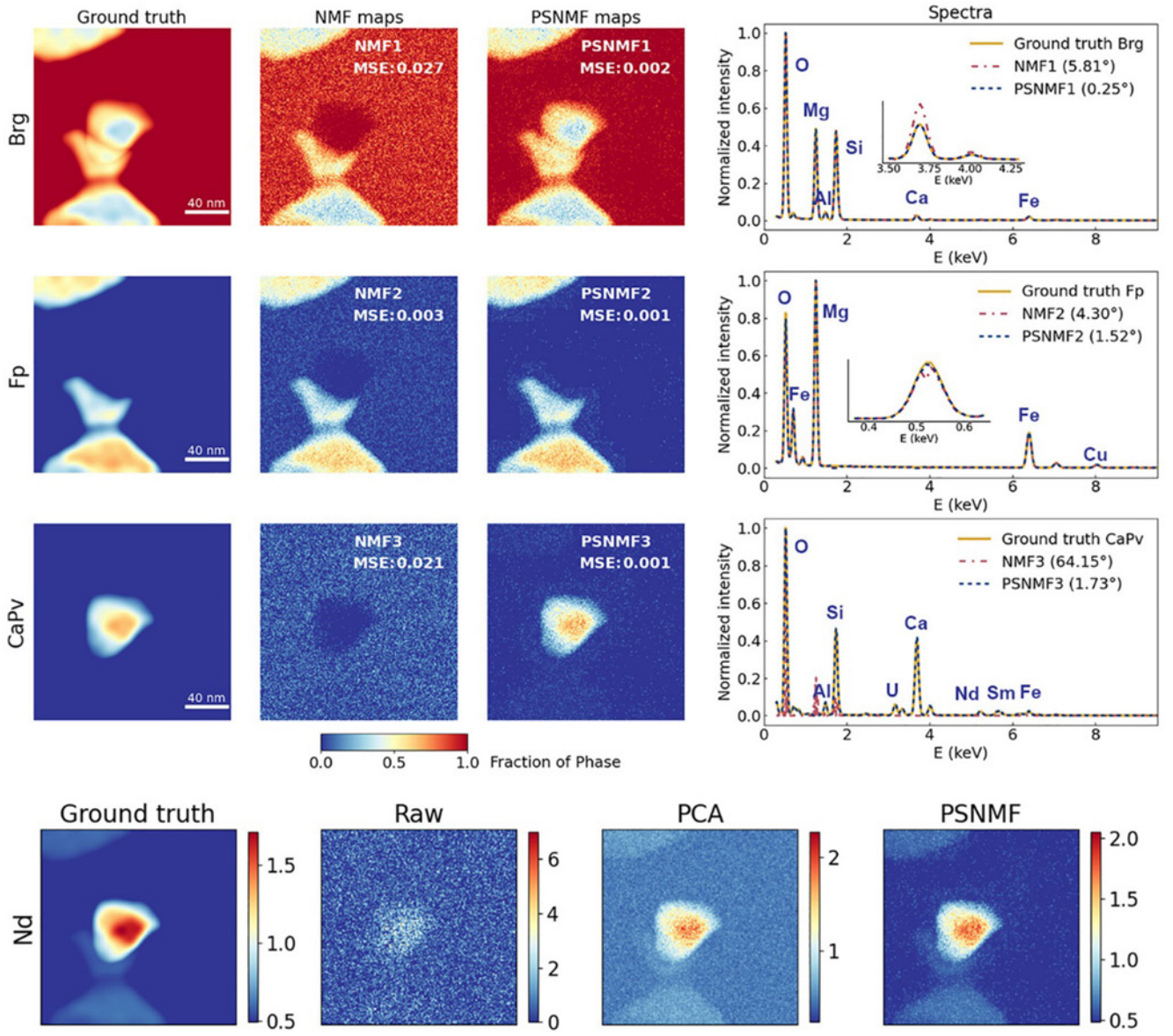


Figure 22. PSNMF validation on synthetic STEM-EDS data with three overlapping mineral phases. Compared to standard NMF (spectral angles up to 64°), PSNMF retrieves all phase spectra (angles <2°) and abundance maps (lower MSE), while also outperforming PCA denoising (Nd map). Synthetic data with known ground truth enable simultaneous assessment of chemical unmixing, denoising, and phase quantification. Brg, Fp, and CaPv denote the three phases; PCA is principal component analysis. The figure is reproduced from Chen *et al.*[131] according to Open Access Creative Commons License CC BY 4.0.

controlled perturbations, known truth, and repeatable nuisance factors. These ingredients are sufficient to estimate stability ranges, bias trends, breakdown points, and sensitivity to missing contrast or overlapping signals — precisely the information needed to decide whether a measurement is scientifically trustworthy.

9.8. Inverse problems and structure/property inference

The most conceptually ambitious use of synthetic data is to move beyond image interpretation toward inferring latent physics, process variables, or material performance. In materials microstructure evolution, variational system identification was demonstrated on synthetic datasets that mimic the properties of experimental microscopy images, enabling inference of the partial differential equations governing pattern formation from sparse and spatially unrelated observations.[168] Here synthetic data serve not as surrogate images but as an intermediary between observation and governing law.

Several recent studies make this transition from image space to property space explicit. In cement emulsified asphalt composites, a GAN was trained to generate synthetic X-ray diffraction (XRD) spectra and SEM images from design parameters, and a downstream deep network predicted dry-shrinkage behavior with an average error of 2.70%.[181] In cast-forged AZ80 magnesium alloy, a deep model predicted yield strength, ultimate tensile strength, and constitutive parameters directly from SEM microstructure images, while diffusion-based synthetic oversampling improved class balance and helped the model achieve low image-out prediction errors (Fig. 23).[182] In fatigue-crack prediction for polycrystalline super alloys, rule-based synthetic data, extended finite element method simulations, and *in situ* SEM observations were combined in a transfer-learning framework, improving the correct prediction rate from 57.3% to 70.6% when adapting from synthetic to simulated domains.[183]

Inverse inference can also target hidden functional information embedded in extremely large image collections. The hardware-assurance framework for integrated circuits used SEM image analysis, synthetic data, and multiple deep-learning modules to infer circuit-level attributes that are not directly visible in a single image.[177] Likewise, synthetic-data-based pretraining of EBSD transformers and unsupervised EM representation learning support anomaly detection and pattern mining in settings where the relevant target may not even be known in advance.[172,173] In this sense, synthetic data do not simply answer established scientific questions; they increasingly help formulate new ones.

Viewed across tasks, the literature in this section suggests a clear hierarchy of synthetic-data utility. The simplest role is dataset expansion for tasks such as detection, counting, and segmentation. A deeper role is stress-testing: synthetic data expose how performance changes with overlap, blur, section thickness, class imbalance, or noise level. The most advanced role is generative or mechanistic mediation, where synthetic observations connect latent structure to measurable images and then to physical quantities, reconstruction parameters, or material properties. These levels are not mutually exclusive; strong workflows often combine all three.

This task-oriented survey also reveals why generic claims about realism are insufficient. For classification, the decisive issue may be whether the latent feature space preserves class boundaries. For counting, it may be whether overlap and small-object statistics are realistic. For registration, the key variable may be deformation with section thickness. For metrology, it is bias and variance under controlled perturbation. For inverse problems, it is whether the synthetic generator preserves the causal link between image observables and the hidden variable to be inferred. Next section therefore follows naturally from the present discussion by asking when synthetic data actually help, what type of sim-to-real transfer must be checked, and which validation criteria distinguish visual plausibility from scientific usefulness.

10. Validation: when do synthetic data actually help?

The central lesson emerging from the literature is that synthetic data are most useful when they improve the validity of the downstream scientific task, not merely when they look convincing to the eye. Across electron microscopy, synthetic data have already been shown to support segmentation, classification, metrology, pose estimation, 3D reconstruction, and inverse modeling. Yet the same literature also shows that success is conditional: the synthetic distribution must cover the relevant experimental variability, the labels must remain physically meaningful, and evaluation must be performed on carefully separated real data rather than on synthetic holdouts alone. Thus, in this section of the review, the papers previously discussed from the perspective of data generation will be summarized in the context of validation tasks.

Taken together, the papers considered here argue for a multi-level validation framework. One must validate sim-to-real transfer, distinguish visual realism from task realism, assess fidelity at the pixel, object, and physics levels, expose performance dependencies through ablation and benchmark design, identify failure modes, and finally quantify uncertainty and out-of-distribution (OOD) awareness (see Fig. 1).

10.1. Sim-to-real transfer: what should be validated

Sim-to-real validation begins with a simple but often overlooked principle: the target of validation is not the synthetic image itself, but the mapping from synthetic training data to real-data performance. Early studies already made this point in practical form. In fibrous media analysis, a stochastic 3D fiber model combined with SEM simulation enabled thickness-measurement algorithms to be tested against known ground truth, precisely because real SEM images do not provide exact fiber diameters.[63] Likewise, artificial SEM images generated from simulated particle assemblies were valuable only after a segmentation network trained on them was compared against models trained on real data.[72]

A similar logic appears in rendering-based NP studies. Synthetic image rendering was proposed as a way to solve the annotation bottleneck, but the key claim was not that the renderings appeared realistic; rather, it was that networks trained on them achieved segmentation quality comparable to manually annotated workflows for toxicologically relevant NP ensembles.[71] UTILE-Gen extended this reasoning by showing that synthetic instance-segmented data can improve both classification and segmentation when real training data are scarce, with gains depending on model class and task definition.[41] In steel microstructure recognition, synthetic micrograph-annotation pairs improved extensibility to unseen micrographs under extreme class imbalance and very small data regimes.[79]

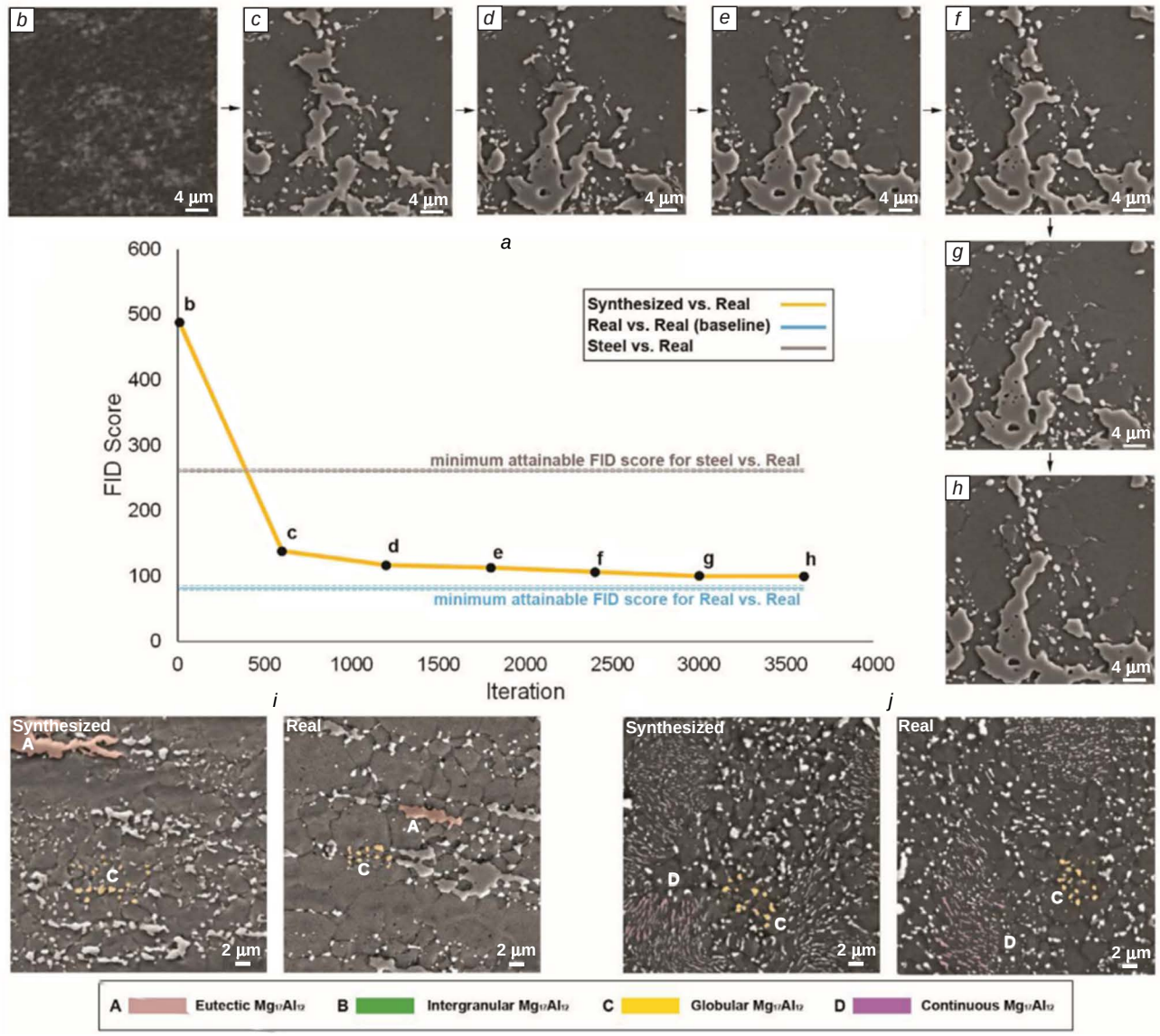


Figure 23. Comparison between synthetic and real SEM microstructures in a microstructure-to-property workflow for cast-forged AZ80 magnesium alloy, including the evolution of Fréchet inception distance (FID) during training and examples of synthesized *vs.* real images (*a*–*j*). This type of figure is especially informative because it ties visual similarity to the actual downstream goal of property prediction. The figure is reproduced from Azqadan *et al*.[182] according to Open Access CC BY-NC-ND License.

The most convincing studies validate sim-to-real transfer under explicitly varying experimental conditions. CycleGAN-based augmentation for STEM defect identification was designed to bridge the gap between multislice simulation and experimental images whose contrast, drift, and other nuisance factors vary substantially in practice.[81] In SEM-based 3D surface reconstruction for semiconductor metrology, unsupervised domain adaptation was required because simulated training images and experimental SEM images differed systematically; the method was then anchored to a second sensor to obtain physically meaningful height predictions.[45] In micro/nano-robotic pose estimation, GAN-based translation and feature-space matching were introduced precisely because direct transfer from simulation to SEM or OM imagery was inadequate.[46]

Microscope calibration provides an especially direct sim-to-real test. A recent VAE-EM framework trained a variational autoencoder on simulated Ronchigrams and then jointly estimated the mapping to real measurements and the aberration state through expectation maximization; symmetry constraints were used to restore identifiability, and real-STEM experiments showed approximately a twofold reduction in estimation error with fewer observations.[184] In this formulation, the simulation-to-reality gap becomes part of the calibration model rather than an unmodeled deployment error.

For cryo-EM and HRTEM, the same question becomes even more precise: which latent experimental variables must be matched? Roodmus illustrates the value of fully controlled benchmark datasets with known conformational ground truth for heterogeneous cryo-EM reconstruction.[37] The robust HRTEM framework of Rangel DaCosta and co-workers explicitly studies how performance depends on the distribution of atomic structures, imaging conditions, and simulation fidelity across benchmark experimental datasets.[64] Thus, sim-to-real validation should not ask whether synthetic data are generally helpful, but whether the synthetic distribution covers the operative domain of the intended experimental problem.

10.2. Visual realism *vs.* task realism

A recurring theme in the literature is that visually impressive synthetic images are not automatically useful, whereas visually simplified images can sometimes be surprisingly effective. A generative model DiffRenderGAN makes the point explicitly: its goal is to reduce the domain gap by learning rendering parameters that produce synthetic images and masks closer to real microscopy data, but the success criterion remains segmentation performance on real test images (Fig. 24).[185] The same lesson is implicit in earlier rendering work, where the point was not photorealism *per se*, but whether rendering could generate annotations that preserved class-defining features for the task at hand.[71] At the opposite extreme, the study based on self-generated simple figure patterns demonstrates that for two-class microscopy segmentation, task realism can occasionally matter more than image realism. The synthetic training data in that work are intentionally simple and highly abstract, yet the resulting models generalize surprisingly well across several microscopy datasets.[80] Synthetic dual-image generation for semantic segmentation makes a similar point from the perspective of evaluation: if labels are imperfect, a customized metric may be more informative than a standard one, because the validation task itself must reflect what is scientifically relevant.[66]

The limitations of visual realism are equally instructive. GAN-based synthetic SEM images can appear convincing yet still omit subtle factors that govern downstream robustness, such as texture irregularities, annotation errors, or scale-dependent cues.[72,185] Conversely, a simulator that looks only moderately realistic may nevertheless be highly useful if it preserves the statistics that control detection, segmentation, or measurement accuracy.

10.3. Pixel-level, object-level, and physics-level fidelity

The appropriate validation metrics depend on the downstream task. For segmentation, pixel-wise overlap measures such as Dice and IoU are standard.[64,186] For object detection, precision, recall, and F1 are common, often supplemented by application-specific counting or sizing errors.[45,187] For denoising and restoration, perceptual similarity metrics (Structural Similarity Index Measure (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Deep Image Structure and Texture Similarity (DISTS)) and distributional alignment scores (FID, Kernel Inception Distance (KID)) complement pixel-wise Peak Signal-to-noise Ratio (PSNR).[186,188] With these task-specific measures in mind, validation becomes more rigorous when fidelity is decomposed into distinct levels. Pixel-level fidelity concerns local correspondence between the synthetic image and what the model sees during inference: edges, contrast transitions, noise texture, blur, and boundary definition. This level matters strongly for segmentation. Papers focused on NP or microstructure segmentation therefore report Dice-type overlap, IoU-derived scores, or related pixelwise metrics to evaluate whether masks generated from synthetic data remain useful on real images.[66,71,185]

Object-level fidelity is more demanding. Here the question is whether objects are counted correctly, localized correctly, and assigned plausible size or shape descriptors. UTILE-Gen showed that synthetic expansion can improve precision, recall, and F1 for NP instance segmentation, which is an object-level gain rather than only a pixel-level one.[41] Synthetic data were also used to evaluate particle-size measurement pipelines in TEM and related microscopy contexts.[127] Importantly, object-

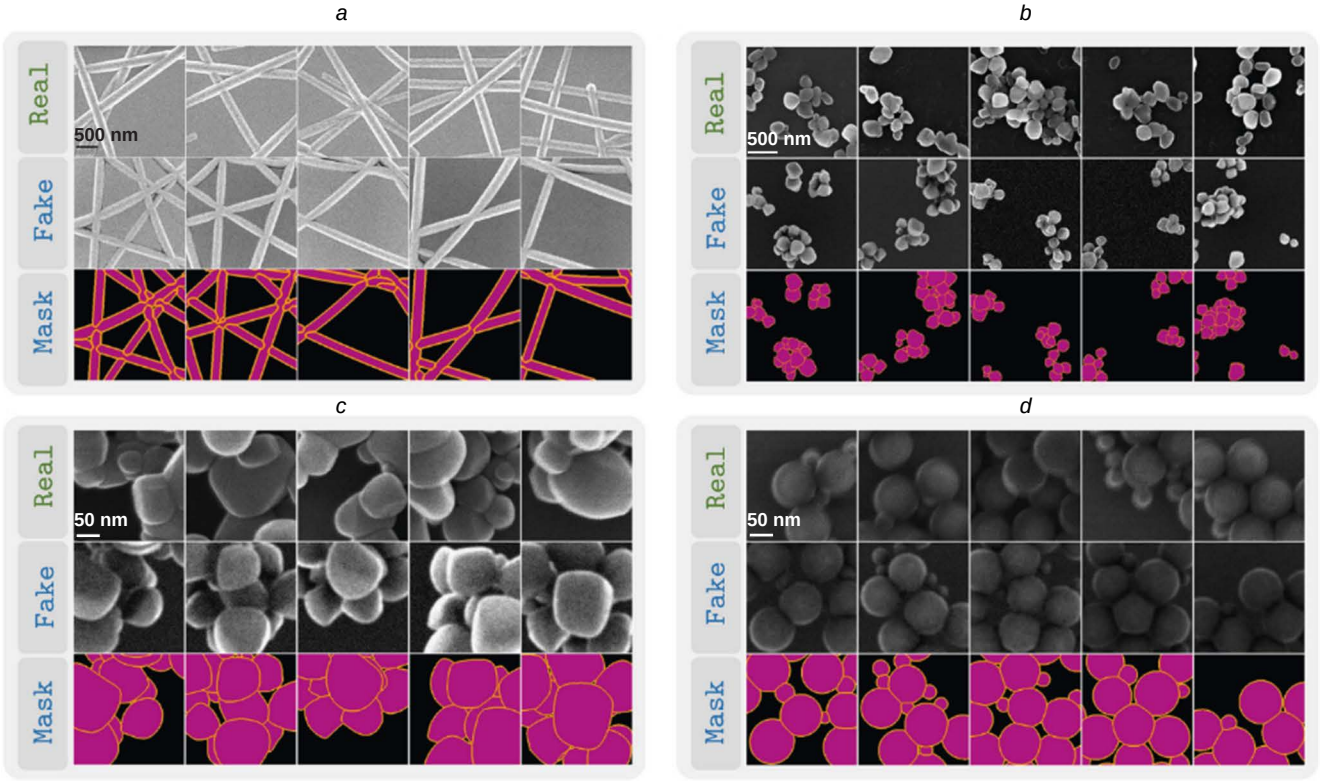


Figure 24. Representative real and DiffRenderGAN-generated synthetic microscopy patches with corresponding segmentation masks, illustrating how improved appearance matching can narrow the domain gap, while still requiring downstream task validation. The figure is reproduced from Possart *et al.*[185] according to Open Access Creative Commons Attribution 4.0 International License.

level fidelity is often where domain gaps first become visible: a model may segment acceptable masks but still merge neighboring objects, split single objects, or distort the recovered size distribution.

Physics-level fidelity is stricter still. The point is no longer simply to match annotations, but to preserve physically meaningful quantities. The fiber thickness study is exemplary because the entire simulation pipeline was built so that thickness values were known exactly, allowing measurement bias to be quantified against ground truth.[63] In semiconductor metrology, synthetic-domain training was only meaningful insofar as the final 3D reconstruction agreed with external measurements in nanometer units.[45] The study of metrology variability in proximity-effect modeling makes the same argument in reverse: injected measurement bias can be absorbed by a calibrated model, leaving a deceptively small MSE while actually degrading predictivity on other patterns.[105]

This multi-level view also highlights a subtle but important trade-off. Recent work on rough surface characterization has proposed numerical and topological metrics that exhibit scale invariance across diverse measurement techniques and length scales.[189] While such universal metrics are attractive for comparing surfaces across instruments and experimental conditions, their generality may render them insufficient for task-specific fidelity assessment: a synthetic dataset might preserve global topological statistics yet fail catastrophically on a particular downstream task such as NP counting or defect segmentation.[189] Thus, the choice of validation metrics must remain grounded in the specific scientific objective rather than in universal descriptors alone.

The analysis of particle-size distributions provides another caution. Synthetic data can validate counting or sizing pipelines, but the final scientific claim may still be distorted by downstream statistical choices. Alxneit demonstrated that inappropriate histogram bin widths can conceal important distributional structure, including missed bimodality, even when the underlying particle measurements are accurate.[190] Thus, true validation must follow the entire chain from pixels to objects to physically interpretable outputs.

10.4. Benchmark design, ablation studies, and volume effects

Because synthetic data are easily generated in large quantities, benchmark design is critical. The most trustworthy studies separate training on synthetic data from testing on real experimental datasets, and they report how performance changes under controlled variations of the training distribution. Recent HRTEM framework is particularly valuable here: it does not treat synthetic data as a monolithic resource, but investigates how simulation fidelity, the distribution of atomic structures, and the distribution of imaging conditions each influence performance on three benchmark experimental datasets.[64]

Volume effects are equally important. More synthetic data do not automatically imply better performance; the marginal utility depends on diversity, class balance, and complementarity to the real data. In steel microstructure recognition, synthetic augmentation was especially effective because it compensated for an extremely small and imbalanced real dataset, and the authors explicitly examined the effect of synthetic-data proportion.[79] UTILE-Gen likewise showed that the benefit of synthetic data depends on the task and model architecture, with some models benefiting more than others from mixed real-synthetic training.[41] The synthetic dual-image study further found that combining original and synthetic images was better than using either type alone.[66]

Ablations can also reveal which synthetic priors matter. The simple-figure-pattern study is instructive because it systematically removed certain pattern types and showed that some design choices degrade generalization more than others.[80] In cryo-EM, Roodmus provides a different benchmark logic: by generating micrographs with known conformational ground truth, it enables heterogeneous reconstruction algorithms to be compared against a reference that is unavailable in ordinary experimental data.[37]

Furthermore, a critical methodological requirement is that evaluation should be deliberately conservative. DiffRenderGAN, for example, limited postprocessing during testing so that the synthetic training data would be judged on their raw ability to support segmentation rather than on downstream clean-up steps that might mask data-quality deficiencies.[185]

10.5. Failure modes: shortcut learning, domain gaps, and hallucinated realism

The failure modes reported across these papers are highly instructive because synthetic-data workflows are especially vulnerable to hidden shortcuts. One common failure mode is domain-gap overconfidence: the model performs well on synthetic holdouts, or on real images close to the synthetic distribution, but degrades sharply when experimental conditions drift. This risk motivated domain-adaptation methods in SEM pose estimation,[46] CycleGAN-based realism enhancement in STEM,[81] and differentiable rendering in NP segmentation.[185]

A second failure mode is hallucinated realism: synthetic images may look authentic while carrying label errors, oversimplified textures, or hidden biases. This concern is visible in the comparison between earlier synthetic NP methods and DiffRenderGAN, where underperformance was attributed not merely to insufficient realism but also to visual artifacts, training instability, or omission of class-important features.[185] In cryo-EM benchmarking, the distinction between visually similar projections with sharply different downstream utility was emphasized in the need for benchmark datasets whose labels and heterogeneity are known rather than merely plausible.[37]

A third failure mode is analytical shortcut learning. In materials-by-design settings, a generator may capture visually plausible microstructure while still failing to preserve the mechanistic variables that govern property prediction.[75] In measurement problems, low error on a fitted calibration set can conceal biased parameter estimation and poor extrapolation, as the metrology-variability study demonstrated.[105] Likewise, the paper highlighting the risks in particle-size-distribution analysis reminds us that seemingly benign post-processing choices, such as selection of the bin width, can create false confidence in quantitative outputs.[190] The recent HRTEM OOD study reinforces this perspective by showing that generalization outside the training distribution is not binary but typically degrades smoothly as imaging conditions move away from the training domain (Fig. 25).[191]

10.6. Confidence scoring, uncertainty, and out-of-distribution detection

If synthetic data models are to be used in practice, they must signal when they are likely to fail. The clearest example in the present literature is image-wide confidence scoring for synthetic-data-trained electron microscopy models. In the study,[187]

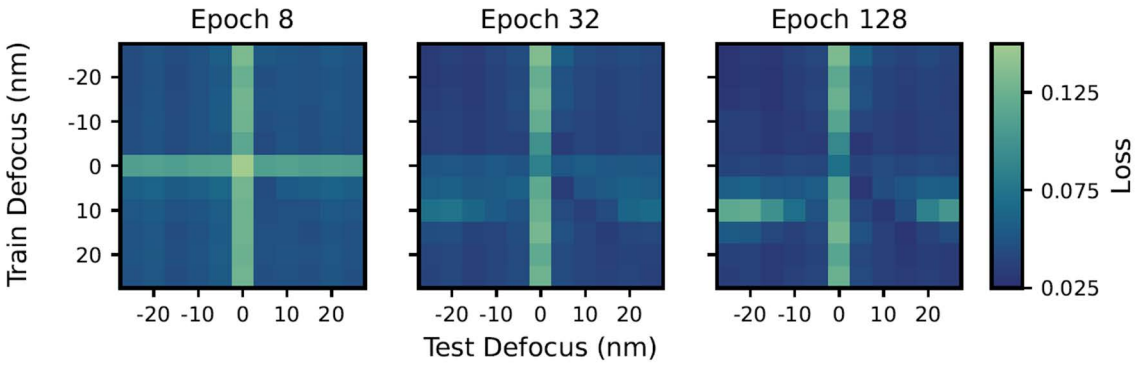


Figure 25. Out-of-distribution generalization and loss maps for HRTEM segmentation models across defocus conditions, showing how benchmark design can expose the dependence of performance on controlled shifts in imaging parameters. The figure is reproduced from DaCosta *et al.*[191] according to Open Access Creative Commons License CC BY 4.0.

feature-level confidences were aggregated into an image-level score that correlated with F1 performance and enabled problematic or out-of-domain images to be filtered, yielding performance gains of 5–30% at the cost of excluding roughly one quarter of the images (Fig. 26).

Confidence-aware validation is especially important for scientific imaging because many end users care less about mean benchmark scores than about whether specific images or regions can be trusted. A segmentation or metrology model may be excellent on average yet unreliable exactly in the rare cases that are scientifically decisive. Confidence scoring, uncertainty quantification, and OOD detection should therefore be regarded as part of validation itself rather than as optional deployment add-ons.

The HRTEM OOD study contributes a complementary perspective by linking generalization to measurable shifts in information content across datasets, which in turn offers a route toward anticipating when performance will worsen.[191]

Across the papers discussed within the whole section, the evidence converges on a clear conclusion: synthetic data help when they are validated against the scientific objective at multiple levels. Helpful synthetic data do not simply resemble microscopy images; they preserve the information needed for classification, detection, segmentation, metrology, reconstruction, or inverse inference on real experimental data. This requires benchmark designs with true real-data holdouts, ablation studies over data diversity and volume, physically meaningful end metrics, and explicit mechanisms for uncertainty and domain awareness.

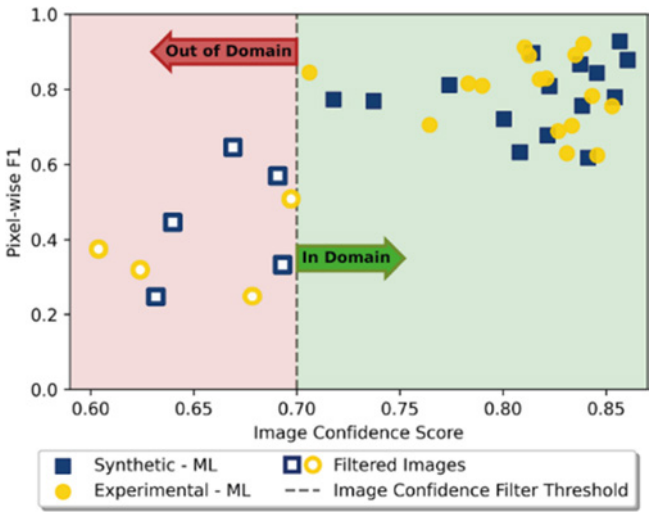


Figure 26. Image-level confidence scoring can predict per-image performance and support filtering of ambiguous or out-of-domain electron microscopy images. The figure is reproduced from Lynch *et al.*[187] according to Open Access CC BY-NC-ND License.

Once validation clarifies which kinds of mismatch matter most, the key practical question becomes how synthetic datasets should be designed in the first place: what should be randomized, what should be held fixed, how experimental variability should be sampled, how labels should be constructed, and how much realism is actually needed for a given task.

11. Design principles for useful synthetic datasets in electron microscopy

The main lesson emerging from the literature is that a synthetic dataset becomes useful only when its design is anchored to the scientific task, the operative image formation conditions, and the statistics of real microstructures. The decisive criterion is not whether a synthetic image looks impressive to a human observer, but whether it teaches a model the correct invariances, failure modes, and uncertainty structure for deployment on real EM data.[42,47,169]

11.1. Matching the data generator to the scientific question

The first design choice is not the network architecture but the generator archetype. A generator meant to support classification of global microstructural states can tolerate far more abstraction than one intended for pixelwise segmentation, quantitative metrology, or inverse structure-property inference. Broad reviews of microscopy AI consistently emphasize that task complexity and scalability determine which synthetic-data strategy is defensible.[42,192] In practical terms, this means that dataset generation should be built backward from the target output: class labels, object centers, instance masks, contours, physical parameters, or uncertainty scores.

Several papers make this task dependence explicit. UTILE-Gen deliberately uses domain randomization because the target tasks are NP classification, instance segmentation, and size analysis, for which controllable variation in shape, crowding, and background is more important than strict first-principles simulation.[41] In contrast, DiffRenderGAN introduces

differentiable rendering because the goal is segmentation of agglomerated NPs across SEM and ion microscopy conditions; there, geometry, occlusion, textural rendering, and automatically extractable masks must remain coupled (Fig. 27).[185] EMcopilot goes one step further for complex HAADF-STEM images by using morphology priors from real images, contrast-aware generative translation, and an EM-specific domain adapter, because the scientific question is not merely object presence but robust segmentation under heterogeneous support thickness, clustered particles, and microscope-induced perturbations.[186]

The same principle appears in studies that pursue quantitative inference rather than annotation alone. Synthetic filter-media datasets were built from simulated fibrous microstructures specifically to train CNNs that infer filtration efficiency and airflow resistance from SEM-like images, so the generator had to preserve those descriptors that control transport rather than reproduce every visual detail of real media.[92] Likewise, synthetic microstructures coupled to mechanics models have been used so that imaging data become inputs for direct property prediction, showing that the design of the synthetic dataset must align with the causal chain ultimately queried by the model.[168]

11.2. Choosing the right balance between physical accuracy and throughput

A second recurrent principle is that physical fidelity is not an absolute good. It is valuable only insofar as it improves sim-to-real transfer for the downstream task. DiffRenderGAN makes this point explicitly: the objective is not a fully physical simulation of materials, but a simulation that captures the characteristics necessary for a segmentation network to generalize.[185] This formulation is important because it reframes realism from an aesthetic target into an operational one. When throughput, parameter sweep breadth, and label quality matter more than full electron-scattering rigor, simplified renderers, Monte Carlo morphology generators, cut-and-paste domain randomization, or image-to-image translation can be preferable to expensive forward models.[41,193]

The literature therefore suggests a spectrum. At one end, physics-based or concept-driven simulators encode morphology distributions, imaging geometry, and noise sources directly. The concept-oriented synthetic TEM framework for grain extraction is a strong example: it begins from explicit concepts such as convex grains, boundary occurrence rules, and lognormal grain-size distributions, then constructs binary targets and corresponding TEM-like inputs that embody those concepts without pretending to be exhaustive first-principles microscopy.[78] At the other end, semi-synthetic workflows deliberately degrade experimental images to produce paired low/high-quality data. PSSR2 improves point-scanning super-resolution by semi-synthetic controlled degradation, because for restoration tasks the key requirement is faithful degradation of sampling, blur, and noise rather than de novo scene generation.[121]

Hybrid approaches often offer the best compromise. Ly *et al*.[193] combine a modified Potts model with image-to-image translation so that simulated uranium-oxide particles inherit realistic SEM appearance before being used for pretraining; importantly, pretraining on synthetic data followed by finetuning on experimental data performs better than naïvely pooling the two domains. This result illustrates a broader design rule: when the synthetic domain is structurally informative but visually incomplete, it should often be used to initialize representation learning rather than treated as a substitute for the real target domain.

11.3. Encoding morphology distributions, not only idealized objects

Many early synthetic datasets failed because they encoded idealized particles or defects but not the population statistics of

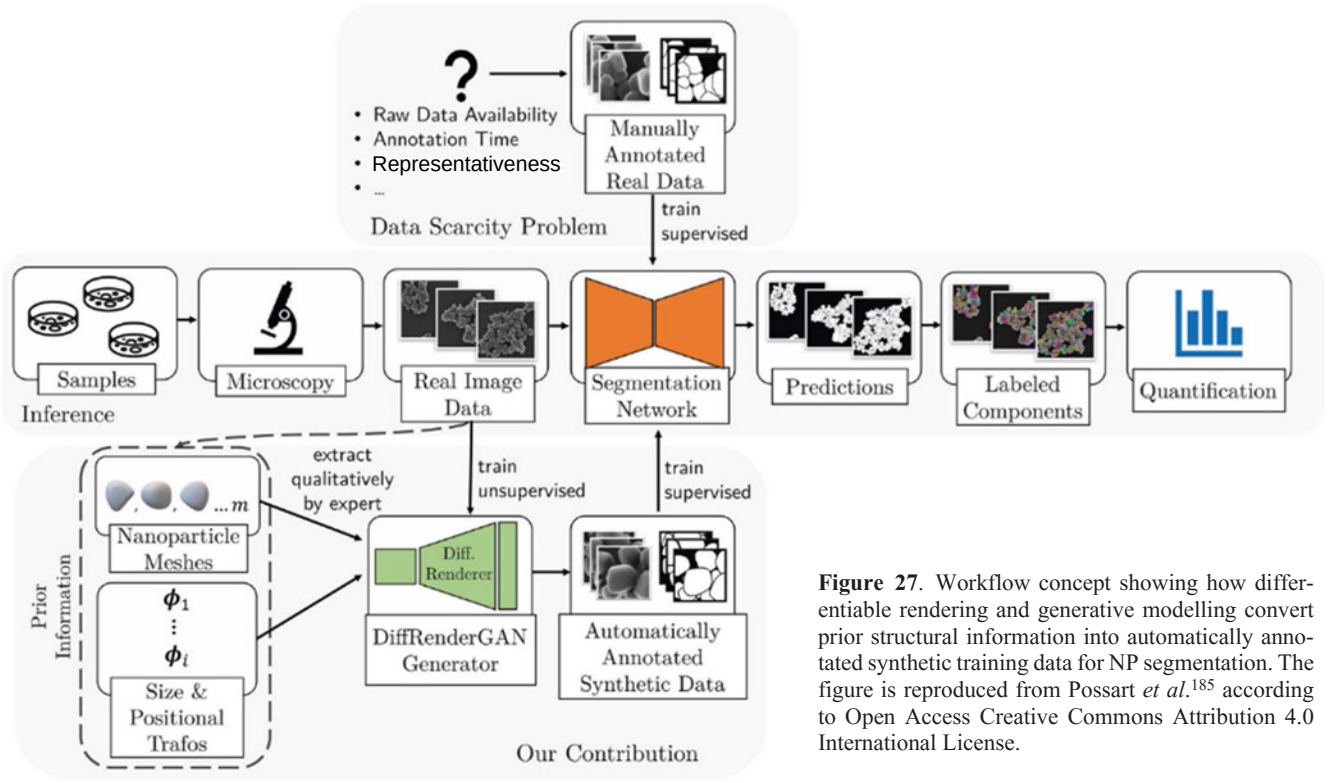


Figure 27. Workflow concept showing how differentiable rendering and generative modelling convert prior structural information into automatically annotated synthetic training data for NP segmentation. The figure is reproduced from Possart *et al*.[185] according to Open Access Creative Commons Attribution 4.0 International License.

the real sample. Distribution-level validation therefore matters: metrics such as the Kolmogorov-Smirnov test, Wasserstein distance, or Kullback-Leibler divergence provide quantitative ways to compare size, area, or shape distributions between synthetic and real data, complementing per-instance overlap measures.[63,186] The field has now moved toward distribution-aware generation. EMcopilot validates generated masks against distributions of area, perimeter, eccentricity, circularity, aspect ratio, and solidity, showing that preserving morphological heterogeneity is essential for reducing hallucinations and domain gap (Fig. 28).[186] DiffRenderGAN similarly uses meshes, placement rules, and size distributions to ensure that object ensembles, not only isolated objects, resemble the measured systems.[185]

The micrograph-annotation strategy for additively manufactured steel makes the same point from a different angle. There, the generator is used not only to produce more training images, but to rebalance the distribution of martensite/austenite fractions in an extremely small and imbalanced dataset.[79] This is highly instructive for EM practice: useful synthetic data should preserve the frequency distribution of meaningful structural states, not merely augment image count. In a related way, the concept-oriented TEM work explicitly chooses a lognormal grain-size distribution and weighted Voronoi construction because grain-size statistics are themselves part of the knowledge the model must learn.[78]

For segmentation and metrology, this requirement extends to overlap, crowding, clustering, and rare topologies. UTILE-Gen introduces domain randomization exactly to span variation in NP shape families, spatial distributions, and overlap conditions.[41] The V_2O_5 nanowire segmentation study likewise shows that synthetic images should reproduce geometrical and visual features of intertwined, elongated, and partially overlapping objects if the model is expected to transfer across modalities, including SEM.[128] Recent work on balancing alloy microstructure classes with diffusion-generated SEM images further confirms that synthetic oversampling is useful when it restores class proportions without erasing morphometric realism.[182]

11.4. Modelling imaging artifacts explicitly

A synthetic dataset becomes substantially more useful when it includes the artifacts that dominate real failure cases. Reviews of microscopy deep learning repeatedly identify noise, charging, drift, focus variation, blur, and reconstruction artifacts as major obstacles to robust training.[42,47] Consequently, artifact modeling should not be an afterthought appended after image generation; it should be part of the generative specification.

EMcopilot formalizes this by introducing a domain adapter that models noise, drift, aberrations, and radiation-induced perturbations after contrast-aware synthesis.[186] DiffRenderGAN reduces label inaccuracies and visual artifacts that afflicted earlier GAN-only pipelines by constraining image formation through rendering parameters tied to scene structure.[185] For FIB-SEM specifically, GAN-based curtaining generation demonstrates that some artifacts are difficult to simulate analytically and may need dedicated data-driven modules if they are to be represented credibly in benchmark datasets.[83]

Semi-synthetic degradation remains especially powerful for restoration problems because it preserves the content of real images while injecting controlled corruption. PSSR2 shows that degradation models must reflect the actual undersampling, blur, and noise that generate the low-quality measurements; otherwise the super-resolution model learns to invert an unrealistic corruption operator.[121] Even simple perturbations can matter: work on synthetic cell-counting data reported improved real-image performance when salt-and-pepper noise was injected into the synthetic training set, underscoring that the right noise model may improve transfer more than additional nominally clean images.[159]

11.5. Dataset diversity, class balance, and rare-event representation

The usefulness of a synthetic dataset depends not on its absolute size but on how well it allocates examples across common, rare, and ambiguous cases. This is a particularly acute issue in EM because the events of greatest scientific interest, *e.g.*, unusual defects, clustered NPs, minority phases, or pathological morphologies, are often precisely those that are hardest to annotate and least frequent in routine data. The steel micrograph-annotation study demonstrates that synthetic data can correct severe class imbalance and improve extensibility when real data contain only a tiny and skewed sample of phase fractions.[79] Similar conclusions arise in axon segmentation, where varying the amount of synthetic data materially changes performance under limited real-data conditions.[157]

However, diversity should be structured rather than indiscriminate. EMcopilot's ablation studies show that removing morphology-aware constraints widens the domain gap and increases hallucinations, even if the nominal number of synthetic samples rises (see Fig. 28).[186] Domain randomization, therefore, is most effective when random variables correspond to scientifically plausible degrees of freedom (shape family, size, orientation, spacing, support thickness, intensity variation, or background complexity) rather than arbitrary perturbations.[41,185]

Rare-event design also includes explicit domain awareness. Lynch *et al.*[187] show that image-wide confidence scoring can identify ambiguous or out-of-domain images, boosting performance after filtering and revealing that even strong synthetic generators must be paired with mechanisms that recognize when a sample lies outside the learned support. Thus, a mature synthetic benchmark should include both difficult in-domain cases and deliberately shifted out-of-domain cases if one wants to evaluate robustness rather than merely interpolation.

11.6. Annotation strategy and label ontology

Synthetic data are attractive because labels can be generated automatically, but label automation is not synonymous with label adequacy. The ontology of labels — what exactly is annotated, at what granularity, and under which conventions — directly conditions what a model can learn. For classification tasks, image-level labels may suffice. For detection tasks, center points or boxes may be enough. For segmentation, one must specify whether labels are semantic, instance-wise, boundary-enhanced, or hierarchical. For metrology, one often needs object contours and downstream measurement definitions that are themselves standardized.[42,169]

Several studies illustrate that ontology design is integral to dataset design. The V_2O_5 nanowire work uses polygon annotations suited to elongated, overlapping rods.[128] DiffRenderGAN generates multiclass masks that separate material classes and boundaries, which is more informative than binary foreground/background labels for crowded NPs.[185] In cryo-EM curation workflows, Class Aggregator for Key Electron-microscopy Data (CAKED) and related PERC tools emphasize that labels, classes, transforms, and data loading

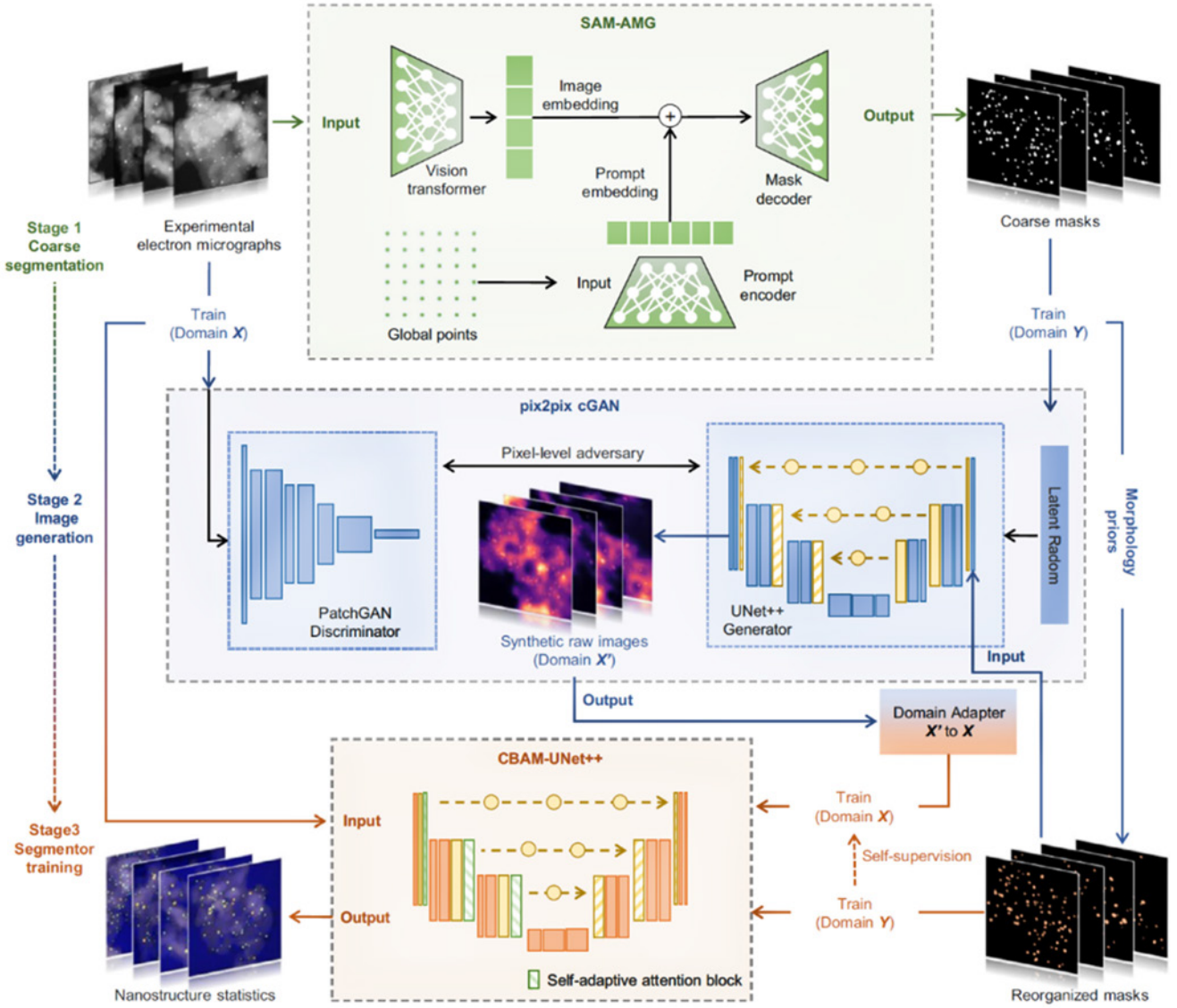


Figure 28. EMcopilot framework illustrating morphology-prior extraction, conditional image generation, and EM-specific domain adaptation. The figure highlights why useful synthetic datasets must jointly encode morphology, contrast heterogeneity, and microscope perturbations. The figure is reproduced from Yuan *et al*.[186] according to Open Access CC BY-NC-ND License.

conventions must be machine-readable and pipeline-compatible if synthetic and experimental datasets are to interoperate cleanly (Fig. 29).[48]

The soot-agglomerate study also illustrates why label representation is part of the scientific design: a 32-bit instance bitmask allowed several overlapping particles to retain independent pixel-level identities, avoiding the ambiguity that conventional single-label masks create in dense TEM aggregates.[126]

For quantitative biology-oriented EM, Python-based toolbox MyeliMetric is instructive because it shows that downstream measurement validity depends on biologically informed aggregation rules rather than raw segmentation alone; binning by axon caliber avoids misleading summary statistics and pseudo-replication.[194] The implication for synthetic dataset design is that labels should be created with the intended measurement protocol in mind. If the final scientific readout is a g-ratio distribution, a pore-size histogram, a defect-density estimate, or a grain-size law, then the ontology should support that inferential chain rather than stop at a visually pleasing mask.

11.7. Reproducibility, metadata, and open synthetic benchmarks

Finally, useful synthetic datasets must be reproducible. This requires more than releasing images and masks. One must also preserve generator parameters, morphology priors, simulation assumptions, rendering settings, augmentation schedules, train/validation/test partitions, and the exact evaluation protocol. Without this metadata, synthetic benchmarks become visually inspectable but scientifically irreproducible. The broad microscopy and materials reviews argue for community data ecosystems, curated benchmarks, and explicit reporting of data-preparation pipelines rather than opaque one-off datasets.[42,169,192]

PERC is particularly relevant here because it treats data curation, lazy loading, augmentation, and machine-learning compatibility as first-class parts of the workflow, enabling synthetic and experimental cryo-EM data to be assembled in interoperable pipelines.[48] Open release of representative data and code, as done for struct2im and several recent synthetic-data pipelines, should increasingly be considered a minimum

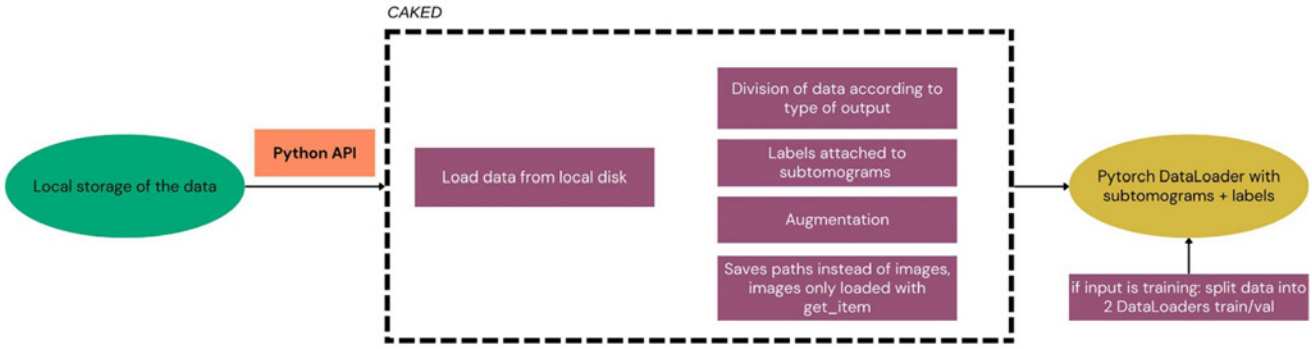


Figure 29. CAKED data-loading workflow: local images and labels are processed with cryo-EM-specific transforms and fed into a PyTorch DataLoader. The design enforces consistent class definitions, metadata handling, and preprocessing, enabling seamless interoperability between synthetic and experimental data in ML pipelines. API is Application Programming Interface. The figure is reproduced from Costa-Gomes *et al*.[48] according to Open Access Creative Commons License CC BY 4.0.

standard rather than an optional supplement.[193] Reproducibility also means publishing ablations: which morphology constraints mattered, which artifact models were included, what proportion of synthetic *vs.* real data was optimal, and how confidence or OOD filters changed performance.[121,186,187]

Taken together, these studies suggest a practical checklist for synthetic dataset release in EM: document the scientific task, disclose the generative assumptions, report morphology and artifact distributions, define the label ontology, specify balance across common and rare cases, release code and metadata, and benchmark on real external data rather than only on synthetic holdouts. When these principles are respected, synthetic datasets cease to be mere substitutes for missing annotations and become instruments for disciplined model design.

Once synthetic datasets are defined more rigorously — with explicit morphology priors, artifact models, uncertainty handling, and reproducible metadata — the field can move beyond narrow task-specific augmentation toward broader synthetic ecosystems: foundation datasets, closed-loop self-supervision, simulator-generator hybrids, and adaptive data engines that interact with microscopes in real time. Thus, careful design is the condition that makes the emerging directions of synthetic EM both credible and scalable.

Key take-home message. In electron microscopy, the best synthetic dataset is not the most photorealistic one. It is the one whose morphology statistics, artifact model, label ontology, and benchmark protocol are all aligned with the scientific question being asked (Table 1).

12. Emerging directions

The next phase of synthetic data in electron microscopy is no longer defined only by larger training sets or by better visual imitation of experimental images. A more consequential transition is underway: synthetic data are becoming an active computational layer that helps represent, regularize, infer, optimize, and even steer microscopy workflows. Across the literature surveyed here, several converging directions are visible. First, synthetic data are being used to pretrain representations and to inject priors into general-purpose models rather than only to provide supervised labels. Second, the boundary between simulator and learner is softening through differentiable rendering, surrogate forward models, and simulation-based inference, so that image formation and downstream interpretation can be co-designed. Third, synthetic data are increasingly embedded into closed experimental loops, where they help choose what to acquire next, how to reconstruct it, and how to connect experimental observables to latent physical variables. Finally, the most ambitious studies point toward digital twins of specimens, microscopes, and experiments, where synthetic data are generated adaptively rather than statically and are tied to uncertainty, metadata, and decision-making.[91,185,186,195]

12.1. Foundation models and synthetic pretraining for microscopy

A major emerging shift is the move from task-specific synthetic datasets toward synthetic pretraining resources that shape the

Table 1. Practical synthesis checklist.

Principle	What to encode	Representative examples[a]	Refs
Generator-task match	Labels, observables, and difficulty that mirror the downstream scientific question	UTILE-Gen DiffRenderGAN	41 185
Operational realism	Enough physics and appearance fidelity for transfer, but not gratuitous simulation cost	PSSR2 struct2im	121 193
Morphology distributions	Size, shape, overlap, clustering, and phase-fraction statistics, not just ideal objects	(DCGAN+Pix2PixHD) EMcopilot	79 186
Artifact modelling	Noise, drift, charging, curtaining, blur, undersampling, contrast variation	(Adapted Pix2Pix+CycleGAN) EMcopilot	83 186
Rare events and balance	Minority classes, ambiguous cases, and explicit OOD situations	Synthetics SIGMA-ML	157 187
Ontology and reproducibility	Machine-readable labels, metadata, code, splits, and benchmark protocol	PERC MyeliMetric	48 194

[a] All software tools discussed here were freely available at the time of writing. Their licensing terms and availability may change over time.

representation space before any downstream fine-tuning. In conventional EM workflows, synthetic data were often generated for one narrowly defined endpoint, such as segmentation or defect detection. The newer literature instead suggests that synthetic data can be used to learn broad visual priors, disentangled latent factors, and morphology-aware embeddings that remain useful across many tasks.

This trend is visible in several complementary forms. In large-scale EM volume analysis, unsupervised latent-feature learning has been used to create embeddings that preserve semantic similarity by generating synthetic positive pairs through augmentations that do not change biological meaning. In this setting, the learned representation enables query-by-example and clustering of previously unknown structures in massive EM datasets, shifting synthetic data from a labeling substitute to a mechanism for exploratory discovery.[173] A related direction appears in disentangled representation learning, where synthetic-data-based transfer is used to improve interpretability in microscopy image classification. The key point is not merely that the model becomes more accurate, but that synthetic factors of variation can be learned in a more controlled environment and then transferred to real data, providing a more transparent basis for downstream decisions.[196]

The integration of general vision models into EM-specific synthetic-data workflows further strengthens this foundation-model perspective. EMcopilot is particularly illustrative because it does not start from manual annotation; instead, it uses the Segment Anything automatic mask generator to extract coarse morphological priors from experimental EM images, then converts these into synthetic mask-image pairs for self-supervised training of a downstream segmentation model.[186] In other words, a broad visual model supplies the initial prior, while synthetic generation and domain adaptation specialize that prior for EM. This architecture shows that future 'foundation models for microscopy' may not be trained exclusively on real microscopy data; they may instead arise from hybrid pretraining pipelines in which general visual knowledge, synthetic microscopy generation, and limited experimental supervision are fused.

The likely implication is that synthetic data will increasingly be judged by how well they support transferable representations rather than by whether they maximize performance on one benchmark. For EM, this is especially important because many practical tasks suffer from low annotation volume, high instrument specificity, and rapidly changing sample classes. A useful synthetic pretraining corpus therefore needs to encode variability in morphology, contrast, artifacts, and acquisition conditions in a way that supports adaptation across domains. The literature suggests that this goal is realistic, but only if the synthetic generator is designed to preserve the factors that matter scientifically rather than simply producing visually plausible images.[185,186,196]

12.2. Differentiable simulators and simulation-aware learning

A second emerging direction is the replacement of rigid offline simulators by differentiable or learnable forward models that can interact directly with optimization and inference. This development is important because many microscopy problems are limited not by the absence of physics, but by the computational cost of deploying physically faithful simulation at the scale needed for machine learning.

DiffRenderGAN provides a clear example of this new logic. Rather than treating rendering as a fixed upstream step, it integrates a differentiable renderer within a GAN framework and optimizes rendering parameters to generate annotated synthetic images that better match real nanomaterial microscopy data across several modalities.[185] The conceptual advance here is substantial: the generator is no longer forced to choose between naive geometric drawing and brute-force physical simulation. Instead, rendering itself becomes an optimizable bridge between morphology prior and image realism. This is a strong indication that future EM synthetic-data pipelines will rely increasingly on differentiable image-formation modules.

A related but complementary strategy is the use of surrogate simulators. For HAADF-STEM NP analysis, Eliasson and Erni developed a 3D convolutional surrogate model that predicts STEM images from voxelized atomic models, achieving a roughly 100-fold speed-up over traditional multislice simulation while maintaining high image quality. This speed gain enabled generation of 100 000 synthetic multislice images, and a model trained with 4000 real plus 100000 synthetic images reduced the median size-estimation error from 9.89% to 5.26%.[69] The significance of this work extends beyond one metrology task. It demonstrates that when forward simulation is made fast enough, synthetic data cease to be a small auxiliary resource and become a scalable substrate for training.

Simulation-aware learning is also advancing from image synthesis toward scientific inference. In cryo-EM, cryoSBI uses simulation-based inference to amortize Bayesian template matching: training is performed once on simulated images, after which conformational inference on individual cryo-EM images takes only milliseconds and includes uncertainty estimates. Crucially, the method removes the need to estimate particle pose and imaging parameters explicitly for each image, showing how simulation and probabilistic learning can be merged into a practical inference engine.[144] In materials microscopy, variational system identification has been extended to infer governing PDEs from sparse, spatially unrelated, and multi-source microscopy observations, with the framework demonstrated on synthetic datasets that reflect the characteristics of experimental material micrographs.[168] These studies indicate that the most powerful role of synthetic data may be neither classification nor segmentation, but identification of latent mechanisms and dynamics.

Figure 9 highlights this broader movement through the CryoGEM workflow. There, physically grounded cryo-EM simulation is coupled with generative noise translation to create annotated training data that improve particle picking and *ab initio* reconstruction.[65] The key message is that future synthetic datasets will not be useful merely because they are large; they will be useful because they are linked to learnable forward models that preserve the operational structure of the microscope.

12.3. Self-supervised and generative representation learning

The third emerging direction is the maturation of self-supervised and generative learning as core strategies for EM analysis. Earlier use of synthetic data frequently assumed that the final goal was supervised learning on manual or simulated labels. The recent literature shows a broader agenda: synthetic data can serve as the source of invariances, contrast transformations, morphology priors, or latent traversals that make representation learning itself more robust.

EMcopilot is again a central example because it is explicitly designed as a label-free, closed-loop generative learning framework for EM segmentation (see Fig. 28).[186] The framework first extracts coarse masks from real images, then reorganizes them to increase morphological diversity, uses a pix2pix conditional GAN to generate contrast-aware synthetic electron micrographs, and finally applies an EM-specific domain adapter to model drift, noise, and related perturbations. The resulting synthetic pairs train a downstream segmentation model without manual annotation. Particularly important is the conceptual inversion of the usual workflow: instead of asking experts to annotate images and then training a model, the system distills morphology and contrast structure directly from the data and turns them into a reusable synthetic training resource. This is not simply data augmentation; it is synthetic self-supervision.

Other studies reinforce the same trend from different angles. DiffRenderGAN reduces the manual burden to selecting target meshes and training parameters while preserving the annotation-rich nature of synthetic data.[185] CryoGEM combines physical simulation with unpaired noise translation so that synthetic cryo-EM micrographs reproduce realistic corruption patterns rather than only ideal particle projections.[65] In a different domain, Lee *et al*.[74] used deep convolutional generative adversarial network (DCGAN), CycleGAN, and Pix2Pix to generate realistic micrographs of steels and even to convert hand-drawn sketches into plausible microstructural images, showing how generative models can turn human design intent into micrograph-like outputs.

These advances imply that self-supervised microscopy will increasingly depend on synthetic operations that preserve or expose the right latent factors. A synthetic image is no longer only a substitute for a scarce labeled real image. It can be a view, a perturbation, a counterfactual, a morphology-controlled sample, or a physically constrained corruption model. For EM, where contrast formation is often complicated and annotation is expensive, this broader generative view is likely to be decisive. It also suggests that benchmarking should extend beyond final task accuracy to include how well a synthetic-learning pipeline captures latent morphology distributions, artifact families, and uncertainty structure.

EMcopilot architecture combines a general vision prior, contrast-aware generation, and an EM-specific domain adapter into a single self-supervised loop.[186] This kind of architecture foreshadows a class of microscopy models in which synthetic generation is not a preprocessing convenience but the central mechanism by which training data are created and refined.

12.4. Synthetic data for multimodal microscopy and correlative workflows

A fourth direction is the extension of synthetic-data logic beyond one microscope or one image modality. As EM increasingly participates in correlative and multimodal workflows, synthetic datasets must bridge differing contrast mechanisms, scales, and information types rather than optimize only one image appearance.

DiffRenderGAN is especially important here because it was demonstrated across helium ion microscopy, SEM, and multibeam SEM datasets, showing that a single rendering-and-generation framework can support segmentation in materially and instrumentally distinct settings.[185] This is a meaningful step toward modality-agnostic synthetic pipelines. EMcopilot also points in this direction by demonstrating adaptability across magnifications, materials systems, HAADF-STEM, and even bright-field TEM scenarios.[186] Such results suggest that future synthetic-data frameworks will be valued not only for realism within a modality, but also for the extent to which they preserve a common representation of structure across modalities.

The recent MXene SEM-EDS workflow provides a concrete example of this multimodal direction: synthetic SEM training data generate morphology masks that are then intersected with co-acquired EDS maps to obtain per-morphology stoichiometry and impurity distributions.[103] Here synthetic data connect image segmentation directly to chemically resolved analysis rather than ending at morphology alone.

Synthetic multimodality also includes linking images to non-image observables. Beckman *et al*.[92] predicted air-filtration efficiency and airflow resistance from SEM images by training a convolutional model on synthetic media generated from a simulation framework combined with a single-fiber-efficiency analytical model. Although this study is outside classical EM segmentation, it is highly instructive for emerging microscopy workflows because it links morphology, image formation, and functional properties in one synthetic loop. This is precisely the kind of bridge required for correlative microscopy, where one wants not only to identify structures but to relate them to transport, reactivity, mechanics, or biology.

The broader materials-intelligence literature reinforces this trajectory. Current reviews emphasize intelligent characterization, multimodal data fusion, and cross-scale modeling as central pillars of next-generation materials R&D, explicitly moving toward digital-twin paradigms that connect synthesis, characterization, and application.[195] For EM, the practical consequence is that synthetic datasets should increasingly be designed as interoperable assets. They must align not only with detector intensities, but also with metadata, physical descriptors, correlative modalities, and downstream property models. This is a much more demanding objective than simple sim-to-real transfer, but it is also where synthetic data can add the greatest scientific value.

12.5. Human-in-the-loop and active learning pipelines

Another emerging direction is the redefinition of the human role. In older workflows, the human expert primarily supplied labels, curated data, and tuned reconstruction parameters. In the new generation of methods, synthetic data help shift human effort from repetitive annotation toward higher-level supervision of model behavior, uncertainty, and experimental goals.

The clearest example is the integration of active learning with simulation surrogates for autonomous experiments. Saranathan *et al*.[91] developed a workflow in which deep-kernel-learning surrogates are trained on simulated microscopy images to predict structure-property relationships and then iteratively select informative samples from molecular-dynamics trajectories. The explicit aim is to connect microscopy observables with simulated quantities such as defect energies and to use uncertainty-aware selection to reduce simulation cost while keeping the experiment-simulation loop responsive. Here, synthetic data are not merely supplemental training examples; they define the search space in which the model learns where information is most valuable.

A similar logic appears in autonomous tomography reconstruction. Millsaps *et al*.[164] showed that Bayesian optimization with Gaussian processes can automate parameter selection for compressed-sensing electron tomography, reducing compute time by about 80% for reconstruction of $SrTiO_3$

nanocubes while avoiding over-smoothing through momentum-based optimization. Although this study focuses on reconstruction hyperparameters rather than synthetic image generation in the narrow sense, it exemplifies an important emerging principle: synthetic or simulated benchmark environments can be used to teach the system how to make better decisions before or during analysis of real data.

EMcopilot again provides an intermediate model between manual and fully autonomous practice. Because it can rapidly self-train on newly generated synthetic data and because its inference time on graphics processing unit is on the order of 0.02 s per frame, it supports near-real-time analysis in live STEM experiments.[186] The human expert is therefore repositioned. Instead of drawing every contour, the expert can judge whether the model's generated priors, segmentation behavior, or uncertainty patterns are scientifically sensible. Synthetic data thus become a tool for elevating the human from annotator to supervisor of an adaptive system.

This logic is illustrated through a workflow that couples experiment control, prior experience, simulation-surrogate training, and simulation feedback.[91] The important insight is that active learning in microscopy will likely depend on adaptive synthetic data generation: the dataset must change as the experiment changes.

12.6. Autonomous experimentation and adaptive data generation

The transition from active learning to autonomous experimentation introduces a further change: synthetic data are no longer assembled once and reused indefinitely, but generated on demand in response to uncertainty, novelty, or target properties. Adaptive synthetic-data generation is therefore emerging as a critical capability.

The studies reviewed above point to several mechanisms by which this may occur. In CryoGEM, the synthetic training set is generated from a coarse reconstruction and refined through realistic noise translation, thereby augmenting weak points in the cryo-EM processing pipeline such as particle picking and *ab initio* reconstruction.[65] In EMcopilot, synthetic masks are reorganized statistically and then translated into image space with domain-specific perturbations, allowing the training resource to be rebuilt around the current morphology distribution and imaging context.[186] In the active-learning framework, simulated structures are sampled adaptively according to uncertainty and expected informativeness, producing a targeted rather than indiscriminate synthetic training set.[91]

This adaptive view is especially powerful when the synthetic loop is coupled to inverse or property-oriented tasks. The air-filter study by Beckman *et al*.[92] is notable because synthetic media were not generated merely to classify images, but to train a model that maps SEM-observed morphology to filtration efficiency and airflow resistance (Fig. 30). Conversely, Lee *et al*.[74] showed that generative models can map design intent and sketches into realistic micrograph-like outputs, hinting at the reverse route from desired structure or property to candidate morphology. Together these works suggest that autonomous EM pipelines may increasingly generate the next synthetic sample not because it looks realistic, but because it is expected to be maximally informative for a property, mechanism, or design objective.

Fast inference is also moving synthetic data closer to instrument-level operation. PtyRANNOSAUR converts large simulated 4D-STEM training corpora into near-live ptychographic reconstruction,[117] while VAE-EM calibration adapts a simulation-trained latent representation to day-to-day variation in a real STEM.[184] Together, these studies point toward closed loops in which synthetic models support not only interpretation, but also rapid reconstruction and instrument-state estimation.

More broadly, recent reviews of machine-learning-driven materials research frame closed-loop autonomous systems as a central organizing principle for intelligent materials discovery.[195] Within such systems, EM is not a passive characterization endpoint; it becomes one sensor inside an iterative decision process. Synthetic data will therefore need to be generated at the tempo of experimentation and with explicit awareness of what the system has already learned and what it still does not know.

12.7. Digital twins of microscopes, samples, and experiments

The most far-reaching emerging direction is the development of digital twins. In the context of electron microscopy, a true digital

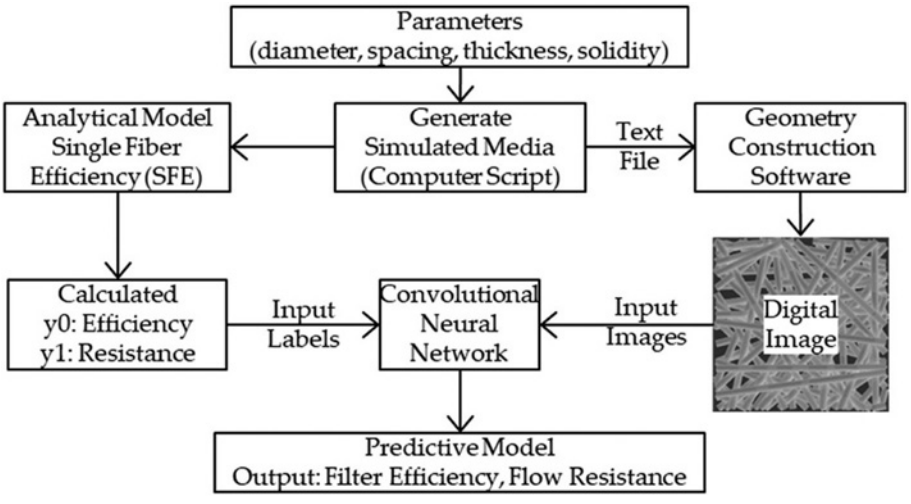


Figure 30. CNN-based prediction pipeline: synthetic media images and analytically computed properties (efficiency, resistance) train a regression model, which then maps a real SEM image to the same physical properties. The figure is published from Beckman *et al*.[92] according to Open Access Creative Commons Attribution (CC BY) license.

twin would not be a static simulation of image formation alone. It would couple specimen structure, morphology distributions, microscope physics, acquisition settings, reconstruction algorithms, latent physical variables, and uncertainty estimates into one continuously updated computational counterpart of the experiment.

The literature reviewed here mostly presents partial rather than complete digital twins, but the broader trajectory is increasingly visible. Surrogate multislice models for STEM dramatically reduce the cost of generating physically grounded image simulations.[69] CryoSBI shows how simulated measurement models can be converted into amortized Bayesian inference engines for conformational assignment.[144] CryoGEM demonstrates that physics-based simulation and learned noise translation can be coupled in a realistic cryo-EM synthetic-data engine.[65] EMcopilot adds morphology extraction, contrast generation, domain adaptation, and fast downstream analysis, bringing the synthetic loop closer to live operation.[186] The workflow closes the loop further by linking microscopy images with simulation-informed property surrogates inside an autonomous experimental framework.[91]

At the conceptual level, the digital-twin idea also requires a broader representation of structure-property relations. The steels study of virtual microstructure design illustrates one side of this by generating realistic micrographs from learned generative models and by envisioning image-based quantitative microstructure-property relationships rather than parameter-only descriptors.[74] The filtration study illustrates the other side by inferring functional performance directly from synthetic image-property pairs.[92] When these two directions are combined with adaptive simulation, uncertainty-aware inference, and real-time feedback, the outline of an EM digital twin becomes visible: a model that can predict how a structure should appear, what property it should imply, how the microscope will distort it, how confident the inference is, and what measurement should be acquired next.

The remaining gap is integration. Most current systems still treat generation, reconstruction, segmentation, inverse analysis, and control as separate modules. Digital twins will require these modules to exchange metadata and uncertainty in a principled way, remain stable across instruments and modalities, and update continuously as the experiment progresses. Nevertheless, the recent literature suggests that several of the necessary building blocks are already emerging.[91,168,185,186,195]

Taken together, these emerging directions indicate that synthetic data in EM are evolving from static stand-ins for scarce annotations into adaptive computational interfaces between microscope, specimen, and model. This evolution is intellectually exciting, but it also raises a harder question that cannot be avoided: when such systems become more capable, how should one judge realism, trustworthiness, transferability, and failure? That question leads directly to the unresolved challenges discussed in the next section.

13. Challenges and unresolved questions

Synthetic data are increasingly moving from a niche methodological aid toward a core enabler of electron-microscopy analytics, although the literature surveyed in this review also makes clear that the field is entering a more demanding phase. The central question is no longer whether synthetic data can help, but under what conditions they remain scientifically reliable, operationally transferable, and epistemically trustworthy. Studies on rendered NP ensembles, physics-based simulation, generative translation, differentiable rendering, and cryo-EM synthesis all report substantial gains in segmentation, detection, picking, metrology, or reconstruction.[41,64,65,68,71,73,81,185] At the same time, these studies repeatedly encounter the same unresolved issues: how much realism is enough, when annotation-free pipelines truly replace human labels, how far trained models travel across instruments and laboratories, how one should quantify domain boundaries, and how to prevent realistic-looking but scientifically misleading outputs.

13.1. How realistic is realistic enough?

A recurring lesson across SEM, TEM/STEM, FIB tomography, and cryo-EM is that visual plausibility alone is an insufficient criterion for synthetic data quality. Early successes in SEM NP rendering showed that networks trained on synthetic images can approach human-level annotation quality when object boundaries, background texture, and particle overlap are rendered in a controlled way.[73] However, later studies made the realism criterion more explicit: the generator must preserve the image statistics that matter for the downstream task, not merely produce images that look convincing to the eye.[81,185] DiffRenderGAN is especially instructive in this respect because it states directly that the goal is not a fully physically exact simulation, but a synthesis process that captures the image characteristics necessary for a segmentation model to generalize (see Fig. 24).[185]

The consequence is that realism in electron microscopy must be understood as task-conditional realism. For morphology-driven segmentation, the decisive ingredients may be contour sharpness, overlap statistics, substrate contrast, blur, and noise distributions. For atomic-resolution TEM/STEM, realism additionally depends on transfer-function-limited information content, defocus, thickness, aberrations, and beam-sensitive contrast behavior.[64,82,191] For cryo-EM, realism is inseparable from pose distributions, ice background, particle crowding, heterogeneity, CTF modulation, and the structured noise that simple Gaussian corruption fails to reproduce.[65,68,145] Thus, ‘realistic enough’ cannot be reduced to a universal aesthetic threshold; it must be benchmarked against the invariants of the intended scientific measurement.

This reframing also reveals why morphology-aware priors are so important. EMcopilot showed that removing morphology constraints widens the domain gap and increases hallucination risk, even when the generator remains visually expressive.[186] Similarly, synthetic microstructure pretraining works best when the generator encodes experimentally plausible particle ensembles rather than idealized toy objects.[193] In practice, the unresolved problem is to formalize realism as a hierarchy: first, object realism; second, artifact realism; third, information-content realism; and fourth, decision realism. The field has good heuristics for the first two, but much weaker theory for the latter two.

13.2. Can synthetic data replace manual annotation, or only reduce it?

The literature now shows that synthetic data can substantially reduce dependence on manual annotation and, in selected benchmark settings, can match or occasionally outperform models trained on limited human-labeled datasets. Rendered SEM NP datasets, domain-randomized generators, and closed-loop generative frameworks have all demonstrated competitive

or superior segmentation relative to human-annotated baselines.[41,71,186] Synthetic STEM or HRTEM workflows similarly show that physically generated labels can support state-of-the-art segmentation of experimental atomic-resolution images.[64,81] In nuclear materials microscopy, synthetic training even yielded quantitative performance close to experimental-label pipelines while avoiding the inconsistency of human markup.[187]

Yet the strongest current conclusion is not that manual annotation has become obsolete, but that its role is changing. Human labels remain indispensable at three stages. First, they still anchor benchmark creation and allow one to estimate the residual domain gap. Second, they expose ontology problems, such as boundary ambiguity, missed weak-contrast objects, and disagreements over what counts as a valid object or defect. Third, they are often needed for failure analysis, especially when the model seems correct numerically but wrong scientifically. The radiation-cavity study is highly revealing here: the synthetic and experimental models performed similarly, but both also inherited problems from ground-truth uncertainty and partial over-labeling, underscoring that the real bottleneck is often not only data volume but label validity itself.[187]

Accordingly, the more realistic near-term position is that synthetic data can replace large fractions of repetitive annotation while shifting human effort toward curation, ontology design, targeted auditing, and challenge-set construction. This is already implicit in pretraining/fine-tuning strategies where synthetic data provide coverage and real labels provide refinement.[193] It is also implicit in generative cryo-EM pipelines, where synthetic micrographs support particle picking or pose estimation, but final structural claims still require validation against experimental reconstructions and biological plausibility.[65,68,145]

13.3. Transferability across instruments, laboratories, and sample classes

Transferability remains perhaps the most practically consequential unresolved question. Electron microscopy is never a single domain: even nominally similar datasets can differ because of instrument optics, detector response, dwell time, scan strategy, sample thickness, preparation route, contamination, charging, or operator choices. This is why methods that generalize across magnifications, imaging modes, and material systems are especially noteworthy.[186] EMcopilot demonstrated zero-shot transfer across several catalyst systems and imaging conditions, suggesting that some kinds of morphology-aware and artifact-aware generation do capture reusable priors. Likewise, DiffRenderGAN emphasized modality-flexible synthesis across electron and ion microscopy settings.[185]

Still, these successes should not be mistaken for universal portability. Many studies achieve strong transfer only after carefully restricting the morphology space or after adapting the synthetic generator to the specific target distribution.[71,73,81] The catalyst-particle literature illustrates the point sharply: a generator that works well for bright, isolated particles on relatively uniform backgrounds is not automatically adequate for clustered particles, sub-nanometer species, nonuniform supports, or mixed imaging modes.[82,186] In cryo-EM, the same issue appears as protein-family transfer, preferred orientation, or the gap between purified particles and *in situ* environments.[68,145]

Recent 4D-STEM polarization benchmarking makes this limitation explicit: strong performance on simulated diffraction patterns did not guarantee stable experimental classification, and the domain gap could only be partially mitigated by augmentation, filtering, and alternative representation strategies.[174] This is a direct reminder that transferability must be treated as an experimentally tested property of the workflow, not as an assumed consequence of physically generated training data.

A deeper problem is that transferability is usually judged after training, not designed before training. The field still lacks a principled rule for deciding when two laboratories or microscopes belong to the same domain for a given task. Progress on out-of-distribution analysis in HRTEM is therefore especially important: by relating generalization loss to contrast transfer functions, this work suggests that transferability can be expressed in terms of information overlap rather than superficial appearance (see Fig. 25).[191] This is a substantial conceptual advance, but it currently applies most cleanly to controlled imaging-condition shifts. General transfer across sample classes, chemistries, or emergent morphologies remains much less understood.

13.4. Generalization under changing imaging conditions

Even within a single laboratory and a single application, imaging conditions drift over time. Defocus changes, alignment drifts, dose budgets shift, detectors age, contamination accumulates, and sample preparation varies. For this reason, generalization under changing conditions is not a secondary robustness problem; it is the everyday operating condition of EM analysis. Several of the papers in this corpus confront this directly. GAN-refined STEM images were proposed precisely because simulated data alone cannot track the dynamically changing appearance of automated microscopy streams.[81] Domain-adaptation layers in EMcopilot serve the same purpose by injecting noise, drift, and contrast perturbations.[186] Noise classification studies in SEM remind us that not all degradation is generic Gaussian noise; different noise families may demand different restoration or inference strategies.[58]

The HRTEM OOD study again provides a rare quantitative perspective by showing that performance can degrade smoothly and predictably as defocus and related imaging parameters move away from the training distribution.[191] This is encouraging because it suggests that generalization failures need not always be abrupt or mysterious. But it also implies that one must characterize the operational envelope explicitly. In other words, synthetic data do not abolish the need for calibration; they make calibration more analyzable.

A parallel issue concerns representation learning. If a model has learned microscopy-specific causal features, modest shifts in acquisition settings should not catastrophically alter its predictions. If, however, it relies on shallow shortcuts tied to brightness, local texture, or artifact co-occurrence, then even visually minor perturbations may cause failure. This distinction is central to current unresolved questions because generative pipelines can inadvertently stabilize exactly the wrong cues. The more powerful the generator becomes, the more important it is to ensure that the learner is not rewarded for exploiting synthetic regularities that are absent, unstable, or scientifically meaningless in real experiments.[64,185]

13.5. Standard benchmarks *vs.* application-specific workflows

Benchmarking is essential, but benchmark success is not identical to scientific utility. Standardized datasets enable

reproducibility, comparison, and cumulative progress, as seen in cryo-EM particle-picking collections and benchmark micrograph sets.[145] Yet the field's most consequential deployments are often highly application-specific: catalyst particle sizing on one support, radiation-cavity counting in one alloy family, mineral-phase classification for a forensic workflow, or fiber measurement under a narrow industrial protocol.[82,170,187]

This tension matters because synthetic data can make a benchmark look solved while leaving the real workflow underconstrained. A benchmark typically freezes the label ontology, target morphology space, and imaging conditions. Real practice rarely does. It is therefore possible to optimize a generator for benchmark alignment while neglecting rare events, unknown unknowns, or edge cases that dominate scientific risk. The problem is particularly acute in inverse and predictive analysis, where the output is not only a mask or detection but a measurement or mechanistic claim.[82,197]

A productive way forward is to treat benchmarks as tier-1 tests rather than final certification. Tier 1 should cover canonical public tasks. Tier 2 should contain laboratory-specific challenge sets. Tier 3 should evaluate whether downstream scientific conclusions are stable under synthetic-perturbation analysis, annotation uncertainty, and instrument shifts. The literature already contains the ingredients of this framework, but they remain scattered across communities rather than integrated into a shared validation doctrine.[187,191]

13.6. Interpretability and scientific trust in generative pipelines

Interpretability is not an optional add-on in scientific microscopy. When a segmentation, particle pick, size distribution, or reconstruction feeds into mechanistic inference, synthesis design, or structural biology, users need to know why the model should be believed. The problem is amplified in generative pipelines, where one network may first synthesize a training distribution and a second network then learn from it. This creates an additional epistemic layer between experiment and conclusion. The domain-aware cavity study addresses this concern pragmatically by converting per-object confidences into an image-level confidence score, effectively teaching the model to declare when it may be outside its safe operating domain (see Fig. 26).[187] EMcopilot adds saliency-map analysis to compare decision attention between self-supervised and manual-label baselines.[186]

But these are first steps. Trust in generative EM will require more than post-hoc visualization. It will require traceable generation graphs, explicit morphology priors, documented artifact models, sensitivity analysis to generator parameters, and challenge tests that search for hallucinated realism. This concern is especially acute in cryo-EM and cryo-ET, where generative models can improve particle picking or *ab initio* inference, yet the structural output may be interpreted biologically or even pharmacologically.[65,68,145] Integrative and Bayesian approaches remind us that uncertainty weighting and cross-modal consistency remain indispensable counterbalances to black-box confidence.[197]

For electron microscopy more broadly, scientific trust will likely emerge from hybrid rather than purely end-to-end paradigms. Physics-informed generators, morphology-aware priors, explicit domain scores, and downstream uncertainty estimates together form a more credible path than unconstrained image synthesis alone.[87,185] This also explains why the unresolved questions in this section connect naturally to the conclusion of the review: the future of synthetic data in EM depends not only on better generation, but on better accountability.

Taken together, the unresolved questions are not signs of failure; they are signs that synthetic-data research in electron microscopy has matured from proof-of-concept image generation to the more difficult problem of scientific deployment. The literature now shows that synthetic data can rival manual labels, can expose and sometimes reduce human annotation bias, and can materially expand what is feasible in high-throughput EM analysis.[71,186,187] But this same literature also shows that robust use requires explicit attention to morphology priors, artifact modeling, transfer boundaries, OOD behavior, and model self-awareness.[64,185,191] Once the promises, limits, and open questions have been clarified, the remaining task is to synthesize a forward-looking perspective on where the field is heading and what conditions must be met for synthetic data to become an ordinary, trusted part of electron-microscopy practice.

14. Outlook

If the earlier sections of this review established why synthetic data are needed, how they can be generated, where they already work, and where they still fail, the outlook must address a larger shift that is now becoming visible. Synthetic data in electron microscopy are beginning to extend beyond the narrow role of substitute training images. Their future importance will lie less in the ability to populate benchmarks and more in the ability to organize the relation between specimen, microscope, model, and scientific question. This shift matters because the central bottleneck in electron microscopy is rarely image scarcity alone. More often, it is the mismatch between what the instrument records and what the investigator actually wants to know: latent structure, three-dimensional organization, hidden defects, dynamic evolution, functional performance, or the best next experiment to run.

Seen from this perspective, the future of synthetic data is not simply 'more realistic images'. It is a move toward computationally explicit experiments in which image generation, inference, validation, uncertainty, and decision-making are linked in one workflow. The most successful developments reviewed in this manuscript already point in that direction. Across SEM, TEM/STEM, cryo-EM, and volume EM, synthetic data are becoming a medium through which hidden variables are made accessible, rare conditions are made testable, and alternative measurement strategies are explored before they are attempted on the instrument itself. The outlook therefore centers on three connected transitions: from synthetic images to synthetic experiments, from passive analysis to predictive and autonomous microscopy, and from qualitative image interpretation to rigorously quantitative electron microscopy.

14.1. From synthetic images to synthetic experiments

The next conceptual step for the field is to stop treating synthetic data as isolated image products and to begin treating them as components of synthetic experiments. A synthetic image is a snapshot. A synthetic experiment is a structured computational episode: a specimen model, a distribution of latent states, a microscope model, an acquisition protocol, a reconstruction or analysis chain, and a defined scientific objective. This distinction is crucial because many of the hardest problems in electron

microscopy do not arise at the level of one image. They arise from how images are acquired, how information is lost, how noise accumulates, how decisions are made during analysis, and how uncertainty propagates through the pipeline.

In that broader setting, synthetic data can do far more than provide additional training examples. They can help formalize counterfactual questions. What would this sample look like if the defocus distribution were different? How would the segmentation fail if charging increased? Which conformational states would become indistinguishable if the dose were reduced? Which scan path, tilt scheme, reconstruction prior, or detector setting would preserve the quantity of interest most effectively? These are not questions about image realism in the narrow visual sense. They are questions about experiment design, identifiability, and robustness. Once synthetic generation is embedded into that level of reasoning, it begins to function as an experimental planning instrument rather than only as a machine-learning convenience.

This transition is especially natural in electron microscopy because the field has always depended on forward models, implicit or explicit, to interpret what is observed. Synthetic experiments simply make that dependence more operational. In cryo-EM, for example, one increasingly wants to know not only whether a reconstruction pipeline can recover a known structure, but under which heterogeneity regime, orientation distribution, noise level, and contrast-transfer setting it remains reliable. In volume EM, the crucial problem may be not whether one synthetic slice looks plausible, but whether a simulated acquisition reproduces the continuity, anisotropy, deformation, and artifact structure that determine downstream reconstruction quality. In SEM and TEM/STEM, the decisive question may be whether a synthetic pipeline preserves the causal chain from morphology or atomic configuration to image contrast and then to the inferred measurement. In all of these cases, the relevant unit is already larger than a single image. The outlook merely makes that explicit.

One important consequence is that benchmark culture will have to evolve. Static benchmark datasets remain valuable, but they will be increasingly insufficient for the most interesting scientific uses of synthetic data. Future benchmarks will likely need to include controllable latent variables, parameter sweeps, acquisition metadata, failure regimes, and uncertainty annotations. In other words, they will need to behave more like compact experimental worlds than like fixed image collections. This would allow the community to ask stronger questions: not simply which model scores highest, but which model remains stable when the microscope changes, when artifacts intensify, when morphology shifts, or when the true target variable becomes only weakly observable. Such benchmark ecosystems would be more demanding to create, but they would also be more scientifically meaningful.

A second consequence is that synthetic experiments may become central to instrument-aware method development. Instead of learning only from historically collected images, a model could be developed within a loop that explicitly simulates acquisition choices and their consequences. The synthetic layer would then support decisions about sampling density, scan strategy, tilt schedule, dose allocation, reconstruction settings, or annotation priorities. This is the point at which synthetic data cease to be passive assets and become active design variables. The long-term implication is profound: synthetic data may eventually serve as the sandbox in which new microscopy workflows are invented before they are executed on real specimens.

14.2. From passive analysis to predictive and autonomous microscopy

A second major transition concerns the role of image analysis itself. Much of present-day microscopy AI remains reactive. The instrument acquires data, and the algorithm interprets them afterward. Synthetic data have already improved this passive analysis stage by enabling better segmentation, denoising, restoration, registration, and metrology. But the more transformative possibility is that synthetic data will help microscopy become predictive rather than merely descriptive.

Predictive microscopy means that the system does not only label what is already visible. It estimates what is hidden, what is likely to happen next, what measurement would be most informative, and which parts of its own output should not yet be trusted. This is where the deeper logic of synthetic data becomes apparent. Because synthetic pipelines can connect image space to latent variables, they provide the training environment for inference that reaches beyond direct observation. In one setting, that latent variable may be particle pose, three-dimensional depth, composition, topography, or pore connectivity. In another, it may be a conformational coordinate, a defect population, a transport property, a reconstruction hyperparameter, or an expected measurement error. The common thread is that the model learns not just to read images but to reason over an explicit hypothesis space defined by synthetic variation.

Once this inferential capacity is coupled to uncertainty, prediction naturally leads toward autonomy. An autonomous microscopy system is not simply one that automates a task quickly. It is one that can decide, within limits, what additional data are needed, where a current model is extrapolating too far, and how the next acquisition step should be adjusted to improve the scientific outcome. Synthetic data are especially important here because real experiments rarely provide enough diverse, labeled, and risk-free experience for such systems to be trained directly on reality alone. A synthetic environment can expose the system to rare defects, unusual morphologies, failure modes, and hypothetical acquisition choices that would be expensive, destructive, or impractical to collect systematically in the laboratory.

This does not imply that the future microscope will become independent of the human expert. On the contrary, the more autonomous the workflow becomes, the more important it will be to define the human role correctly. The expert will shift from annotator and parameter tweaker toward supervisor of goals, priors, and trust boundaries. Synthetic data will help enable that transition because they can move routine supervision into the model-development stage while leaving conceptual oversight to the scientist. The human will increasingly judge whether the model's operating assumptions, confidence patterns, and chosen measurements remain scientifically reasonable. In this sense, autonomy in microscopy should not be imagined as the disappearance of expertise, but as the relocation of expertise to a higher level of experimental governance.

There is also an important epistemic shift here. Passive analysis often invites the illusion that the image is the final object of study. Predictive and autonomous microscopy instead treat the image as one partial observation within a larger inferential process. Synthetic data reinforce this view by making explicit the model class under which interpretation occurs. A digitally assisted microscope that proposes the next best measurement, adapts its acquisition policy, or warns that a given region is out of distribution is not only faster than a conventional workflow. It is also more explicit about what it assumes and

what it cannot yet resolve. Properly developed, this could improve scientific discipline rather than weaken it.

The caution, however, is equally clear. A predictive system is only as trustworthy as the synthetic world on which its reasoning has been formed. If the synthetic pipeline omits an important failure mode, encodes the wrong morphology distribution, or confuses visual plausibility with causal fidelity, autonomy can amplify error rather than reduce it. For that reason, the path from passive analysis to autonomy must pass through the validation and design principles discussed earlier in this review. The future will favor systems that are not only adaptive, but also uncertainty-aware, instrument-aware, and explicit about the conditions under which their predictions remain valid.

14.3. The future role of synthetic data in quantitative electron microscopy

The most enduring impact of synthetic data may ultimately be on quantitative electron microscopy. For decades, many of the strongest applications of EM have depended on expert qualitative interpretation supplemented by selected measurements. Synthetic data are now helping to change the balance. They make it increasingly plausible to treat EM images as inputs for calibrated inference about quantities that are difficult or impossible to measure directly: particle-size distributions under overlap, surface topography from two-dimensional views, three-dimensional structure from limited projections, elemental or phase information from contrast patterns, and functional or process-related variables inferred from morphology.

The key point is that quantitative EM does not require perfect realism everywhere. It requires controlled relations between latent variables, acquisition conditions, and observable image statistics. That is why the most important future role of synthetic data is likely to be as a metrological interface. They can define what should be measurable, under what assumptions it is measurable, what bias and variance should be expected, and how uncertainty should be reported when those assumptions begin to fail. In this form, synthetic data become part of the measurement model itself. They are no longer peripheral training supplements but operational components of calibration, validation, and uncertainty quantification.

This development could help resolve a longstanding tension in microscopy between visual authority and quantitative authority. Electron micrographs are often compelling images, but compelling appearance is not equivalent to calibrated evidence. Synthetic-data-driven pipelines offer a way to make that distinction sharper. A model may be able to produce a segmentation, a reconstruction, or an inferred property field, but the more important question is whether it can attach those outputs to a known validity regime. The mature future of synthetic data in EM will therefore not be defined by image generation alone, but by traceable quantitative claims: this estimate is reliable within this range of morphology, this dose regime, this detector condition, this domain gap, and this uncertainty threshold.

Another important consequence is that quantitative electron microscopy will become more explicitly multimodal and model-linked. Many scientific questions do not end at the microscope image. They concern reactivity, transport, mechanics, function, or biological state. Synthetic pipelines can help connect these downstream properties to image-derived structural descriptors in a disciplined way, provided that the intermediate representation is physically meaningful. This is one of the reasons why digital-twin thinking is so attractive. A true digital twin of an EM workflow would not only synthesize images; it would couple structure generation, microscope physics, reconstruction, inverse inference, and property prediction within one continuously updated framework. Even if full digital twins remain aspirational, partial versions of this logic are already beginning to appear across the literature surveyed in this review.

The broader significance is that synthetic data may help EM evolve from a largely interpretive science into a more computationally testable one. Instead of asking only what the micrograph seems to show, one can ask which family of hidden states could have produced it, which additional experiment would discriminate among them, and which quantitative conclusion survives realistic perturbation of the imaging process. This is a more demanding scientific standard, but it is also a more powerful one. It moves the field closer to falsifiable, uncertainty-aware, and mechanism-linked image analysis.

For that future to be realized, several conditions will be essential. Synthetic generators must remain tied to scientific questions rather than to generic visual benchmarks. Metadata, uncertainty, and morphology distributions must be treated as first-class components of dataset design. Community benchmarks will need to reward transferability, calibration, and robustness instead of superficial realism alone. And perhaps most importantly, the field must resist the temptation to treat synthetic data as a shortcut around experimental difficulty. Their real value lies elsewhere. They do not abolish the complexity of electron microscopy; they make that complexity computationally negotiable.

Taken together, the outlook is both ambitious and disciplined. Synthetic data in electron microscopy are unlikely to replace experiment, human expertise, or physical understanding. But they are increasingly poised to reorganize how all three interact. They will help convert images into experiments, analysis into prediction, and visual interpretation into quantitative inference with explicit uncertainty. If this transition is handled carefully, synthetic data will not simply make electron microscopy more automated. They will make it more testable, more transferable, and ultimately more scientific.

15. Summary tables and practical reference guide

To complement the section-by-section discussion, the following tables condense the main logic of the review into a compact reference format. They are intended to help readers move quickly from general concepts to practical decisions: which synthetic-data strategies fit each EM modality, which paradigms best support specific downstream tasks, how validity should be judged, and where the most important methodological gaps still remain.

Read together, the Tables 2–6 also make visible a central conclusion of the review: useful synthetic data in electron microscopy are not defined by a single ideal recipe, but by task-aware combinations of image-formation realism, structural priors, annotation design, validation strategy, and uncertainty control.

15.1. Synthetic-data categories

The field of synthetic data in electron microscopy encompasses a diverse set of generation strategies, each with distinct assumptions, resource requirements, and realism bottlenecks. Table 2 provides a consolidated reference that defines the major

Table 2. Synthetic-data generation strategies in electron microscopy: definitions, primary strengths, and key limitations.

Category	Definition	Primary strengths	Key limitations
Fully simulated (physics-based)	Forward models of electron-matter interaction (multislice, Monte Carlo, Contrast Transfer Function)	Full control over imaging parameters; Exact ground-truth labels; Systematic ablation studies	Computational cost can be high for high-fidelity simulations (hours per image); Incomplete modeling of artifacts (charging, drift); Persistent domain gap
Geometry or microstructure-driven	3D structural models (particles, fibers, defects) rendered into images	Fast generation; Full control over morphology distributions; Exact 3D geometry known	Simplified contrast; May not capture experimental noise and detector effects Requires validation against real distributions
Semi-synthetic (real backgrounds + inserted features)	Real micrographs as backgrounds with simulated objects inserted digitally	Realistic background texture and noise; Precise labels; Moderate cost	Does not model object-background interactions; Limited to insertable objects; Background selection may introduce bias
Augmentation-based	Geometric or photometric transforms (rotation, flipping, scaling, noise) applied to existing images	Computationally cheap; Easy to implement; Improves model robustness	Does not create new physical content; Cannot introduce new object types or morphologies; Limited domain-gap bridging
Generative AI (GANs, diffusion, VAEs)	Deep generative models trained on real EM distributions, conditioned on masks or priors	High visual realism; Diverse morphologies and contrast; No physics simulation required	Risk of hallucinations; Requires real training data; Black-box; computationally intensive; Potential mode collapse
Hybrid physics-informed generative	Generative models with physical constraints (differentiable simulators, forward models) in the training loop	Combines physical fidelity with realism; May improve extrapolation beyond training data; Physically consistent outputs	Complex implementation; Requires differentiable simulators (rare); High computational cost; Still under development
Cryo-EM specific (particle + noise models)	Projections of known 3D structures at random orientations with realistic noise	Ground-truth particle positions and orientations; Realistic SNR; Benchmarks picking and alignment	Limited to known structures; Fixed-structure simulators does not capture conformational variability without ensemble models; Requires careful noise modeling
Volume EM and 3D synthetic (FIB-SEM, tomography)	3D volumes simulating serial sectioning, slicing artifacts, or tomographic projections	Full 3D ground truth; Systematic study of segmentation and reconstruction; Simulates realistic artifacts	Computationally expensive; Requires realistic 3D models; Difficult to validate against experimental volumes

categories discussed throughout this review, summarizes their primary strengths and limitations, and offers a practical starting point for researchers choosing an approach for a specific application.

15.2. Electron-microscopy modalities *vs.* synthetic-data strategies

Table 3 maps the major EM modalities to the synthetic-data strategies that are usually most effective. The main message is that modality differences matter at every level. Surface-sensitive SEM, projection-based TEM/STEM, low-dose cryogenic methods, and volumetric workflows do not require the same balance of physics-based simulation, geometric modeling, generative translation, or semi-synthetic augmentation.

15.3. Downstream tasks *vs.* suitable synthetic-data paradigms

Table 4 reorganizes the field by scientific task rather than by instrument. This view is useful because the same downstream objective often recurs across modalities, but the synthetic-data paradigm that works best depends on whether the task requires pixel-level realism, rare-event coverage, physically calibrated labels, latent-variable inference, or uncertainty-aware control.

15.4. Validation criteria and characteristic failure modes

Table 5 emphasizes that validation of synthetic data cannot be reduced to visual inspection. A useful synthetic dataset must be tested at the level at which scientific claims are made: pixel distributions, object statistics, latent-variable recovery, task transfer, uncertainty calibration, and robustness under acquisition changes. Failure to distinguish these layers is one of the main reasons why apparently realistic synthetic datasets can still support weak science.

15.5. Open challenges and methodological directions

Table 6 distills the open problems discussed across the review into a forward-looking agenda. Several themes recur: realism must be judged relative to the scientific task; transferability requires explicit handling of instrument and laboratory variability; benchmarks should test mechanisms rather than only appearance; and future progress will depend on uncertainty-aware links between synthetic data, experiment, and increasingly predictive microscopy workflows.

Taken together, these tables do not replace the detailed discussion in the main text, but they provide a compact decision framework for readers approaching the field from different

Table 3. Electron-microscopy modalities *vs.* the synthetic-data strategies that are typically most useful, together with their dominant strengths and main realism bottlenecks.

EM modality	Most suitable synthetic-data strategies	What synthetic data are especially good for	Main realism bottlenecks and cautions
SEM	Geometry-driven generators; morphology-aware rendering; physics-light contrast/noise simulation; semi-synthetic compositing; domain adaptation	Particle and pore morphology, surface textures, defect detection, segmentation, counting, metrology, dataset balancing	Surface charging, detector-specific contrast, topographic shadowing, preparation artifacts, and over-smooth synthetic textures can distort transferability
TEM/STEM	Physics-driven forward models; multislice or surrogate simulators; atomistic/structural generators; semi-synthetic corruption models; generative refinement	Atom localization, defect identification, phase discrimination, denoising, super-resolution, inversion, thickness/orientation-sensitive studies	Projection ambiguity, strong dependence on thickness and alignment, beam-sample interactions, scan distortions, and mismatch between visual plausibility and quantitative fidelity
Cryo-EM/ cryo-ET	High-fidelity forward simulation of particles and optics; conformational ensemble generation; noise translation; pseudo-labeling; simulation-based inference	Particle picking, orientation estimation, denoising, tomographic reconstruction, template matching, benchmarking of low-SNR workflows	Extremely low signal-to-noise ratio, ice and contamination effects, preferred orientations, missing wedge, conformational heterogeneity, and risk of hallucinated structure
FIB-SEM and volume EM	3D or 4D virtual volumes; slicing-aware simulators; anisotropic-resolution models; registration artifact synthesis; mask-rich synthetic annotation	Volume segmentation, organelle tracing, connectomics-style reconstruction, registration, interpolation, section-loss robustness, rare-object enrichment	Sectioning artifacts, misalignment, anisotropic voxels, staining variability, long-range context requirements, and very high annotation complexity
Correlative/ multimodal EM	Cross-modal paired synthesis; latent-space alignment; domain translation with physical constraints; synthetic metadata fusion	Registration, weakly supervised correspondence, transfer learning across modalities, multimodal feature discovery	Misleading one-to-one translations, inconsistent spatial scales, non-identical fields of view, and overconfident cross-modal labels

Table 4. Major downstream tasks in electron microscopy and the synthetic-data paradigms that are usually most suitable for each task.

Downstream task	Best-matched synthetic-data paradigm	Why this paradigm is useful	Typical failure mode if used uncritically
Segmentation and semantic labeling	Semi-synthetic masks; geometry-aware fully synthetic images; generative augmentation; pseudo-labeling	Provides dense labels, class balancing, and controlled morphology variation where manual annotation is slow or inconsistent	Models learn simplified edges or textures and fail on preparation artifacts, weak boundaries, or previously unseen classes
Detection and counting	Object-centric synthetic placement; compositing; rare-event oversampling; task-aware domain randomization	Enables large controlled populations, overlap scenarios, and label-perfect object coordinates for training and benchmarking	Counts remain accurate only on synthetic object separations; performance collapses when crowding, clutter, or contrast shifts differ experimentally
Denoising and restoration	Paired simulation of clean/noisy data; corruption models; self-supervised pretraining with synthetic priors	Allows known target images, controlled noise regimes, and stress-testing across signal levels unavailable experimentally	Networks remove scientifically meaningful weak features together with noise, producing visually attractive but distorted outputs
Super-resolution and reconstruction	Physics-based forward simulation; differentiable rendering; synthetic low-/high-resolution pairs	Supports inversion tasks by linking latent structure to degraded observations under known acquisition conditions	Apparent detail may be hallucinated when the forward model is incomplete or when synthetic priors are too narrow
Registration and alignment	Synthetic transformations with known ground-truth displacement; section-loss and drift simulation	Useful for benchmarking alignment robustness, uncertainty, and failure thresholds in multi-frame or multi-slice workflows	Training may underestimate non-rigid distortions, specimen damage, or correlated artifacts encountered in real datasets
Tomographic inference and 3D recovery	Volumetric simulators; missing-wedge-aware generation; synthetic projection series; simulation-based inference	Creates fully known 3D truth for evaluating reconstructions, segmentation in volume, and pose or orientation estimation	Synthetic volumes may encode unrealistic morphology priors, leading to overconfident recovery of structures absent in experiment
Metrology and quantitative property inference	Calibrated physics-driven synthetic experiments; parametric sample generators; uncertainty-tagged synthetic datasets	Links image statistics to measurable latent quantities such as size, thickness, porosity, composition, or defect density	Quantitative predictions drift if calibration is weak, metadata are incomplete, or nuisance variables are omitted from synthesis
Representation learning and pretraining	Large diverse synthetic datasets; self-supervised tasks; cross-domain latent alignment	Improves transferable feature learning even when final labels are scarce or task definitions evolve later	Representations become biased toward synthetic regularities and may not cover rare but scientifically important real structures
Active learning and autonomous microscopy	On-demand synthetic generation; adaptive simulation; digital-twin-like closed loops	Helps choose informative acquisitions, estimate counterfactuals, and train controllers under controlled uncertainty	Closed-loop systems can become confidently wrong when the synthetic world omits hidden failure modes or instrument drift

Table 5. Validation criteria for synthetic data in electron microscopy and the characteristic failure modes that each criterion is intended to detect.

Validation criterion	What should be examined	Why it matters scientifically	Characteristic failure mode
Visual plausibility	Noise, contrast, texture, artifacts, and overall appearance relative to held-out experimental data	Useful as an initial sanity check, especially for detecting obviously unrealistic synthesis	Images look convincing to human observers but remain scientifically wrong for the target task
Distributional similarity	Intensity histograms, frequency content, morphology statistics, object-size distributions, spatial correlations, class balance	Checks whether synthetic data occupy approximately the right statistical regime	Global distributions match while task-critical local structures, rare cases, or context relations are missing
Task-level sim-to-real transfer	Performance on fully held-out experimental data using the intended downstream metric	Most direct test of practical usefulness for detection, segmentation, denoising, metrology, or reconstruction	High synthetic benchmark scores but weak transfer to real images from other instruments, sessions, or laboratories
Physics consistency	Agreement with known acquisition physics, imaging geometry, dose effects, detector behavior, and structural constraints	Essential when the goal is quantitative inference rather than only visual augmentation	Model exploits shortcuts unrelated to electron-matter physics and fails outside a narrow acquisition window
Ground-truth adequacy	Whether the provided synthetic labels correspond to the scientific quantity actually being inferred	Prevents confusion between annotation convenience and meaningful truth	Perfect labels for the wrong target variable create deceptively strong model performance
Robustness and stress testing	Performance under changes in noise, blur, dose, defocus, thickness, drift, section loss, contamination, or sample class	Reveals whether the synthetic pipeline teaches stable relationships or only narrow operating conditions	A model performs well only near the synthetic training manifold and degrades sharply under realistic perturbation
Uncertainty calibration	Confidence estimates, error bars, failure detection, and abstention behavior on real data	Necessary for trustworthy deployment in quantitative and semi-autonomous workflows	Predictions remain overconfident when the model encounters out-of-domain data or rare structures
Reproducibility and documentation	Transparent reporting of generators, priors, metadata, preprocessing, splits, and evaluation protocol	Allows the field to compare benchmarks, identify hidden assumptions, and build cumulative knowledge	Apparent gains cannot be reproduced because the generative assumptions or validation details were not disclosed

angles. A microscopist can identify which synthetic strategy is most plausible for a given modality; a method developer can locate the paradigm suited to a specific task; a reviewer or benchmark designer can see which validation layers are necessary; and a forward-looking reader can connect present capabilities to the field's most consequential unresolved questions.

16. Conclusion

Synthetic data have become an important methodological component of electron microscopy because they enable controlled training, validation, and inference when experimental ground truth is incomplete or inaccessible. Across the diverse landscape surveyed in this review — from SEM and TEM/STEM to cryo-EM, cryo-ET, FIB-SEM, and other volume-imaging workflows — the same conclusion emerges in different forms: synthetic data matter not merely because they enlarge datasets, but because they allow the field to formalize, control, and interrogate the relation between specimen structure, image formation, acquisition conditions, and computational inference. In a domain where experimental ground truth is often incomplete, ambiguous, destructive to obtain, or physically inaccessible, synthetic data provide a way to make assumptions explicit and testable. Their real value therefore lies not in imitation alone, but in structured translation between physical reality and algorithmic reasoning.

The review has shown that this translation cannot be reduced to one universal recipe. Electron microscopy is an unusually demanding setting for synthetic-data methodology because signal formation is shaped by multiple coupled variables: electron-matter interactions, projection effects, contrast transfer, detector response, dose constraints, beam damage, charging, drift, reconstruction artifacts, and specimen-dependent sources of variation. As a result, synthetic data in EM must always be judged relative to purpose. The most useful synthetic dataset is not necessarily the one that looks most realistic to a human observer; it is the one that preserves the task-relevant relations needed for a model, a validation protocol, or an inverse problem to remain scientifically meaningful. This shift from visual plausibility to functional adequacy is one of the most important conceptual outcomes of the field and one of the strongest themes running through this review.

A second major conclusion is that the landscape of synthetic data in electron microscopy is now best understood as a continuum rather than a binary division between 'real' and 'simulated'. Physics-driven forward models, geometry-based rendering, semi-synthetic composites, augmentation pipelines, generative models, weak labels, and hybrid workflows each occupy different positions along that continuum. Their strengths and vulnerabilities are complementary. Forward models offer interpretability and parameter control; generative approaches offer flexibility and scale; semi-synthetic strategies often provide the most practical route when full physical realism is unattainable. The most effective future workflows are therefore unlikely to depend on any single class of method. Instead, they will combine physically informed simulation, experimentally anchored calibration, uncertainty-aware learning, and carefully designed transfer across instruments, laboratories, and specimen classes.

Table 6. Open challenges for synthetic data in electron microscopy and suggested methodological directions for addressing them.

Open challenge	Why it remains difficult	Suggested methodological direction	Expected payoff if addressed well
How realistic is realistic enough?	Different tasks depend on different image attributes; maximal photorealism may be unnecessary or even inefficient	Define task-specific realism criteria and validate at pixel, object, and inference levels rather than by appearance alone	More efficient dataset design and clearer scientific justification for synthesis choices
Can synthetic data replace manual annotation, or only reduce it?	Many EM labels are partial, subjective, or indirect rather than fully observable truths	Use synthetic data to complement expert annotation: pretraining, rare-case enrichment, uncertainty estimation, and targeted expert correction	Lower labeling burden without pretending that all scientific truth can be annotated directly
Transferability across instruments, laboratories, and sample classes	Domain shifts arise from optics, detectors, preparation, protocols, and sample heterogeneity	Combine domain adaptation, metadata-aware modeling, broader acquisition randomization, and cross-site benchmark design	More reliable real-world deployment and less benchmark overfitting
Generalization under changing imaging conditions	Noise, dose, thickness, drift, contamination, and alignment conditions vary continuously in practice	Train with stress-tested synthetic perturbations and evaluate robustness envelopes rather than one fixed operating point	Models that remain useful when acquisition settings drift or must be optimized adaptively
Benchmark design for scientific rather than purely visual success	Conventional benchmarks often reward easy labels or narrow metrics disconnected from scientific inference	Create benchmarks with latent-variable recovery, uncertainty, rare events, and mechanism-linked evaluation	A shift from visually impressive results to scientifically meaningful comparison
Interpretability and scientific trust	Generative and predictive pipelines can hide shortcut learning, hallucinated detail, or poorly calibrated confidence	Use uncertainty-aware outputs, interpretable intermediate variables, counterfactual tests, and failure disclosure protocols	Greater acceptance of synthetic-data methods in high-stakes quantitative workflows
Integration with predictive and autonomous microscopy	Closed-loop systems need fast, adaptive, and trustworthy synthetic models linked to real instruments	Develop synthetic experiments, surrogate simulators, and digital-twin-like frameworks coupled to active decision-making	Microscopy workflows that move from passive analysis toward controlled, adaptive inference
Long-term data stewardship and reproducibility	Synthetic datasets can become opaque when code, priors, metadata, or preprocessing are not preserved	Adopt release standards covering generators, parameter ranges, labels, metadata schema, validation splits, and versioning	Reusable community resources rather than one-off project-specific datasets

The evidence assembled across application areas suggests that synthetic data are moving beyond a niche role in segmentation or classifier pretraining. They now participate in denoising, super-resolution, localization, tomographic reconstruction, defect analysis, particle picking, metrology, registration, hypothesis testing, and the broader design of benchmark environments. In this wider role, synthetic data become part of the experimental logic of microscopy itself. They can be used to stress-test algorithms under rare conditions, probe failure modes that may scarcely appear in curated real datasets, quantify sensitivity to acquisition settings, and create controlled environments in which competing methods can be compared on equal footing. This is especially important for quantitative electron microscopy, where the central question is not whether an image appears convincing, but whether a measurement, classification, or inferred structure remains reliable when the imaging process changes.

At the same time, the review makes equally clear that synthetic data are not a shortcut around rigor. Poorly matched synthetic distributions can generate false confidence, conceal bias, and reward models for exploiting artifacts that are absent (or differently expressed) in real experiments. For that reason, validation is not an accessory stage added after dataset generation; it is the discipline that determines whether synthetic data deserve scientific trust. The design principles discussed in this review converge on a common message: realism must be decomposed into measurable components, benchmarks must be tied to application context, uncertainty must be reported rather than hidden, and transferability must be tested rather than assumed. In this sense, synthetic-data research in EM is inseparable from a broader movement toward reproducible and quantitatively defensible microscopy.

This review has also had deliberate limits. It has focused on synthetic data as they relate to image-centric electron microscopy and associated computational workflows, rather than attempting exhaustive coverage of every adjacent area such as correlative platforms, spectroscopy-rich modalities, or all possible machine-learning paradigms. The literature is evolving extremely rapidly, and many workflows remain highly instrument-specific, task-specific, or available only in partially standardized forms. Consequently, the present review should be read not as a final catalog of established solutions, but as a framework for understanding where synthetic data are already robust, where they remain fragile, and which conceptual questions are likely to govern the next stage of progress.

The central conclusion is that synthetic data will be most useful when treated not as replacements for experiment, but as formal companions to it. They are powerful when they expose assumptions, organize uncertainty, and enable controlled reasoning about what the microscope can and cannot reveal. Used in this way, synthetic data can help move EM from a regime dominated by scarce labels and visually persuasive outputs toward one grounded in calibrated inference, transferable analysis, and progressively more predictive microscopy. That transition will not eliminate the need for expert judgment; it will elevate the standard by which computational microscopy is

judged. The future role of synthetic data in electron microscopy is therefore likely to be more than auxiliary: they are becoming part of how quantitative EM will be designed, validated, interpreted, and ultimately trusted.

Methodology

This review was prepared using TRACE AI co-expert-assisted technology.[198] The review preparation checklist (Table S1) is provided in the Supporting Information file (https://www.uspkhim.ru/RCR5240).

This work was financially supported by the Russian Science Foundation (Grant No. 24-73-10177).

17. List of abbreviations

AI — artificial intelligence;
BPC — Back-Propagation Counting;
CAKED — Class Aggregator for Key Electron-microscopy Data;
CAVIAR — Correlated Atomic Vibration Imaging with sub-Ångstrom Resolution;
CD-SEM — critical-dimension scanning electron microscopy;
CNN — convolutional neural network;
CRISP — CRyo-EM image Segmentation and Processing;
cryo-EM — cryo-electron microscopy;
cryo-ET — cryo-electron tomography;
CTF — contrast transfer function;
CycleGAN — cycle-consistent generative adversarial network;
DCGAN — deep convolutional generative adversarial network;
4D-STEM — four-dimensional scanning transmission electron microscopy;
DISTS — Deep Image Structure and Texture Similarity;
EBSD — electron backscatter diffraction;
EDS — energy-dispersive X-ray spectroscopy;
EM — electron microscopy;
Faster R-CNN — Faster Region-based Convolutional Neural Network;
FIB — focused ion beam;
FIB-SEM — focused ion beam scanning electron microscopy;
FID — Fréchet inception distance;
GAN — generative adversarial network;
HAADF-STEM — high-angle annular dark-field scanning transmission electron microscopy;
HRTEM — high-resolution transmission electron microscopy;
IoU — intersection over union;
KID — Kernel Inception Distance;
LPIPS — Learned Perceptual Image Patch Similarity;
Mask-RCNN — Mask Region-based Convolutional Neural Network;
MSE — mean squared error;
NMF — non-negative matrix factorization;
NP — nanoparticle;
OOD — out of distribution;
PARSED — PARticle SEgmentation Detector;
PERC — profet, EMPIAReader and CAKED;
PSNR — Peak Signal-to-noise Ratio;
PtyRANNOSAUR — Ptychography with Robust Artificial Neural Networks Optimized for Sub-Angstrom Accuracy and Ultrafast Reconstruction;
RMSE — root-mean-square error;
SAM — Segment Anything Model;
SE — secondary electron;
SEM — scanning electron microscopy;
SNR — signal-to-noise ratio;
SSIM — Structural Similarity Index Measure;
STEM — scanning transmission electron microscopy;
STEM-EDS — scanning transmission electron microscopy-energy-dispersive X-ray spectroscopy;
STEM-EELS — scanning transmission electron microscopy with electron energy-loss spectroscopy;
TEM — transmission electron microscopy;
TREM — Template Rotation Expectation Maximization;
U-Net — U-shaped convolutional network;
UTILE — aUTonomous Image anaLysis to accelerate the discovery and integration of energy matErials;
UTILE-Gen — UTILE synthetic dataset generator;
VAE — variational autoencoder;
VQ-VAE — vector-quantized variational autoencoder;
WRGAN — Wave-Reconstruction Generative Adversarial Network;
XRD — X-ray diffraction;
YOLO — You Only Look Once;
YOLOv8 — You Only Look Once, version 8.

18. References

1. M.Botifoll, I.Pinto-Huguet, J.Arbiol. *Nanoscale Horiz*., **7**, 1427 (2022); https://doi.org/10.1039/d2nh00377e
2. S.Muto, M.Shiga. *Microscopy*, **69**, 110 (2020); https://doi.org/10.1093/jmicro/dfz036
3. Z.Cheng, C.Wang, X.Wu, J.Chu. *J. Semicond*., **43**, 081001 (2022); https://doi.org/10.1088/1674-4926/43/8/081001
4. S.V.Kalinin, M.Ziatdinov, J.Hinkle, S.Jesse, A.Ghosh, K.P.Kelley, A.R.Lupini, B.G.Sumpter, R.K.Vasudevan. *ACS Nano*, **15**, 12604 (2021); https://doi.org/10.1021/acsnano.1c02104
5. S.V.Kalinin, D.Mukherjee, K.M.Roccapriore, B.Blaiszik, A.Ghosh, M.Ziatdinov, A.Al-Najjar, C.Doty, S.Akers, N.S.V.Rao, J.Agar, S.R.Spurgeon. *npj Comput. Mater*., **9**, 227 (2023); https://doi.org/10.1038/s41524-023-01142-0
6. R.Sánchez-García, J.Gómez-Blanco, A.Cuervo, J.M.Carazo, C.Ó.S.Sorzano, J.Vargas. *Commun. Biol*., **4**, 874 (2021); https://doi.org/10.1038/s42003-021-02399-1
7. J.He, T.Li, S.Y.Huang. *Nat. Commun*., **14**, 3217 (2023); https://doi.org/10.1038/s41467-023-39031-1
8. A.Dhakal, R.Gyawali, L.Wang, J.Cheng. *Sci. Data*, **10**, 392 (2023); https://doi.org/10.1038/s41597-023-02280-2
9. X.Zhang, T.Zhao, J.Chen, Y.Shen, X.Li. *Nat. Commun*., **13**, 2468 (2022); https://doi.org/10.1038/s41467-022-29994-y
10. G.Liu, T.Niu, M.Qiu, Y.Zhu, F.Sun, G.Yang. *Nat. Commun*., **15**, 2090 (2024); https://doi.org/10.1038/s41467-024-46041-0
11. K.Jamali, L.Käll, R.Zhang, A.Brown, D.Kimanius, S.H.W.Scheres. *Nature*, **628**, 450 (2024); https://doi.org/10.1038/s41586-024-07215-4
12. R.Gyawali, A.Dhakal, J.Cheng. *Commun. Chem*., **8**, 320 (2025); https://doi.org/10.1038/s42004-025-01718-5
13. M.Harastani, G.Patra, C.Kervrann, M.Eltsov. *Nat. Commun*., **16**, 8833 (2025); https://doi.org/10.1038/s41467-025-63895-0
14. D.Vizoso, A.J.Baker, D.L.Medlin, S.D.House, R.Dingreville. *Measurement*, **258**, 119554 (2026); https://doi.org/10.1016/j.measurement.2025.119554
15. Y.Chen, S.Liu, P.Tong, Y.Huang, H.Tian, F.Lin. *Sci. Rep*., **13**, 2631 (2023); https://doi.org/10.1038/s41598-023-29606-9
16. H.Ni, Z.Wu, X.Wu, J.G.Smith, M.J.Zachman, J.Zuo, L.Ju, G.Zhang, M.Chi. *Nano Lett*., **23**, 7442 (2023); https://doi.org/10.1021/acs.nanolett.3c01892
17. S.Kandel, T.Zhou, A.V.Babu, Z.W.Di, X.Li, X.Ma, M.V.Holt, A.Miceli, C.Phatak, M.J.Cherukara. *Nat. Commun*., **14**, 5501 (2023); https://doi.org/10.1038/s41467-023-40339-1

18. H.Liang, C.Wang, H.Yu, D.J.Kirsch, R.Pant, A.McDannald, A.G.Kusne, J.C.Zhao, I.Takeuchi. *Sci. Adv.*, **11**, eadu7426 (2025); https://doi.org/10.1126/sciadv.adu7426
19. J.Yang, R.A.Yin, C.Jiang, Y.Hu, X.Zhu, X.Hu, S.A.Kumar, S.K.Holmes, X.Wang, X.Zhai, K.Rong, Y.Zhu, T.Zhang, Z.Yin, Y.Cao, H.Tang, A.D.Franklin, J.Kong, N.Z.Gong, Z.Ren, H.Wang. *ACS Nano*, **19**, 35493 (2025); https://doi.org/10.1021/acsnano.5c09057
20. U.Pratiush, H.Funakubo, R.K.Vasudevan, S.V.Kalinin, Y.Liu. *Digit. Discov.*, **4**, 252 (2025); https://doi.org/10.1039/d4dd00277f
21. A.S.Galushko, V.P.Ananikov. *Angew. Chem., Int. Ed.*, **65**, e20712 (2026); https://doi.org/10.1002/anie.202520712
22. A.S.Galushko, V.P.Ananikov. *ACS Catal.*, **14**, 161 (2024); https://doi.org/10.1021/acscatal.3c03889
23. D.M.Arkhipova, D.A.Boiko, A.A.Oganov, A.S.Kashin, *E.G.*Gordeev, V.P.Ananikov. *Sci. Data*, **13**, 570 (2026); https://doi.org/10.1038/s41597-026-06941-w
24. D.A.Boiko, D.M.Arkhipova, V.P.Ananikov. *Small*, **20**, 2403423 (2024); https://doi.org/10.1002/smll.202403423
25. S.Lu, B.Montz, T.Emrick, A.Jayaraman. *Digit. Discov.*, **1**, 816 (2022); https://doi.org/10.1039/d2dd00066k
26. L.Rettenberger, N.J.Szymanski, A.Giunto, O.Dartsi, A.Jain, G.Ceder, V.Hagenmeyer, M.Reischl. *npj Comput. Mater.*, **11**, 289 (2025); https://doi.org/10.1038/s41524-025-01802-3
27. A.Khadangi, T.Boudier, V.Rajagopal. *Bioinformatics*, **37**, 97 (2021); https://doi.org/10.1093/bioinformatics/btaa1094
28. A.A.Holicheva, K.S.Kozlov, D.A.Boiko, M.S.Kamanin, D.V.Provotorova, N.I.Kolomoets, V.P.Ananikov. *npj Biofilms Microbiomes*, **11**, 16 (2025); https://doi.org/10.1038/s41522-025-00647-4
29. L.E.Nersesyan, D.A.Boiko, S.Kurbanalieva, L.U.Dzhemileva, K.S.Kozlov, V.P.Ananikov. *npj Biofilms Microbiomes*, **12**, 109 (2026); https://doi.org/10.1038/s41522-026-00971-3
30. A.Ercetin, I.Arpaci, S.Memiş, A.Akagić, Ö.Özgün. *Arch. Computat. Methods Eng.*, **33**, 11511 (2026); https://doi.org/10.1007/s11831-026-10641-4
31. N.M.Ivanova, A.S.Kashin, V.P.Ananikov. *Chemistry*, **7**, 160 (2025); https://doi.org/10.3390/chemistry7050160
32. S.Anada, Y.Nomura, K.Yamamoto. *Microscopy*, **72**, 461 (2023); https://doi.org/10.1093/jmicro/dfad037
33. M.Suzuki, Y.Igarashi. *Microscopy*, **75**, 318 (2026); https://doi.org/10.1093/jmicro/dfaf049
34. K.Kaufmann, C.Zhu, A.S.Rosengarten, D.Maryanovsky, T.J.Harrington, E.Marin, K.S.Vecchio. *Science*, **367**, 564 (2020); https://doi.org/10.1126/science.aay3062
35. T.Dahmen, P.Trampert, F.Boughorbel, J.Sprenger, M.Klusch, K.Fischer, C.Kübel, P.Slusallek. *AI Perspect*, **1**, 2 (2019); https://doi.org/10.1186/s42467-019-0002-0
36. P.Trampert, D.Rubinstein, F.Boughorbel, C.Schlinkmann, M.Luschkova, P.Slusallek, T.Dahmen, S.Sandfeld. *Crystals*, **11**, 258 (2021); https://doi.org/10.3390/cryst11030258
37. M.Joosten, J.Greer, J.Parkhurst, T.Burnley, A.J.Jakobi. *IUCrJ*, **11**, 951 (2024); https://doi.org/10.1107/S2052252524009321
38. A.Boschet, A.Collin, N.Katoch, J.Cohen-Adad. *arXiv:2412.02858* (2024); https://doi.org/10.48550/arXiv.2412.02858
39. A.Boschet, A.Collin, N.Katoch, J.Cohen-Adad. In *Deep Generative Models. DGM4MICCAI 2024. Lecture Notes in Computer Science*. (Eds A.Mukhopadhyay, I.Oksuz, S.Engelhardt, D.Mehrof, Y.Yuan). (Cham: Springer, 2025). P. 54; https://doi.org/10.1007/978-3-031-72744-3_6
40. P.Harar, L.Herrmann, P.Grohs, D.Haselbach. *Structure*, **33**, 820 (2025); https://doi.org/10.1016/j.str.2025.01.020
41. A.Colliard-Granero, J.Jitsev, M.H.Eikerling, K.Malek, M.J.Eslamibidgoli. *ACS Nanosci. Au*, **3**, 398 (2023); https://doi.org/10.1021/acsnanoscienceau.3c00020
42. K.Alrfou, T.Zhao, A.Kordijazi. *Integr. Mater. Manuf. Innov.*, **13**, 703 (2024); https://doi.org/10.1007/s40192-024-00369-z
43. Q.Li, B.Kailkhura, R.Anirudh, J.Zhang, Y.Zhou, Y.Liang, T.Y.-J.Han, P.K.Varshney. *SIAM J. Math. Data Sci.*, **3**, 1197 (2021); https://doi.org/10.1137/20m1344299
44. R.Lin, R.Zhang, C.Wang, X.-Q.Yang, H.L.Xin. *Sci. Rep.*, **11**, 5386 (2021); https://doi.org/10.1038/s41598-021-84499-w
45. T.Houben, T.Huisman, M.Pisarenco, F. van der Sommen, P.de With. *J. Micro/Nanopattern. Mater. Metrology*, **22**, 031208 (2023); https://doi.org/10.1117/1.jmm.22.3.031208
46. D.Zhang, A.Barbot, F.Seichepine, F.P.W.Lo, W.Bai, G.-Z.Yang, B.Lo. *Commun. Phys.*, **5**, 80 (2022); https://doi.org/10.1038/s42005-022-00844-z
47. M.Ragone, R.Shahabazian-Yassar, F.Mashayek, V.Yurkiv. *Prog. Mater. Sci.*, **138**, 101165 (2023); https://doi.org/10.1016/j.pmatsci.2023.101165
48. B.Costa-Gomes, J.Greer, N.Juraschko, J.Parkhurst, J.Mirecka, M.Famili, C.Rangel-Smith, O.Strickson, A.Lowe, M.Basham, T.Burnley. *Acta Crystallogr. F Struct. Biol. Commun.*, **81**, 441 (2025); https://doi.org/10.1107/S2053230X25007575
49. R.R.Lederman, J.Andén, A.Singer. *Inverse Probl.*, **36**, 044005 (2020); https://doi.org/10.1088/1361-6420/ab5ede
50. M.Wolf, D.J.DeRosier, N.Grigorieff. *Ultramicroscopy*, **106**, 376 (2006); https://doi.org/10.1016/j.ultramic.2005.11.001
51. B.D.Bunday, S.Klotzkin, D.Patriarche, M.Mukhtar, K.Maruyama, S.-K.Kang, Y.Yamazaki, presented in part at the Metrology, Inspection, and Process Control XXXVI, 2022/5/26, 2022; https://doi.org/10.1117/12.2615753
52. E.Sezestre, J.Scoarnec, J.Pradelles, L.Perraud, A.Fay, S.Bérard-Bergery, J.Bustos, J.-B. Henry, O.Dubreuil, I.Mendes, C.Valade, A.Moly, A.Batte, N.G.Schuch, F.Robert, T.Figueiro, presented in part at the Metrology, Inspection, and Process Control XXXVI, 2022/5/26, 2022; https://doi.org/10.1117/12.2616527
53. A.Gladyshev, B.Haas, T.C.Pekin, T.M.Boland, M.Schloz, P.Rez, C.T.Koch. *Nat. Commun.*, **17**, 5173 (2026); https://doi.org/10.1038/s41467-026-74135-4
54. J.Renner, M.A.Wright, K.Bouchard, B.E.Cohen, P.Ercius, A.Goldschmidt, C.C.S.Pedroso, A.Saha, P.Denes. *npj Comput. Mater.*, **12**, 244 (2026); https://doi.org/10.1038/s41524-026-02126-6
55. Y.Zhang, R.Tammaro, P.J.Peters, R.B.G.Ravelli. *J. Chem. Inf. Model.*, **60**, 2605 (2020); https://doi.org/10.1021/acs.jcim.9b01176
56. K.Wang, P.C.Doerschuk, presented in part at the Image Reconstruction from Incomplete Data VI, 2010/8/19, 2010; https://doi.org/10.1117/12.862065
57. E.-M.Papia, A.Kondi, V.Constantoudis. *Microporous Mesoporous Mater.*, **394**, 113675 (2025); https://doi.org/10.1016/j.micromeso.2025.113675
58. S.S.M.M.Rahman, M.Salomon, S.Dembélé, presented in part at the 2023 3rd International Conference on Computer, Control and Robotics (ICCCR), 2023/3/24, 2023; https://doi.org/10.1109/icccr56747.2023.10194053
59. S.P.Tiwari, S.Chhabra, F.Tama, O.Miyashita. *J. Chem. Inf. Model.*, **60**, 2570 (2020); https://doi.org/10.1021/acs.jcim.9b01107
60. X.Li, O.Dyck, S.V.Kalinin, S.Jesse. *Microsc. Microanal.*, **24**, 623 (2018); https://doi.org/10.1017/S143192761801543X
61. L.Evans, O.-V.Murad, L.Dingeldein, P.Cossio, R.Covino, M.Meila. *PRX Life*, **3**, 033025 (2025); https://doi.org/10.1103/txrb-fw3z
62. K.Wang, P.C.Doerschuk, presented in part at the 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2011/3, 2011; https://doi.org/10.1109/isbi.2011.5872788
63. P.Easwaran, M.J.Lehmann, O.Wirjadi, T.Prill, S.Didas, C.Redenbach. *Chem. Eng. Technol.*, **39**, 395 (2016); https://doi.org/10.1002/ceat.201500342
64. L.R. DaCosta, K.Sytwu, C.K.Groschner, M.C.Scott. *npj Comput. Mater.*, **10**, 165 (2024); https://doi.org/10.1038/s41524-024-01336-0
65. J.Zhang, Q.Chen, Y.Zeng, W.Gao, X.He, Z.Liu, J.Yu. *arXiv:2312.02235* (2023); https://doi.org/10.48550/ARXIV.2312.02235
66. M.O.Volman Stern, D.Hohs, A.Jansche, T.Bernthaler, G.Schneider. *Methods Microsc.*, **2**, 165 (2025); https://doi.org/10.1515/mim-2024-0016
67. M.Vuijk, G.Ducci, L.Sandoval, M.Pietsch, K.Reuter, T.Lunkenbein, C.Scheurer. *Microsc. Microanal.*, **31**, ozae130 (2025); https://doi.org/10.1093/mam/ozae130

68. R.Jiang, G.Zhang, Y.Yang, S.Wu, M.Wu, W.Feng, Y.Zhao, X.Xiao, X.Wang, T.Wang, X.Li, M.Chen, M.Xu. *arXiv:2505.23444* (2025); https://doi.org/10.48550/ARXIV.2505.23444
69. H.Eliasson, R.Erni. *Nano Lett*., **25**, 2474 (2025); https://doi.org/10.1021/acs.nanolett.4c06025
70. T.Sardhara, C.J.Cyron, M.Ritter, R.C.Aydin. *Mach. Learn. Sci. Technol*., **6**, 025023 (2025); https://doi.org/10.1088/2632-2153/adc870
71. L.Mill, D.Wolff, N.Gerrits, P.Philipp, L.Kling, F.Vollnhals, A.Ignatenko, C.Jaremenko, Y.Huang, O. De Castro, J.-N.Audinot, I.Nelissen, T.Wirtz, A.Maier, S.Christiansen. *Small Methods*, **5**, 2100223 (2021); https://doi.org/10.1002/smtd.202100223
72. J.Bals, M.Epple. *Adv. Intell. Syst*., **5**, 2300004 (2023); https://doi.org/10.1002/aisy.202300004
73. A.Y.Kharin. *Ultramicroscopy*, **219**, 113125 (2020); https://doi.org/10.1016/j.ultramic.2020.113125
74. J.-W.Lee, N.H.Goo, W.B.Park, M.Pyo, K.-S. Sohn. *Eng. Rep*., **3**, e12274 (2021); https://doi.org/10.1002/eng2.12274
75. S.Chun, S.Roy, Y.T.Nguyen, J.B.Choi, H.S.Udaykumar, S.S.Baek. *Sci. Rep*., **10**, 13307 (2020); https://doi.org/10.1038/s41598-020-70149-0
76. J.D.López Gutiérrez, I.M.Abundez Barrera, N.Torres Gómez. *Nanomaterials*, **12**, 1818 (2022); https://doi.org/10.3390/nano12111818
77. L.Fu, Z.Liu, J.Li, J.Simmons, H.Yu, S.Wang. *arXiv:2302.05541* (2023); https://doi.org/10.48550/ARXIV.2302.05541
78. A.S.Saleh, K.Croes, H.Ceric, I.D.Wolf, H.Zahedmanesh. *Comput. Mater. Sci*., **251**, 113723 (2025); https://doi.org/10.1016/j.commatsci.2025.113723
79. C.Shen, J.Zhao, M.Huang, C.Wang, Y.Zhang, W.Xu, S. Zheng. *Mater. Charact*., **217**, 114407 (2024); https://doi.org/10.1016/j.matchar.2024.114407
80. S.Iijima, T.Ohashi, Y.-C.Lin, Y.Ng, presented in part at the 2025 19th International Conference on Machine Vision and Applications (MVA), 2025/7/26, 2025; https://doi.org/10.23919/mva65244.2025.11175084
81. A.Khan, C.-H.Lee, P.Y.Huang, B.K.Clark. *npj Comput. Mater*., **9**, 85 (2023); https://doi.org/10.1038/s41524-023-01042-3
82. H.Eliasson, A.Lothian, I.Surin, S.Mitchell, J.Pérez-Ramírez, R.Erni. *Small Methods*, **9**, 2401108 (2025); https://doi.org/10.1002/smtd.202401108
83. D.Roldán, L.Barbosa-Torres. *Opt. Commun*., **574**, 131029 (2025); https://doi.org/10.1016/j.optcom.2024.131029
84. J.Boehm, B.Dey. *arXiv:2606.24817* (2026); https://doi.org/10.48550/arxiv.2606.24817
85. N. Da Silva De Sa, A.Stewart. *BIO Web Conf*., **129**, 10011 (2024); https://doi.org/10.1051/bioconf/202412910011
86. C.-H.Lee, A.Khan, B.Clark, P.Y.Huang. *Microsc. Microanal*., **31**, ozaf048.1074 (2025); https://doi.org/10.1093/mam/ozaf048.1074
87. S.Zahran, Z.Ao, Z.Zhang, C.Chi, C.Yuan, Y.Wang. *arXiv:2602.01710* (2026); https://doi.org/10.48550/arXiv.2602.01710
88. N.Davydiuk, E.Krieg, J.Gaitzsch, P.M.McCall, G.K.Auernhammer, M.Yang, J.B.Tracy, S.Bals, W.J.Parak, N.A.Kotov, L.M.Liz-Marzán, A.Fery, M.Faria, Q.A.Besford. *Nat. Nanotechnology*, **20**, 1174 (2025); https://doi.org/10.1038/s41565-025-02009-9
89. Y.Wang, X.Tang, J.Fan, G.Xiao, presented in part at the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020/12/16, 2020; https://doi.org/10.1109/bibm49941.2020.9312978
90. C.Lu, K.Chen, H.Qiu, X.Chen, G.Chen, X.Qi, H.Jiang. *bioRxiv* (2023); https://doi.org/10.1101/2023.07.12.548636
91. G.Saranathan, A.Ghosh, M.Foltin, A.J.Koomthanam, A.Tripathy, M.Ziatdinov, S.Bhattacharya, K.Roccapriore, P.Faraboschi, presented in part at the SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2024/11/17, 2024; https://doi.org/10.1109/scw63240.2024.00270
92. I.P.Beckman, G.Berry, J.Ross, G.Riveros, H.Cho. *J. Aerosol Sci*., **171**, 106164 (2023); https://doi.org/10.1016/j.jaerosci.2023.106164
93. A.P.Tafti, A.B.Kirkpatrick, Z.Alavi, H.A.Owen, Z.Yu. *Micron*, **78**, 54 (2015); https://doi.org/10.1016/j.micron.2015.07.005
94. A.P.Tafti, J.D.Holz, A.Baghaie, H.A. Owen, M.M.He, Z.Yu. *Micron*, **87**, 33 (2016); https://doi.org/10.1016/j.micron.2016.05.004
95. C.García-Gascón, J.Bas-Bolufer, P.Castelló-Pedrero, F.Chinesta. *Fiber. Polym*., **27**, 435 (2026); https://doi.org/10.1007/s12221-025-01218-2
96. H.Altendorf, S.Didas, T.Batt. In *Forum Bildverarbeitung*. (Eds F.P.Leon, M. Heizmann). (Regensburg: KIT Scientific Publishing, 2010). P. 59
97. J.G.Quijada-Pioquinto, E.I.Kurkin, E.Y.Minaev, A.V.Gavrilov, presented in part at the 2023 IX International Conference on Information Technology and Nanotechnology (ITNT), 2023/4/17, 2023; https://doi.org/10.1109/itnt57377.2023.10139073
98. E.Kurkin, E.Minaev, A.Sedelnikov, J.G.Q.Pioquinto, V.Chertykovtseva, A.Gavrilov. *Technologies*, **12**, 249 (2024); https://doi.org/10.3390/technologies12120249
99. N.A.Marnautov, M.V.Matveev, A.A.Gulin, T.Kálai, B.Bognár, A.T.Rebrikova, N.A.Chumakova. *J. Phys. Chem. C*, **128**, 2543 (2024); https://doi.org/10.1021/acs.jpcc.3c07127
100. L.Lv, F.He, J.Zhang, L.Mao, J.Fan, G.Xiao, X.Tang, presented in part at the 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021/12/9, 2021; https://doi.org/10.1109/bibm52615.2021.9669432
101. F.Liang, Y.Zhang, C.Zhou, H.Zhang, G.Liu, J.Zhu. *PLoS One*, **19**, e0311228 (2024); https://doi.org/10.1371/journal.pone.0311228
102. M.d.M.Juez Lorenzo, V.Kolarik, K.Sethia, P.Strakos. *High Temp. Corros. Mater*., **101**, 1497 (2024); https://doi.org/10.1007/s11085-024-10321-3
103. B.Limanond, P.Moraille, K.Boonpalit, K.Suk-ieam, V.Nganpipatmongkon, P.Pakawatpanurut, A.Badia, T.Okatani, S.Deebansok. *Zenodo*, version 1 (2026); https://doi.org/10.5281/zenodo.20928897
104. P.P.Shinde, P.P.Pai, S.P.Adiga, K.Subramanya Mayya, Y.Seo, M.Hwang, H.Go, C.Park. *Heliyon*, **11**, e43377 (2025); https://doi.org/10.1016/j.heliyon.2025.e43377
105. T.Figueiro, M.Saib, K.-H.Choi, C.Hohle, M.J.Thornton, C.Vannufel, J.-H.Tortai, P.Schiavone, presented in part at the Photomask Technology 2013, 2013/9/23, 2013; https://doi.org/10.1117/12.2026423
106. A.P.Tafti, A.B.Kirkpatrick, H.A.Owen, Z.Yu. In *Advances in Visual Computing. ISVC 2014. Lecture Notes in Computer Science*. (Eds G. Bebis, R.Boyle, B.Parvin, D.Koracin, R.McMahan, J.Jerald, H.Zhang, S.M.Drucker, Ch. Kambhamettu, M.Choubassi, Zh.Deng, M.Carlson). (Cham: Springer, 2014). P. 141; https://doi.org/10.1007/978-3-319-14364-4_14
107. J.Lindblad, N.Sladoje, A.Suveer, A.Dragomir, I.-M.Sintorn, presented in part at the 2015 International Conference on Image Processing Theory, Tools and Applications (IPTA), 2015/11, 2015; https://doi.org/10.1109/ipta.2015.7367090
108. C.Aguilar, M.Comer, I.Hanhan, R.Agyei, M.Sangid, presented in part at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2021/6, 2021; https://doi.org/10.1109/cvprw53098.2021.00415
109. G.Kumar, S.Kumar, A.Sharma, A.Raturi, A.Shrivastava, A.L.N.Rao, A.K.Khan. *J. Nano- Electron. Phys*., **16**, 06008 (2024); https://doi.org/10.21272/jnep.16(6).06008
110. K.P.Treder, C.Huang, C.G.Bell, T.J.A.Slater, M.E.Schuster, D.Özkaya, J.S.Kim, A.I.Kirkland. *npj Comput. Mater*., **9**, 18 (2023); https://doi.org/10.1038/s41524-022-00949-7
111. I.Z.Borshon, M.Ragone, A.H.Phakatkar, L.Long, R.Shahbazian-Yassar, F.Mashayek, V.Yurkiv. *npj Comput. Mater*., **10**, 275 (2024); https://doi.org/10.1038/s41524-024-01461-w
112. M.A.Ziatdinov, M.Y.Yaman, Y.Liu, D.Ginger, S.V.Kalinin. *Digit. Discov*., **3**, 1213 (2024); https://doi.org/10.1039/d3dd00196b

113. S.Chen, S.Lin, J.Liu, Y.Huang, Y.Sima, F.Lin. *Microstructures*, **6**, 2026055 (2026); https://doi.org/10.20517/microstructures.2025.153
114. K.Govind, D.Oliveros, A.Dlouhy, M.Legros, S.Sandfeld. *Mach. Learn. Sci. Technol.*, **5**, 015006 (2024); https://doi.org/10.1088/2632-2153/ad1a4e
115. R.K.Vasudevan, M.Ziatdinov, S.Jesse, S.V.Kalinin. *Nano Lett.*, **16**, 5574 (2016); https://doi.org/10.1021/acs.nanolett.6b02130
116. M.Leibovich, M.Tan, R.Manzorro, A.Marcos-Morales, S.Mohan, P.A.Crozier, C.Fernandez-Granda. *arXiv:2601.17046* (2026); https://doi.org/10.48550/arXiv.2601.17046
117. K.Loehr, R.Raja, X.Ding, J.Huang, G.Nolan, S.h.Bae, B.K.Clark, P.Y.Huang. *arXiv:2606.27587* (2026); https://doi.org/10.48550/arxiv.2606.27587
118. E.Monier, T.Oberlin, N.Brun, X.Li, M.Tencé, N.Dobigeon. *Ultramicroscopy*, **215**, 112993 (2020); https://doi.org/10.1016/j.ultramic.2020.112993
119. I.Pinto-Huguet, M.Botifoll, X.Chen, M.B.Eriksen, J.Yu, G.Isella, A.Cabot, G.Merino, J.Arbiol. *Adv. Intell. Syst.*, **8**, e202501077 (2026); https://doi.org/10.1002/aisy.202501077
120. S.Mohan, J.L.Vincent, R.Manzorro, P.A.Crozier, C.Fernandez-Granda, E.P.Simoncelli. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. (Eds M.Ranzato, A.Beygelzimer, Y.Dauphin, P.S.Liang, J.W.Vaughan). (Red Hook, NY, USA: Curran Associates Inc., 2021). P. 23727; https://doi.org/10.5555/3540261.3542078
121. H.C.Stites, U.Manor. *BMC Methods*, **2**, 1 (2025); https://doi.org/10.1186/s44330-024-00020-5
122. W.Thornley, S.Sullivan–Allsop, R.Cai, N.Clark, R.Gorbachev, S.J.Haigh. *npj Comput. Mater.*, **12**, 68 (2026); https://doi.org/10.1038/s41524-025-01939-1
123. K.Kamijo, M.Shiga, S.Kanomi, T.Miyata, H.Jinnai. *Res. Sq.*, (2026); https://doi.org/10.21203/rs.3.rs-9374144/v1
124. H.-F. Yang, Y.Choe. In *Advances in Visual Computing. ISVC 2010. Lecture Notes in Computer Science*. (Eds G.Bebis, R.Boyle, B.Parvin, D.Koracin, R.Chung, R.Hammound, M.Hussain, Tan Kar-Han, R.Crawfis, D.Thalmann, D.Kao, L.Avila). (Berlin, Heidelberg: Springer, 2010). P. 322; https://doi.org/10.1007/978-3-642-17274-8_32
125. C.-M.Svensson, K.G.Bondoc, G.Pohnert, M.T.Figge. *Comput. Vis. Image Underst.*, **154**, 64 (2017); https://doi.org/10.1016/j.cviu.2016.08.003
126. X.Tao, Z.Ma, C.Gao, X.Li. *Powder Technol.*, **479**, 122609 (2026); https://doi.org/10.1016/j.powtec.2026.122609
127. E.Walsh, A.Stewart. *Microsc. Microanal.*, **28**, 3046 (2022); https://doi.org/10.1017/s1431927622011370
128. B.Lin, N.Emami, D.A.Santos, Y.Luo, S.Banerjee, B.-X.Xu. *npj Comput. Mater.*, **8**, 88 (2022); https://doi.org/10.1038/s41524-022-00767-x
129. C.Ly, C.A.Nizinski, A.Toydemir, C.Vachet, L.W.McDonald IV, T.Tasdizen. *Microsc. Microanal.*, **27**, 1465 (2021); https://doi.org/10.1017/s1431927621012915
130. A.Teurtrie, N.Perraudin, T.Holvoet, H.Chen, D.T.L.Alexander, G.Obozinski, C.Hébert. *Mach. Learn. Sci. Technol.*, **5**, 045050 (2024); https://doi.org/10.1088/2632-2153/ad9192
131. H.Chen, D.T.L.Alexander, C.Hébert. *Nano Lett.*, **24**, 10177 (2024); https://doi.org/10.1021/acs.nanolett.4c02446
132. A.E.Leschziner, E.Nogales. *J. Struct. Biol.*, **153**, 284 (2006); https://doi.org/10.1016/j.jsb.2005.10.012
133. P.Chandramouli, R.Hernandez-Lopez, H.-W.Wang, A.E.Leschziner. *J. Struct. Biol.*, **175**, 85 (2011); https://doi.org/10.1016/j.jsb.2011.04.012
134. S.H.W.Scheres, R.Melero, M.Valle, J.-M.Carazo. *Structure*, **17**, 1563 (2009); https://doi.org/10.1016/j.str.2009.10.009
135. B.Turoňová, F.K.M.Schur, W.Wan, J.A.G.Briggs. *J. Struct. Biol.*, **199**, 187 (2017); https://doi.org/10.1016/j.jsb.2017.07.007
136. C.J.Prust, P.C.Doerschuk, J.E.Johnson. Presented in part at the 2006 IEEE International Conference on Electro/Information Technology, 2006/5, 2006; https://doi.org/10.1109/eit.2006.252192
137. C.J.Prust, P.C.Doerschuk, J.E.Johnson. Presented in part at the 2006 International Conference on Image Processing, 2006/10, 2006; https://doi.org/10.1109/icip.2006.312624
138. B.B.Cheikh, E.Baudrier, G.Frey. Presented in part at the 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015/4, 2015; https://doi.org/10.1109/isbi.2015.7164049
139. A.K.Louis, M.Riplinger, M.Spiess, E.Spodarev. *Inverse Probl.*, **27**, 035015 (2011); https://doi.org/10.1088/0266-5611/27/3/035015
140. R.Yao, J.Qian, Q.Huang. *Bioinformatics*, **36**, 1252 (2020); https://doi.org/10.1093/bioinformatics/btz728
141. H.Lei, Y.Yang. *Front. Genet.*, **11**, 627746 (2021); https://doi.org/10.3389/fgene.2020.627746
142. Z.Yang, H.Li, D.Zang, R.Han, F.Zhang. In *Bioinformatics research and applications. ISBRA 2023. Lecture Notes in Computer Science*. (Eds X.Guo, S.Mangul, M.Patterson, A.Zelikovsky). (Singapore: Springer, 2023). P. 325; https://doi.org/10.1007/978-981-99-7074-2_25
143. S.-C.Chung, P.-C.Chou. *J. Struct. Biol.*, **217**, 108239 (2025); https://doi.org/10.1016/j.jsb.2025.108239
144. L.Dingeldein, D.Silva-Sánchez, L.Evans, E.D'Imprima, N.Grigorieff, R.Covino, P.Cossio. *Proc. Natl. Acad. Sci. U.S.A.*, **122**, e2420158122 (2025); https://doi.org/10.1073/pnas.2420158122
145. A.Levy, R.Raghu, J.R.Feathers, M.Grzadkowski, F.Poitevin, J.D.Johnston, F.Vallese, O.B.Clarke, G.Wetzstein, E.D.Zhong. *Nat. Methods*, **22**, 1486 (2025); https://doi.org/10.1038/s41592-025-02720-4
146. B.Koo, J.Martel, A.Peck, A.Levy, F.Poitevin, N.Miolane. *arXiv:2306.07274* (2023); https://doi.org/10.48550/ARXIV.2306.07274
147. D.Elmlund, H.Elmlund. *J. Struct. Biol.*, **167**, 83 (2009); https://doi.org/10.1016/j.jsb.2009.04.009
148. R.Lian, B.Huang, L.Wang, Q.Liu, Y.Lin, H.Ling. *Acta Crystallogr. D Struct. Biol.*, **78**, 174 (2022); https://doi.org/10.1107/S2059798321011761
149. N.A.Anoshina, D.V.Sorokin. *Comput. Math. Model.*, **34**, 1 (2023); https://doi.org/10.1007/s10598-024-09591-y
150. N.A.Anoshina, D.V.Sorokin. Presented in part at the Proceedings of the 33rd International Conference on Computer Graphics and Vision, 2023; https://doi.org/10.20948/graphicon-2023-643-653
151. E.Jansson, J.Krook, K.Modin, O.Öktem. *SIAM J. Imaging Sci.*, **18**, 2509 (2025); https://doi.org/10.1137/24m1704336
152. T.Prill, K.Schladitz, D.Jeulin, M.Faessel, C.Wieser. *J. Microsc.*, **250**, 77 (2013); https://doi.org/10.1111/jmi.12021
153. C.Fend, A.Moghiseh, C.Redenbach, K.Schladitz. *J. Microsc.*, **281**, 16 (2021); https://doi.org/10.1111/jmi.12944
154. H.-F.Yang, Y.Choe. Presented in part at the 2009 IEEE Symposium on Computational Intelligence for Multimedia Signal and Vision Processing, 2009/3, 2009; https://doi.org/10.1109/cimsvp.2009.4925647
155. N.A.Sokolov, E.P.Vasiliev, A.A.Getmanskaya. *Comput. Opt.*, **47**, 778 (2023); https://doi.org/10.18287/2412-6179-CO-1273
156. N.A.Sokolov, E.P.Vasiliev, A.A.Getmanskaya. Presented in part at the Proceedings of the 32nd International Conference on Computer Graphics and Vision, 2022/09/19, 2022; https://doi.org/10.20948/graphicon-2022-706-714
157. A.I.Mironov, N.A.Sokolov, A.A.Serebrjakova. Presented in part at the Proceedings of the 35th International Conference on Computer Graphics and Vision, 2025/12/1, 2025; https://doi.org/10.25686/978-5-8158-2474-4-2025-828-834
158. K.E.Yashin, N.A.Sokolov, E.P.Vasiliev, A.A.Getmanskaya. Presented in part at the Proceedings of the 33rd International Conference on Computer Graphics and Vision, 2023, 2023; https://doi.org/10.20948/graphicon-2023-654-662
159. V.Patakvölgyi, L.Kovács, D.A.Drexler. Presented in part at the 2025 IEEE 29th International Conference on Intelligent Engineering Systems (INES), 2025/6/11, 2025; https://doi.org/10.1109/ines67149.2025.11078023
160. H.Kniesel, P.Rapp, P.Hermosilla, T.Ropinski. Presented in part at the 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2025/10/19, 2025; https://doi.org/10.1109/iccvw69036.2025.00604

161. B.Chen, T.Xin, H.Han, X.Chen. Presented in part at the Medical Imaging 2022: Image Processing, 2022/4/4, 2022; https://doi.org/10.1117/12.2611465
162. D.B.Ress, M.L.Harlow, R.M.Marshall, U.J.McMahan. *Structure*, **12**, 1763 (2004); https://doi.org/10.1016/j.str.2004.07.022
163. T.Leuliet, L.Friot-Giroux, W.Baaziz, E.Bretin, O.Ersen, F.Peyrin, B.Sixou, V.Maxim. Presented in part at the 2020 28th European Signal Processing Conference (EUSIPCO), 2021/1/18, 2021; https://doi.org/10.23919/eusipco47968.2020.9287762
164. W.Millsaps, J.Schwartz, Z.W.Di, Y.Jiang, R.Hovden. *Microsc. Microanal.*, **29**, 1650 (2023); https://doi.org/10.1093/micmic/ozad083
165. M.Kalinowski, A.Daurat, G.T.Herman. In *Graph-Based Representations in Pattern Recognition. GbRPR 2007. Lecture Notes in Computer Science*. (Eds F.Escolano, M.Vento). (Berlin, Heidelberg: Springer, 2007). P. 263; https://doi.org/10.1007/978-3-540-72903-7_24
166. E.Rosten, S.Cox. In *Advances in Visual Computing. ISVC 2006. Lecture Notes in Computer Science*. (Eds G.Bebis, R.Boyle, B.Parvin, D.Koracin, P.Remagnino, A.Nefian, G.Meenakshisundaram, V.Pascucci, J.Zara, J.Molineros, H.Theisel, T.Malzbender). (Berlin, Heidelberg: Springer, 2006). P. 373; https://doi.org/10.1007/11919476_38
167. J.G.Tarolli, L.M.Jackson, N.Winograd. *J. Am. Soc. Mass Spectrom.*, **25**, 2154 (2014); https://doi.org/10.1007/s13361-014-0927-7
168. Z.Wang, X.Huan, K.Garikipati. *Comput. Methods Appl. Mech. Eng.*, **377**, 113706 (2021); https://doi.org/10.1016/j.cma.2021.113706
169. R.Jacobs. *Comput. Mater. Sci.*, **211**, 111527 (2022); https://doi.org/10.1016/j.commatsci.2022.111527
170. M.Jonák, J.Dorazil, V.Myška, R.Burget, M.Kotrlý. Presented in part at the 2025 17th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2025/11/3, 2025; https://doi.org/10.1109/icumt67815.2025.11268630
171. L.Drumetz, M.D.Mura, S.Meulenyzer, S.Lombard, J.Chanussot. *J. Electron. Imaging*, **24**, 061109 (2015); https://doi.org/10.1117/1.jei.24.6.061109
172. P.Belamri, H.Proudhon, D.Texier, D.Ryckelynck. *Mater. Des.*, **258**, 114599 (2025); https://doi.org/10.1016/j.matdes.2025.114599
173. G.B.Huang, H.-F.Yang, S.Takemura, P.Rivlin, S.M.Plaza. *arXiv:2012.12175* (2020); https://doi.org/10.48550/arXiv.2012.12175
174. M.Martinc, G.Dražić, A.Kokalj, K.Žiberna, J.Roknić, M.Poberžnik, S.Džeroski, A.B.Golob. *Sci. Rep.*, **16**, 27448 (2026); https://doi.org/10.1038/s41598-026-57754-1
175. K.Shaga Devan, P.Walther, J.von Einem, T.Ropinski, H.A.Kestler, C.Read. *Cell. Microbiol.*, **23**, e13280 (2021); https://doi.org/10.1111/cmi.13280
176. H.Zhang, N.Söderholm, L.Sandblad, K.Wiklund, M.Andersson. *Microsc. Microanal.*, **25**, 711 (2019); https://doi.org/10.1017/S1431927619000308
177. T.Lin, Y.Shi, N.Shu, D.Cheng, X.Hong, J.Song, B.H.Gwee. *Microelectron. Reliab.*, **123**, 114196 (2021); https://doi.org/10.1016/j.microrel.2021.114196
178. H.Kniesel, T.Ropinski, T.Bergner, K.S.Devan, C.Read, P.Walther, T.Ritschel, P.Hermosilla. Presented in part at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022/6/18, 2022; https://doi.org/10.1109/cvpr52688.2022.02010
179. T.-Y.Lan, T.Bendory, N.Boumal, A.Singer. *IEEE Trans. Signal Process.*, **68**, 1589 (2020); https://doi.org/10.1109/tsp.2020.2975943
180. T.M.Mayhew. *J. Exp. Bot.*, **62**, 4101 (2011); https://doi.org/10.1093/jxb/err176
181. Z.Tong, Z.Wang, X.Wang, Y.Ma, H.Guo, C.Liu. *Constr. Build. Mater.*, **274**, 121898 (2021); https://doi.org/10.1016/j.conbuildmat.2020.121898
182. E.Azqadan, A.Arami, H.Jahed. *J. Magnes. Alloy.*, **13**, 4231 (2025); https://doi.org/10.1016/j.jma.2025.07.019
183. Z.Lei, J.Zhou, Y.Wang, S.Chen, Y.Zhang, Z.Zhang. *Int. J. Fatigue*, **200**, 109072 (2025); https://doi.org/10.1016/j.ijfatigue.2025.109072
184. J.S.v.Hulst, W.P.M.H.Heelmes, D.J.Antunes. *arXiv:2603.16549* (2026); https://doi.org/10.48550/arxiv.2603.16549
185. D.Possart, L.Mill, F.Vollnhals, T.Hildebrand, P.Suter, M.Hoffmann, J.Utz, D.Augsburger, M.Thies, M.Gu, F. Wagner, G.Sarau, S.Christiansen, K.Breininger. *npj Comput. Mater.*, **11**, 197 (2025); https://doi.org/10.1038/s41524-025-01702-6
186. W.Yuan, B.Yao, S.Tan, F.You, Q.He. *npj Comput. Mater.*, **11**, 322 (2025); https://doi.org/10.1038/s41524-025-01800-5
187. M.J.Lynch, R.Jacobs, G.A.Bruno, P.Patki, D.Morgan, K.G.Field. *npj Comput. Mater.*, **11**, 261 (2025); https://doi.org/10.1038/s41524-025-01756-6
188. K.P.Treder, C.Huang, J.S.Kim, A.I.Kirkland. *Microscopy*, **71**, i100 (2022); https://doi.org/10.1093/jmicro/dfab043
189. A.S.Aglikov, M.V.Zhukov, T.A.Aliev, V.I.Maslii, P.V.Gelfenshtein, D.A.Kozodaev, D.V.Andreeva, M.Nosonovsky, E.V.Skorb. *Mater. Horiz.*, **12**, 10713 (2025); https://doi.org/10.1039/d5mh01327e
190. I.Alxneit. *J. Phys. Chem. A*, **124**, 10075 (2020); https://doi.org/10.1021/acs.jpca.0c07840
191. L.R.DaCosta, M.C.Scott. *arXiv:2512.09067* (2025); https://doi.org/10.48550/arXiv.2512.09067
192. G.Abdelal, C.-W.Chan, S.McLoone. *Appl. Sci.*, **16**, 975 (2026); https://doi.org/10.3390/app16020975
193. C.Ly, W.Frazier, A.Olsen, I.Schwerdt, L.W.McDonald IV, A.Hagen. *Comput. Mater. Sci.*, **249**, 113639 (2025); https://doi.org/10.1016/j.commatsci.2024.113639
194. I.Ahmad, F.S.Chowdhury, A.I.Boullerne, A.Gow, D.L.Feinstein. *ASN Neuro*, **18**, 2603411 (2026); https://doi.org/10.1080/17590914.2025.2603411
195. H.Zhou, J.Xu, X.Qin, J.Zhang, W.Zou, M.Shakouri, J.Xu, L.Li, J.Ding, H.Pang. *Nano Res.*, **18**, 94908095 (2025); https://doi.org/10.26599/nr.2025.94908095
196. J.Dapueto, V.P.Pastore, N.Noceti, F.Odone. Presented in part at the 2025 International Joint Conference on Neural Networks (IJCNN), 2025/6/30, 2025; https://doi.org/10.1109/ijcnn64981.2025.11228114
197. M.Bonomi, S.Hanot, C.H.Greenberg, A.Sali, M.Nilges, M.Vendruscolo, R.Pellarin. *Structure*, **27**, 175 (2019); https://doi.org/10.1016/j.str.2018.09.011
198. V.P.Ananikov, D.M.Arkhipova, A.S.Kashin, A.P.Vorozhtsov. *Russ. Chem. Rev.*, **95** (11), RCR5239 (2026); https://doi.org/10.59761/RCR5239