npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-026-02243-2

Improving atomic force microscopy structure discovery via style-translation

Check for updates

Jie Huang[1], Niko Oinonen[1,2], Fabio Priante[3], Filippo Federici Canova[1,2], Lauri Kurki[1], Chen Xu[1] & Adam S. Foster[1,4] ✉

Atomic force microscopy (AFM) is a key tool for characterising nanoscale structures, with functionalised tips now offering detailed images of the atomic structure. In parallel, AFM simulations using the probe particle model provide a cost-effective approach for rapid AFM image generation. Using state-of-the-art machine learning models and substantial simulated datasets, properties such as molecular structure, electrostatic potential, and molecular graph can be predicted from AFM images. However, transferring model performance from simulated to experimental AFM images poses challenges due to the subtle variations in real experimental data compared to the seemingly flawless simulations. In this study, we explore style translation to augment simulated images and improve the predictive performance of machine learning models in surface property analysis. We reduce the style gap between simulated and experimental AFM images and demonstrate the method's effectiveness in enhancing structure discovery models through local structural property distribution comparisons. This research presents a novel approach to improving the efficiency of machine learning models in the absence of labelled experimental data.

Understanding material design and properties at the nanoscale relies heavily on visualising atomic structures. Among many characterisation tools, atomic force microscopy (AFM) stands out for its ability to image surfaces with atomic resolution[1,2]. In particular, frequency-modulated non-contact AFM provides contrast that reflects tip-sample interactions, enabling indirect visualisation of atomic configurations through frequency shift signals[3–5]. However, interpreting AFM images, especially for complex molecules, requires both experience and intuition[6,7]. While experts can propose reasonable structure hypotheses, validating these guesses typically involves density functional theory (DFT) calculations and AFM simulations like the probe particle model (PPM)[8–10] to compare simulated AFM images with experiments[11,12]. This simulation-aided process has proven valuable, but it is iterative, computationally demanding, and challenging to scale for large systems. Consequently, it is appealing to develop machine learning (ML) models that are capable of directly predicting and inferring atomic configurations from AFM images, thereby possibly automating the structure discovery process.

Recent advances in ML have shown promise for automating structure discovery from scanning probe microscopy images. A wide range of methods has been explored. Convolutional neural networks (CNNs) have been used to map AFM images to atomic descriptors such as van der Waals (vdW) spheres[13], and to predict atomic positions and orientations in ice structures[14,15]. CNN-based approaches have also been applied to automated molecular recognition in scanning tunnelling microscope (STM) images[16,17]. By integrating CNNs with graph neural networks (GNNs), it is possible to predict complete atomic graphs, including 3D coordinates, directly from AFM data[18,19]. Other approaches leverage AFM image fingerprints to identify molecular candidates and assign confidence scores to predictions[20]. In parallel, multimodal recurrent neural networks (mRNNs) have been used to infer chemical nomenclature, such as IUPAC names, from AFM images[21]. Generative models have also contributed to the progress. Variational autoencoders (VAEs) have been used to synthesise AFM images to improve molecular classification tasks[22]. Conditional generative adversarial networks (cGANs) have been used to convert AFM images into interpretable ball-and-stick representations[23]. More broadly, GANs have been applied to generate realistic microscopy images, such as in scanning transmission electron microscopy (STEM), for training defect detection models[24], and to simulate STM images[25]. Additionally, support vector machines (SVMs) have been used for real-time classification and feature recognition during AFM measurements[26].

Despite these advances, two key challenges continue to limit the real-world deployment of ML-based structure discovery tools. First, the simulation-to-real domain gap poses a fundamental obstacle. Although PPM-based simulation can produce AFM images that closely resemble

[1]Department of Applied Physics, Aalto University, Helsinki, Finland. [2]Nanolayers Research Computing Ltd., London, UK. [3]Department of Chemistry and Materials Science, Aalto University, Helsinki, Finland. [4]WPI Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kakuma-machi, Kanazawa, Japan. ✉e-mail: adam.foster@aalto.fi

experimental observations, the style mismatch between domains still exists: real AFM images often contain noise, artefacts, and subtle distortions that are absent in idealised simulations. This domain gap hinders the generalisation of simulation-trained models to experimental data. Second, the scarcity of ground-truth atomic structures for experimental AFM images makes it extremely difficult to systematically evaluate the structure prediction performance on real datasets, let alone training the model on experimental data.

To tackle these challenges, we propose an approach that integrates style translation[27] with structure prediction[13,18,19] to bridge the gap between simulated and experimental AFM images. By translating simulated images into experimental-like ones, we improve model performance on real experimental data. Additionally, we introduce structure-based evaluation metrics that do not rely on ground-truth atomic structures. Our results show that style translation enhances prediction accuracy and enables evaluation in the absence of direct validation.

Here, we outline the challenges and hypotheses in ML-based structure discovery from AFM images and introduce our style translation framework. We then describe the dataset construction and present prediction comparisons across models. Finally, we evaluate model performance based on physically meaningful properties.

Results

Problem definitions and hypothesis

As shown in Fig. 1, we let m, u, v represent an atomic configuration, a simulated 3D AFM image, and an experimental 3D AFM image, respectively. Correspondingly, $\mathcal{M}$, $\mathcal{U}$, and $\mathcal{V}$ denote the domains of atomic configurations, simulated AFM images, and experimental AFM images. The forward problem is defined as: given an atomic configuration, what is the corresponding 3D AFM image? This question can be addressed in two ways. The first is by physically performing the AFM experiment for the given sample, thereby obtaining experimental 3D AFM images, i.e., realising the mapping from $\mathcal{M}$ to $\mathcal{V}$. The second approach involves simulating 3D AFM images from atomic configurations using the PPM, representing the mapping from $\mathcal{M}$ to $\mathcal{U}$. In contrast, the inverse problem is defined as: given a 3D AFM image, what is the corresponding atomic structure? Assuming that the simulated and experimental AFM images are drawn from the same underlying data distribution, a structure discovery model $F_{\mathcal{U}}$ is trained using pairs from the domain $\mathcal{U}$ and their corresponding atomic configurations in $\mathcal{M}$. The trained model $F_{\mathcal{U}}$ is then applied directly to experimental AFM images from domain $\mathcal{V}$. This assumption is primarily driven by two practical limitations: (1) it is costly and time-consuming to collect sufficient experimental AFM samples, and (2) more critically, it is often unfeasible to obtain the corresponding ground-truth atomic configurations for experimental images, particularly for large or complex samples. As a result, training is conducted using ($\mathcal{U} \to \mathcal{M}$) rather than directly using experimental data ($\mathcal{V} \to \mathcal{M}$).

However, in this study, we do not assume that simulated and experimental AFM images are drawn from the same distribution. While the simulated AFM images resemble experimental ones quite closely, notable discrepancies remain, especially in image noise and artefacts introduced by real-world experimental conditions. We refer to this discrepancy as the style gap between simulated and experimental AFM images. The first hypothesis of this work is that the presence of a style gap degrades the performance of a model $F_{\mathcal{U}}$ trained solely on simulated data when applied to the experimental AFM images. The second hypothesis is that reducing this style gap in the training data can improve the model's performance on the experimental AFM images, motivating a data-driven style translation approach, which is described in the next section.

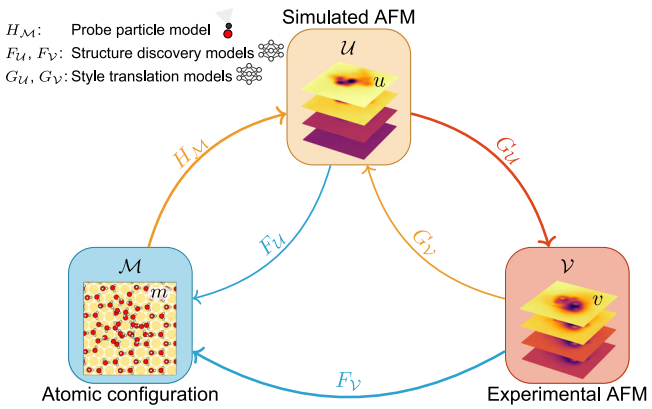


Fig. 1 | Overview of structure discovery from experimental AFM images. Simulated AFM images $\mathcal{U}$ are generated from atomic configurations $\mathcal{M}$ using the probe particle model (PPM) $H_{\mathcal{M}}$. A machine learning-based structure discovery model $F_{\mathcal{U}}$ is trained to recover the atomic configuration from a simulated 3D AFM image, learning the mapping from $\mathcal{U}$ to $\mathcal{M}$. Ideally, the structure discovery model $F_{\mathcal{V}}$ is trained on experimental style AFM images $\mathcal{V}$ and their corresponding atomic configurations $\mathcal{M}$, enabling structure prediction directly from experimental data. To bridge the style gap between simulated and experimental AFM images, style translators $G_{\mathcal{U}}$ and $G_{\mathcal{V}}$ are used to translate domains $\mathcal{U}$ and $\mathcal{V}$, respectively.

Data-driven style translation

To reduce the style gap between the simulated and experimental AFM image domains, we employ an image-to-image translation model that takes simulated images as input and generates images in the style of experimental data. This model is inspired by the GAN[28] framework, a class of machine learning models designed to generate new data samples that resemble a target distribution. Unlike conventional GANs, which generate data from random noise vectors, our image-to-image translation model uses source domain images, simulated AFM images in our case, as direct inputs. Specifically, we adopt a cycle-consistent generative adversarial network (CycleGAN) framework[27,29], which extends the GAN architecture by introducing a cycle consistency constraint. CycleGAN has demonstrated effectiveness across a wide range of applications, including statistical tasks such as density estimation and manifold fitting[30,31], medical image synthesis[32,33], and high-fidelity image generation in STEM and STM images[24,25].

Figure 2 illustrates the CycleGAN architecture in this study[27,31] (for details please refer to Methods). We define two domains: $\mathcal{U}$ and $\mathcal{V}$, representing the distributions of simulated and experimental AFM images, respectively. Two datasets are constructed: $\mathcal{U}_M = \{u_i\}_{i=1}^{M}$ and $\mathcal{V}_N = \{v_i\}_{i=1}^{N}$, where u and v are 2D AFM image slices, and $M = 729$, $N = 728$ denote the number of image samples in each dataset. Each 2D AFM slice represents the frequency shift of the cantilever over a horizontal plane at a fixed height. We intentionally set $M \neq N$ to emphasise that the datasets are unpaired, i.e., there is no one-to-one correspondence between images in the

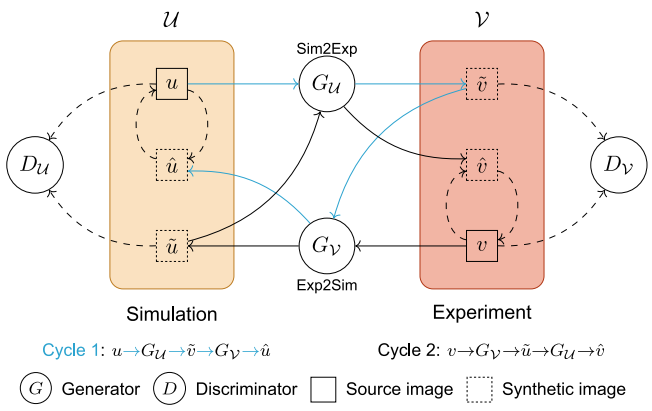


Fig. 2 | Schematic representation of the framework of CycleGAN. The shaded rectangles indicate the domains of simulated and experimental AFM images, while the circular components represent the individual sub-networks. The solid arrows are the transfer paths within the generators, while the dashed arrows denote the paths for computing the loss function.

two domains. In order to learn the bidirectional mapping between domains, two image-to-image generators are defined: $G_{\mathcal{U}} : \mathcal{U} \to \mathcal{V}$ and $G_{\mathcal{V}} : \mathcal{V} \to \mathcal{U}$. Given real input images u and v, the generators produce the synthetic output points $\tilde{v} = G_{\mathcal{U}}(u)$ and $\tilde{u} = G_{\mathcal{V}}(v)$, respectively. To enforce cycle consistency, the outputs are further transformed back to $\hat{u} = G_{\mathcal{V}}(\tilde{v})$ and $\hat{v} = G_{\mathcal{U}}(\tilde{u})$, with the aim that $\hat{u} \approx u$ and $\hat{v} \approx v$. Hence, there are two cycle loops as shown in Fig. 2. To distinguish real from generated images, two discriminators are employed: $D_{\mathcal{V}}$ and $D_{\mathcal{U}}$. Each discriminator receives an image as input and outputs a number indicating whether the image is real or synthetic. The adversarial training process encourages the generators to produce images indistinguishable from the real domain samples, as judged by the corresponding discriminators.

One of the key advantages of the CycleGAN framework is that it does not require paired images. In the context of AFM, a paired image would consist of simulated and experimental AFM images derived from exactly the same underlying atomic configuration - obtaining sufficient examples in practice for AFM is unfeasible. CycleGAN overcomes this limitation by learning from unpaired datasets: it only requires a sufficient number of samples from each domain, without any explicit correspondence between them. After training the CycleGAN framework, we obtain two domain translation models: the forward generator $G_{\mathcal{U}}$, which maps simulated AFM images from domain $\mathcal{U}$ to the experimental style AFM images in domain $\mathcal{V}$, and the reverse generator $G_{\mathcal{V}}$, which performs the opposite mapping.

As shown in Fig. 3a, two example 2D simulated AFM images u_1 and u_2 are transformed by the forward generator $G_{\mathcal{U}}$ into $\tilde{v}_1, \tilde{v}_2$, respectively, which resemble experimental AFM images. The forward generator learns to add stylistic features like noise into the simulated images to make them visually consistent with experimental images. On the contrary, the reverse generator $G_{\mathcal{V}}$ produces outputs visually closer to simulated AFM images by removing noise and experimental artefacts, as shown in Fig. 3b, suggesting potential applications in denoising or artefact reduction in experimental images. The backward translation provides an alternative route for AFM structure discovery: experimental AFM images can first be translated into simulated-style images and then analysed by a structure discovery model trained on simulated data. This reverse-path workflow is conceptually appealing because it treats the inverse generator as a denoising or artefact-reduction step. We tested this possibility as shown in Figs. S24–S35 in the Supporting Information (SI), but found that it did not provide more reliable predictions than the forward workflow for the present dataset and model setting as shown in Figs. S36–S39.

In this study, we focus on the forward translation ($\mathcal{U} \to \mathcal{V}$) to reduce the style gap and generate structure discovery training data that better matches the characteristics of experimental AFM images. For comparisons, Fig. 3c shows several representative handcrafted perturbations applied to u_1: (1) Gaussian noise, (2) random small cutout region, shown as a black rectangle, (3) random gradient background, (4) combinations of the first three perturbations, and (5) salt-and-pepper noise. Except for salt-and-pepper noise, these perturbations have been used as data augmentations in previous studies[17,19]. These handcrafted perturbations are used only as simple baseline augmentations. They mimic generic image degradation, such as noise, background drift, and partial information loss, but do not fully capture experimental AFM artefacts such as scan lines, edge effects, local textures, and nonuniform backgrounds – this emphasises the important role of the CycleGAN approach in capturing more realistic experimental artefacts. To evaluate the quality of the style translation, we design a quantitative and data-driven evaluation workflow including the following steps:

Step 1: Machine expert training. We first train a binary classifier, referred to as a machine expert, that takes a 2D AFM image as the input and outputs an authenticity score $s \in [0, 1]$. The CNN-based classifier is trained on the datasets using two kinds of labelled data: one consisting of simulated AFM images labelled with $s = 0$ and another consisting of experimental AFM images labelled with $s = 1$. The architecture and training details of the binary classifier are described in Methods. As shown in Fig. 3d, the resulting authenticity score distributions, denoted $\rho_{\mathcal{U}}(s)$ for simulated data and $\rho_{\mathcal{V}}(s)$ for experimental data, are well-separated. This indicates that the classifier is effective in distinguishing between the two domains.

Step 2: Authenticity distribution shift after style translation. We apply the forward generator $G_{\mathcal{U}}$ to the set $\mathcal{U}$, producing the style-translated image set $\widetilde{\mathcal{V}} = G_{\mathcal{U}}(\mathcal{U})$. This set is then evaluated using the machine expert to obtain the authenticity distribution $\rho_{\widetilde{\mathcal{V}}}(s)$. As shown in Fig. 3e, the distribution $\rho_{\widetilde{\mathcal{V}}}(s)$ closely aligns with $\rho_{\mathcal{V}}(s)$, demonstrating that the forward style generator $G_{\mathcal{U}}$ successfully bridges the style gap between simulation and experiment. For comparison, Fig. 3f shows the authenticity distributions after applying handcrafted perturbations. These methods fail to produce observable shifts in authenticity, suggesting that these perturbations are less effective in making the simulated images resemble experimental ones, at least from the perspective of the machine expert.

Step 3: Quantitative evaluation with distribution metrics. To quantitatively assess the style similarity, we compute the Wasserstein Distance (also known as Earth Mover's Distance)[34–36]$\mathrm{WD}(\cdot \parallel \mathcal{V})$ between the authenticity distributions of the translated/perturbed sets and the experimental distributions $\mathcal{V}$. Additionally, we measure the corresponding Fréchet Inception Distance (FID)[37]$\mathrm{FID}(\cdot \parallel \mathcal{V})$ between the distributions of different image sets and the distribution of the authentic experimental image set $\mathcal{V}$. Unlike the Wasserstein distance $\mathrm{WD}(\cdot|\mathcal{V})$, which uses features from a lightweight, AFM-specific model (the machine expert), FID is based on features from a large Inception network[38] trained on general images from ImageNet[39]. Although FID is not specifically designed for AFM images, it can still capture statistical differences between AFM domains before and after style translation. Both methods have their respective strengths and limitations; hence, we use both metrics with the motivation to provide a more complete evaluation. For more information on the details of WD and FID, please refer to Methods.

As illustrated in Fig. 3g, we visualise the comparison results in a two-dimensional Cartesian coordinate system, where the x-axis represents $\mathrm{FID}(\cdot \parallel \mathcal{V})$ and the y-axis represents $\mathrm{WD}(\cdot \parallel \mathcal{V})$. Each image set is mapped to a point in this space. The origin corresponds to the experimental dataset $\mathcal{V}$, representing zero style gap. The point labelled $\mathcal{U}$ denotes the pure simulated dataset, and the Euclidean distance between $\mathcal{U}$ and the origin captures the initial style gap between simulation and experiment, shown as the shaded blue region. The style-translated and perturbed versions of $\mathcal{U}$ are plotted in the same coordinate space. The ideal translation trajectory, where a simulated image set is perfectly converted into the experimental style, is shown by the dashed arrow. The solid arrow from $\mathcal{U}$ to $\widetilde{\mathcal{V}}$ illustrates the style translation of generator $G_{\mathcal{U}}$. This transformation largely reduces both the FID and Wasserstein distances, effectively reducing the style gap between the original (blue) region and a smaller (orange) region, demonstrating the success of the learned translation. In contrast, the points corresponding to handcrafted perturbations show limited or no reduction in style gap. Notably, adding salt-and-pepper noise actually increases the distance from the experimental distribution, moving the simulated images further away from the target domain.

While these results indicate that the learned translator reduces the distribution-level style gap, they do not by themselves guarantee that all physically relevant image content is perfectly preserved. In particular, CycleGAN is trained as an unpaired image-to-image translation model, and the objective is to match the target image style rather than to explicitly constrain every atom-relevant feature. Therefore, an overly aggressive translation could, in principle, modify not only contrast, background, and noise, but also spatial features that are important for structure prediction. To reduce this risk, we used the cycle-consistency loss during training and tuned the loss weights to keep the translation sufficiently conservative. Additional feature-similarity and slice-to-slice consistency checks are provided in Figs. S19–S23 in SI to assess content preservation after style translation. It is also worth noting that the authenticity score used here serves only as a supplementary indicator of the domain shift from simulated to experimental-like images. A reduced domain shift alone does not guarantee that the translated dataset will lead to the best structure discovery model. Therefore, additional evaluation metrics for structure discovery are essential, and these are discussed in the following sections.

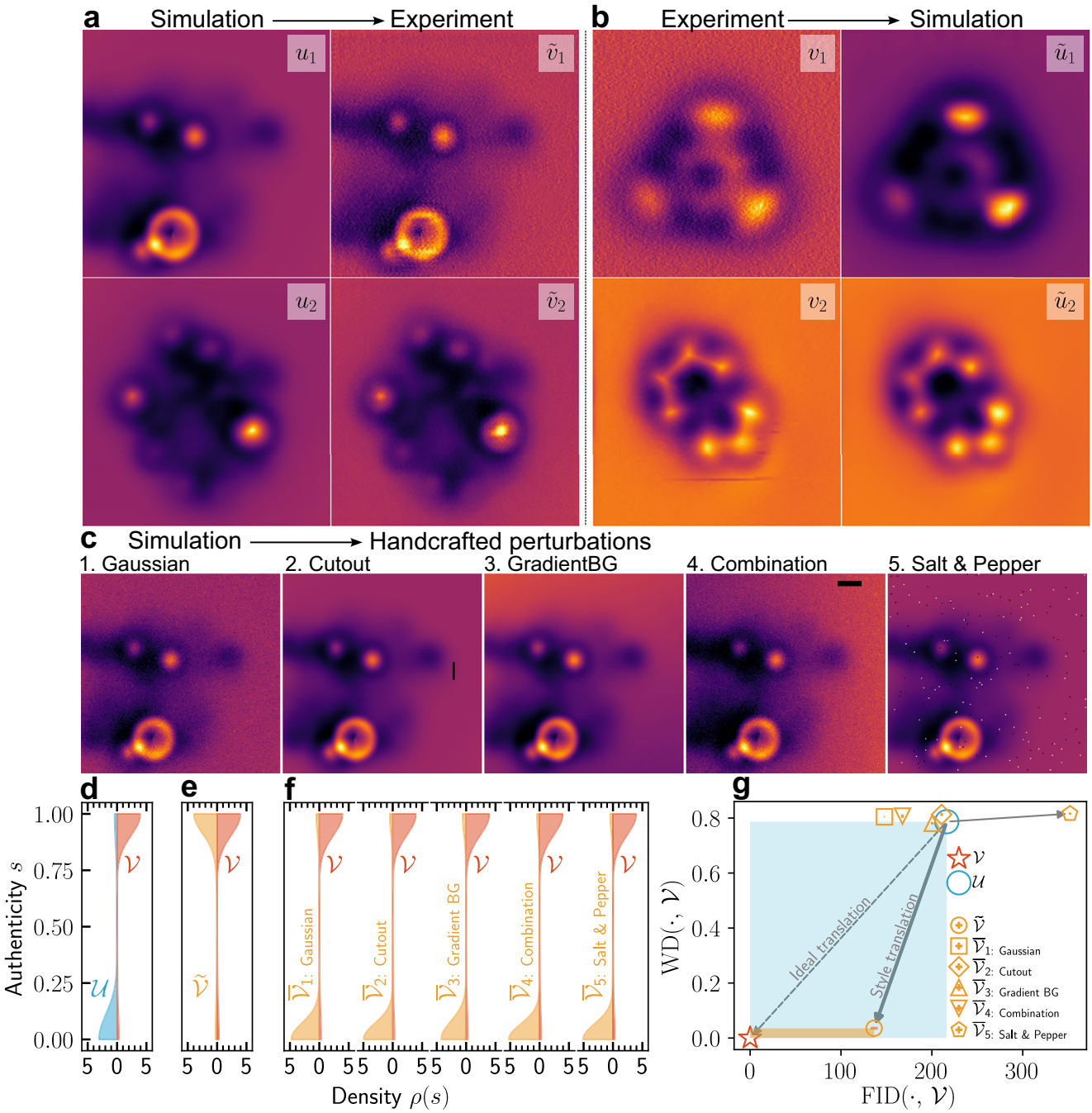


Fig. 3 | Translating AFM image styles and comparing domain shifts using authenticity and distribution metrics. a Simulated AFM images translated into experimental style using generator $G_{\mathcal{U}}$. **b** Experimental AFM images translated into simulation style using generator $G_{\mathcal{V}}$. **c** Simulated images with handcrafted perturbations. **d** Machine-expert authenticity-score s distributions for simulated $\mathcal{U}$ and experimental $\mathcal{V}$ images. **e** Distributions for translated experimental-style $\widetilde{\mathcal{V}}$ and experimental $\mathcal{V}$ images. **f** Distributions for perturbed simulated $\overline{\mathcal{V}}$ and experimental $\mathcal{V}$ images. **g** Style gap comparisons using Wasserstein distance and Fréchet Inception Distance (FID) relative to the experimental domain $\mathcal{V}$.

In summary, the CycleGAN-based forward translator $G_{\mathcal{U}}$ provides an effective solution for bridging the style gap between simulated and experimental AFM images. Unlike the handcrafted perturbations, which may unintentionally move the data distribution further from the experimental style, the learned translation yields significantly improved realism and authenticity, as supported by both qualitative visual comparisons and quantitative distribution metrics.

The following analysis involves several image sets, configuration sets, and structure discovery models. To make the notation clear, we first summarise the key symbols in Table 1.

Structure discovery model training and predictions

With the availability of experimental-style AFM images generated by the forward generator, we construct training datasets for the structure discovery model that maps 3D AFM images to atomic configurations. In this study, we use bilayer water configurations $\mathcal{M}$ adsorbed on the Au(111) surface, obtained from simulations based on DFT and machine learning potentials[19]. Figure 4a demonstrates one example of an atomic configuration m, where the dashed lines indicate the unit cell and the rectangle indicates the xy region used for AFM image simulations via PPM $H_{\mathcal{M}}$. A side view of this configuration is shown in Fig. 4b. The

Table 1 | Key notation for configurations, image sets, and prediction models

m; $\mathcal{M}$	atomic configuration; configuration set
v; $\mathcal{V}$	experimental AFM image; experimental set
u; $\mathcal{U}$	simulated AFM image; simulated set
$\bar{v}$; $\bar{\mathcal{V}}$	handcrafted-perturbed image; perturbed set
$\tilde{v}$; $\tilde{\mathcal{V}}$	experimental-style image; experimental-style set
$v^{\dagger}$; $\mathcal{V}^{\dagger}$	hybrid image; hybrid set
$F_{\mathcal{D}}$; $F_{\mathcal{D}}(v)$	model trained on $\mathcal{D}$; structure prediction for v

Lowercase symbols denote individual atomic configurations or 3D AFM images, while calligraphic symbols denote the corresponding sets. Each 3D AFM image contains multiple 2D slices at different tip heights. Here, $\mathcal{D}$ denotes a generic image set with known configuration labels, such as $\mathcal{U}$, $\bar{\mathcal{V}}$, $\tilde{\mathcal{V}}$, or $\mathcal{V}^{\dagger}$.

corresponding oxygen density profiles along the z axis are plotted in Fig. 4c. The solid line represents the density of the single configuration m, while the dashed line shows the average density across the configuration set $\mathcal{M}$. Two distinct peaks around $z = 3.3$ Å and $z = 5.9$ Å reveal the bilayer nature of the structures. Figure 4d shows three representative slices from a 3D simulated AFM image u, generated from configuration m. Pixel intensities reflect the frequency shift signals in AFM imaging. To create the corresponding experimental-style AFM image $\tilde{v}$, we apply the trained forward style generator $G_{\mathcal{U}}$ to each 2D slice of u and stack them to reconstruct the 3D AFM image, as shown in Fig. 4e. Figure 4f shows pixel intensity profiles for simulated and experimental style AFM images along the P_0P_1 line, from which differences can be inferred. We can see that $\tilde{v}$ is noisier and contains more details compared to the original simulated image u. The pixel-wise differences $\Delta = \tilde{v} - u$ are shown as the shaded regions in Fig. 4f, indicating the style change.

To examine whether decreasing the style gap in training data helps to improve structure prediction on experimental AFM images, we design two kinds of training datasets: 1. Simulation dataset: Each training sample (u, m) contains a 3D simulated AFM image u and its corresponding atomic configuration m. 2. Experimental-style dataset: Each training sample $(\tilde{v}, m)$ contains a 3D experimental-style AFM image $\tilde{v}$ and the same atomic configuration m. Through the style translation, we transform the simulated data distribution into a new experimental-like distribution that approximates the real experimental distribution. It follows logically that a model trained on this distribution, denoted $F_{\tilde{\mathcal{V}}}$, should generalise better to real experimental AFM images than a model $F_{\mathcal{U}}$ trained solely on the simulation dataset. In the following discussion, we evaluate and compare the performance of models trained on different datasets to test our hypothesis and validate the impact of style translation on the accuracy of structure discovery. For more details on simulation data generation, including DFT calculations and AFM simulations, please refer to our previous studies[18,19]. The architecture of the structure discovery model and training details are summarised in Methods.

Figure 5 presents the predicted atomic configurations from different structure discovery models. Each model F receives an input of a 3D experimental AFM image with vertical resolution of $\Delta z = 0.1$ Å along the surface normal. The first two columns display representative 2D AFM slices of the input 3D AFM image at relatively far and close distances to the sample surface, respectively. Column 3 shows the atomic configuration predictions from the model $F_{\mathcal{U}}$, which is trained solely on the simulation dataset. Column 4 shows predictions from model $F_{\bar{\mathcal{V}}}$ trained on the dataset with handcrafted perturbations. Column 5 shows the results from $F_{\tilde{\mathcal{V}}}$, trained on experimental-style images generated via style translation. Column 6 presents predictions from $F_{\mathcal{V}^{\dagger}}$, trained on hybrid images combining style translation and handcrafted perturbations. For the style translation, we use cycle-consistency and identity loss weights of $\lambda_c = 20$ and $\lambda_i = 1$, respectively. The sizes of the atoms in each prediction reveal the relative heights of each atom in the z-axis, allowing for a visual assessment of vertical structure. Rows a–f represent different experimental AFM samples.

Qualitatively, model $F_{\mathcal{U}}$ predicts fewer atoms compared to the other models. This raises a key question: are these atoms genuinely absent in the experiment, or are they simply missed by the model due to limitations in generalising from simulation data? To investigate this, we run PPM simulations using the predicted atomic configurations and compare the resulting AFM images with the experimental inputs. A good prediction should yield a recovered simulation that closely matches the experimental image. The corresponding results are shown in Figs. S1–S4 in SI. These comparisons indicate that $F_{\mathcal{U}}$ indeed misses atoms, particularly evident in sample f (Fig. S1). Among the other models, predictions from $F_{\tilde{\mathcal{V}}}$ and $F_{\mathcal{V}^{\dagger}}$ tend to be more conservative than those from $F_{\bar{\mathcal{V}}}$, especially for the atoms positioned at lower z values. However, the recovered AFM images from all three models appear similar (Figs. S2–S4), suggesting that the suppressed underlying atoms have limited influence on the overall AFM signal. Notably, the experimental-style-trained models exhibit enhanced robustness against image noise, suggesting improved generalisation to experimental data features.

However, since the exact atomic structures from these AFM images are unknown, we cannot perform a direct accuracy evaluation like that on the simulation data. Consequently, visual inspection alone is insufficient to assess model performance. Here, we introduce an evaluation approach based on local structural properties that bypasses the need for ground-truth atomic configurations.

Performance evaluations

Evaluating model performance on experimental AFM images is inherently challenging due to the absence of ground-truth atomic configurations, unlike in simulation datasets, where both AFM images and corresponding atomic structures are known. In contrast to typical machine learning tasks, such as image classification, where humans can often establish ground truth, structure discovery presents a fundamentally harder problem: even experts cannot reliably determine the exact atomic configurations from AFM images alone. Nevertheless, even without exact atomic labels, predicted structures should still exhibit physically meaningful local properties, such as atom-atom distance and angle distributions, that are consistent with those obtained from first-principles simulations such as DFT[40]. Therefore, instead of comparing each prediction to an unavailable ground truth, we assess model performance by comparing the statistical distributions of local structural properties across many predicted configurations. This provides an indirect yet meaningful evaluation of the physical validity of the predicted structures.

As shown in Fig. 6, we calculated local structural distributions for two cases: (1) all the water molecules in the configuration set $\mathcal{M}$, which includes both the top and bottom layers on the Au(111) surface, and (2) only the top layer water molecules. This comparison allows us to see the differences between the two layers.

Figure 6a demonstrates the oxygen-oxygen distance d_{OO} distribution within a 3.5 Å cutoff, characterising intermolecular structure. A dominant peak appears around 2.75 Å for both cases, while an additional peak near 2.95 Å arises from the interaction between the bottom layer of water molecules and the Au(111) surface. Figure 6b, c shows the distributions of oxygen-hydrogen distance d_{OH} and the angle θ_{HOH}, respectively, both within a 1.25 Å cutoff to capture the intramolecular geometry. Structural differences are observed for the top layer, reflecting its greater configurational freedom compared to the bottom layer, which is more constrained by the substrate. We use free OH bonds[41–43] to capture the surface-related orientation of water molecules. Figure 6d demonstrates the distribution of the angle θ_{ZOH} between the free OH bond vector and the surface normal. To analyse hydrogen bonding, we calculate the joint distribution of donor-acceptor oxygen distance $d_{O_dO_a}$ and hydrogen bond angle $\theta_{O_dHO_a}$, shown in Fig. 6e. A hydrogen bond is considered present if $d_{O_dO_a} < 3.5$ Å and $\theta_{O_dHO_a} > 120^{\circ}$. Each point in the space corresponds to a donor-acceptor pair that satisfies this geometric criterion. In addition, inspired by studies on using order parameters to study the local water structures[44,45], to evaluate the

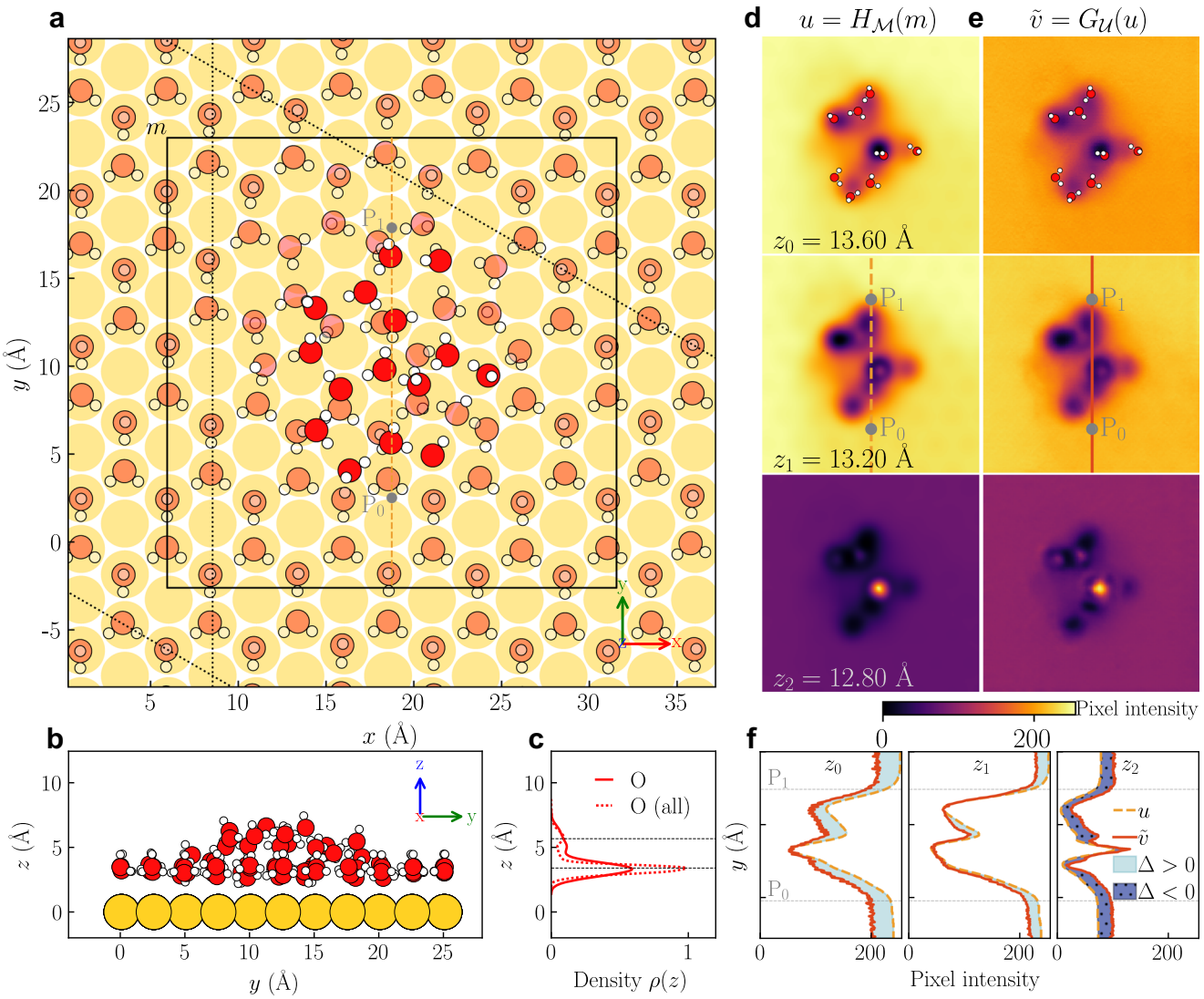


Fig. 4 | Schematic illustration of a training sample. a, **b** One simulated atomic configuration m of a bilayer water molecule cluster on the Au(111) surface from top and side views, respectively. **c** The probability density $\rho(z)$ of the oxygen atoms along the z axis, where solid and dashed lines correspond to the density of this specific configuration m and the mean density of many configurations $\mathcal{M}$. **d** The simulated AFM image u of the configuration m at different heights from z_0 to z_2 through the PPM $H_{\mathcal{M}}$. **e** The experimental style AFM image $\tilde{v}$ obtained by a simulation-to-experiment style translation model $G_{\mathcal{U}}$ with the simulated AFM image u as the input. **f** The pixel intensity comparisons between simulated u and experimental style $\tilde{v}$ AFM images along the direction from P_0 to P_1 at different heights, where the shadow areas indicate the difference of pixel intensity between experimental style $\tilde{v}$ and simulated u AFM images.

structure order within the water network, we calculate the joint distribution of two tetrahedral order parameters: the translational order parameter S_k and the orientational tetrahedral order parameter S_g[46,47]. The translational tetrahedral order S_k is defined as

$$S_k = \frac{1}{3}\sum_{k=1}^{4}\frac{(r_k - \bar{r})^2}{4\bar{r}^2}, \tag{1}$$

where r_k is the radial distance from the central oxygen atom to the k-th peripheral oxygen atom, and $\bar{r}$ is the arithmetic mean of the four radial distances. This parameter quantifies the variance in radial distances, with $S_k = 0$ for a perfect tetrahedron and increasing as the structure becomes distorted. The orientational tetrahedral order S_g is defined as

$$S_g = \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left(\cos\psi_{j,k} + \frac{1}{3}\right)^2 \tag{2}$$

where $\psi_{j,k}$ is the angle between bonds j and k at the central oxygen. A perfect tetrahedron yields $S_g = 0$, while random angular arrangements yield an average $\langle S_g \rangle \approx 1$ due to the normalisation factor. The joint distribution of S_k and S_g is illustrated in Fig. 6f, where each point corresponds to a local environment consisting of a central oxygen atom and its four nearest neighbours within a 3.5 Å cutoff.

To evaluate the performance of a given structure discovery model on experimental AFM images, we first feed the 3D experimental images, together with their rotated versions, into the model to obtain atomic configuration predictions (for predictions of other rotated images, please refer to Figs. S11–S13 in SI). We then calculate the distributions of local structural properties from the predicted structures and compare them with the corresponding theoretical reference distributions. The resulting distributional distances are converted into performance scores. This procedure is applied to all models, and the final scores are averaged over model replicas trained with the same hyperparameter setting.

Figure 7 shows a comparative performance evaluation of three structure discovery models trained on modified datasets: $F_{\overline{\mathcal{V}}}$

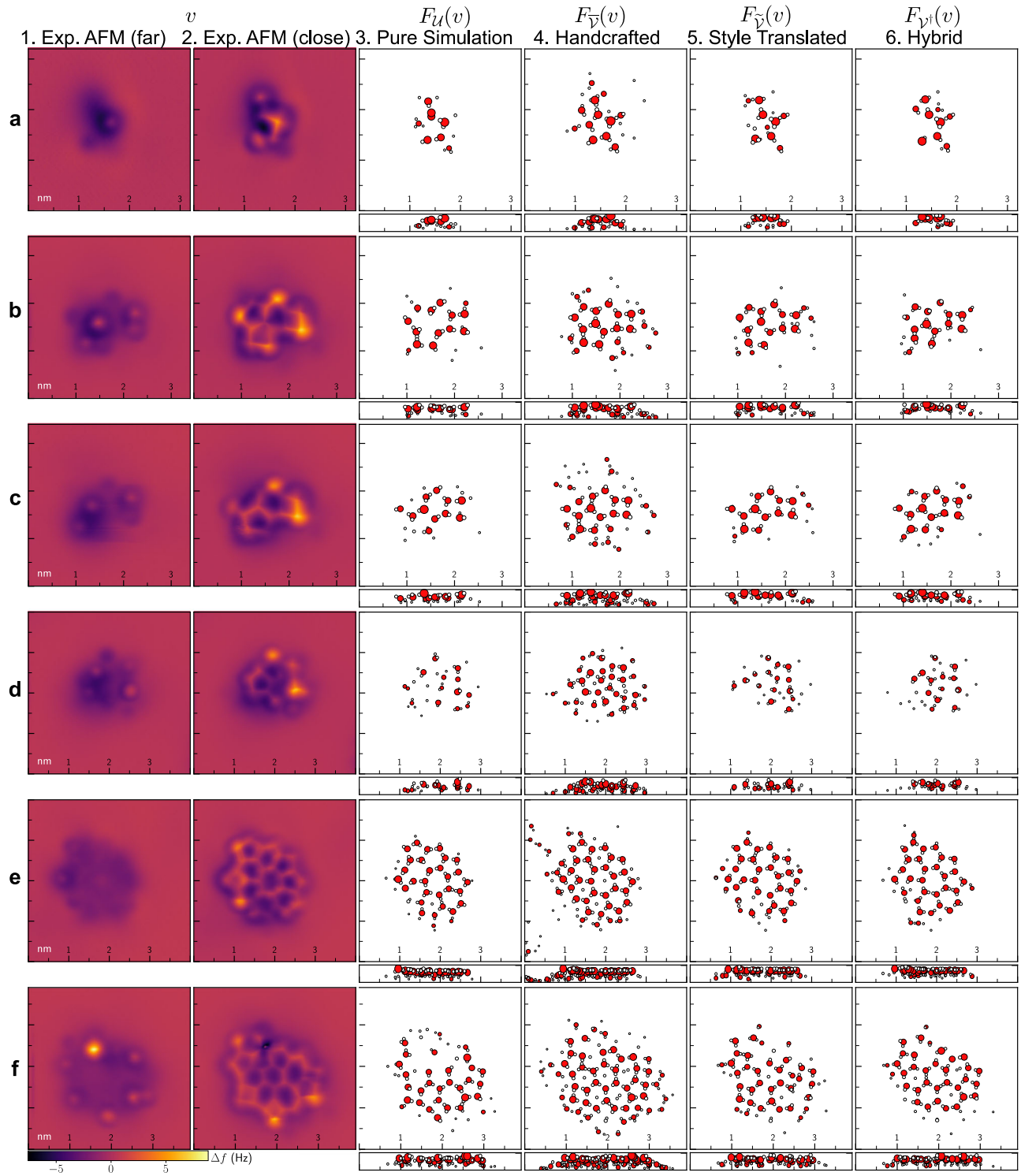


Fig. 5 | Atomic configuration predictions from experimental AFM images using structure discovery models trained on different datasets. Models trained on four types of images, including pure simulated AFM images $\mathcal{U}$, handcrafted perturbed images $\overline{\mathcal{V}}$, style translated images $\widetilde{\mathcal{V}}$, and hybrid images $\mathcal{V}^\dagger$ where both style translation and handcrafted perturbation are applied. Panels **a**-**f** represent different experimental AFM samples. For each predicted structure, the xy view is shown together with the corresponding xz view below it, providing both lateral and height information.

(handcrafted perturbations), $F_{\widetilde{\mathcal{V}}}$ (style-translated) and $F_{\mathcal{V}^\dagger}$ (hybrid), against the baseline model $F_{\mathcal{U}}$ that is trained on pure simulated AFM images. Each row corresponds to one of three distributional distance metrics used to assess structural fidelity: Wasserstein distance (WD, top row), energy distance (ED, middle row), and maximum mean discrepancy (MMD, bottom row). For detailed calculations of these metrics, please refer to Methods.

Each radar chart visualises normalised performance scores across six structural properties, as previously defined in Fig. 6: the oxygen-oxygen distance d_{OO}, the oxygen-hydrogen distance d_{OH}, the angle θ_{HOH}, the free OH orientation angle θ_{ZOH}, the hydrogen bond geometry (d_{OdOa}, θ_{OdHOa}), and the tetrahedral order parameters (S_k, S_g). For each property, we compute the distance between the predicted and theoretical distributions. These distances are normalised using min-max normalisation across all models in

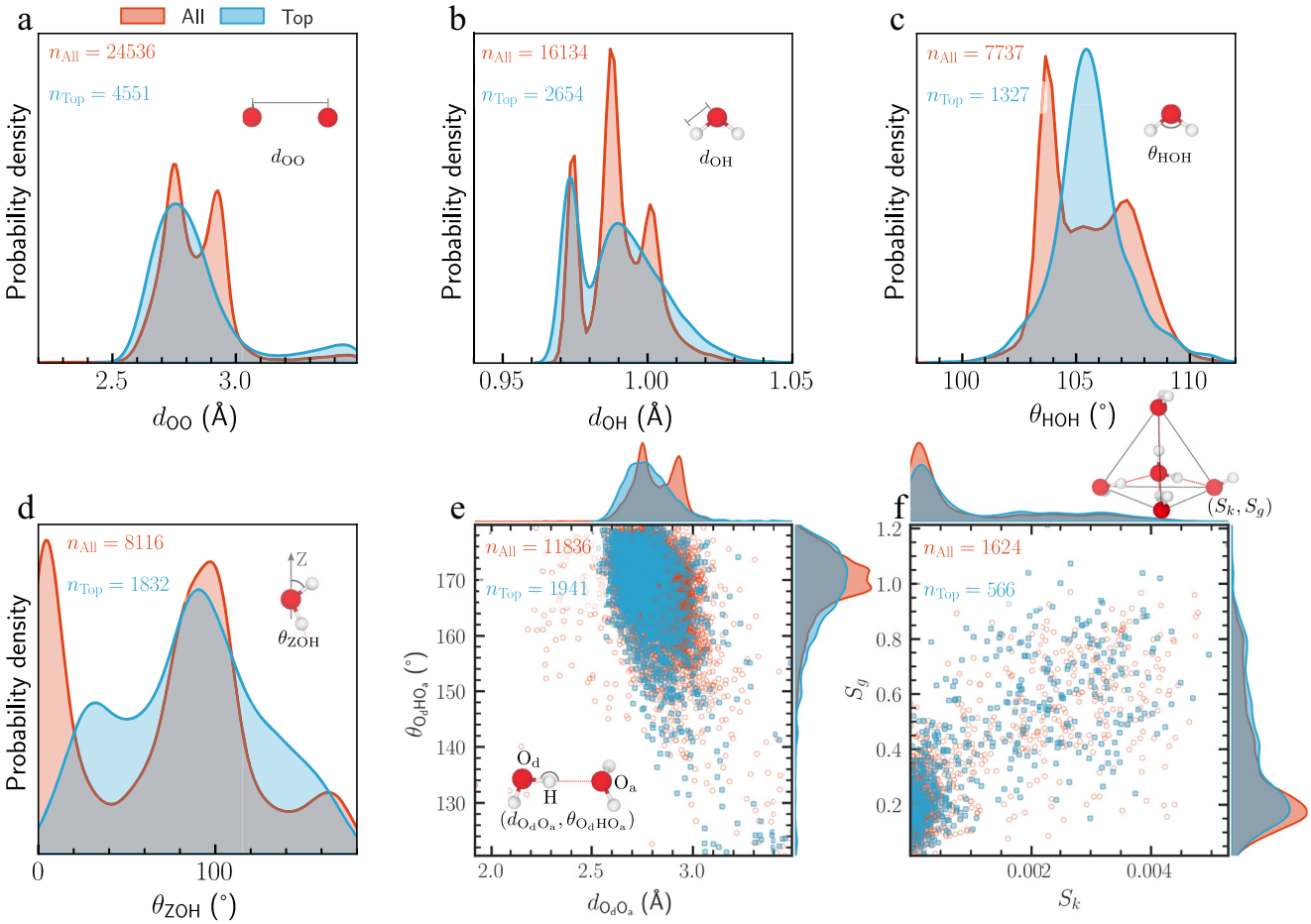


Fig. 6 | Local structural distributions of water molecules in the configuration set $\mathcal{M}$ and in the top interfacial layer obtained from simulations. a Distribution of oxygen-oxygen distances d_{OO} within a cutoff of 3.5 Å. **b** Distribution of oxygen-hydrogen distances d_{OH} within a cutoff of 1.25 Å. **c** Distribution of the intramolecular bond angle θ_{HOH}. **d** Distribution of the angle θ_{ZOH} between the free OH bond and the surface normal (z direction). **e** Joint distribution of the donor-acceptor oxygen distance d_{OdOa} and the hydrogen bond angle θ_{OdHOa}, used to identify hydrogen bonds based on geometric criteria (d_{OdOa} < 3.5 Å and θ_{OdHOa} > 120°). **f** Joint distribution of translational (S_k) and orientational (S_g) tetrahedral order parameters, which together characterise the local structural order. All distributions were computed from 74 water configurations sampled from $\mathcal{M}$.

our computational experiments with varying hyperparameter settings (λ_c, λ_i). To convert distances into performance scores, we use 1 − normalised distance, such that higher values indicate better agreement with the theoretical distributions. The reference theoretical target distributions in Fig. 7 are obtained from the top-layer water molecules in the configuration set $\mathcal{M}$ as shown in Fig. 6. In each plot, the dark-colored polygon represents the performance of the baseline model $F_{\mathcal{U}}$, and the orange, red, and blue polygons represent the performance of models trained with handcrafted, style-translated, and hybrid datasets, respectively. The error bars are obtained from the standard error of different replicas of the structure discovery model trained on the same dataset. For more details on distribution comparisons for these six structural properties, please refer to Figs. S5–S10 in SI. A detailed summary of the scores in Fig. 7d, including the number of model replicas and the mean and standard error for each model with the same hyperparameter setting, is provided in Table S1 in the SI.

Figure 7a shows that handcrafted perturbations provide moderate performance improvements in specific properties like angle θ_{HOH} and hydrogen bond geometry (d_{OdOa}, θ_{OdHOa}), but offer little benefit for other structural metrics. Figure 7b shows style translation yields broad improvements across most structural properties, particularly in oxygen-oxygen distance and hydrogen bond geometry (d_{OdOa}, θ_{OdHOa}). However, performance gains are less evident for order parameters (S_k, S_g). The tetrahedral order parameters are inherently less local than the others, requiring the central oxygen atom to have another four oxygen neighbours within 3.5 Å, posing challenges in accurately computing the distributional distances when limited tetrahedral environments are found in the predicted configurations (see Fig. S10 in SI). Figure 7c shows that hybrid datasets achieve the most balanced and consistent performance, improving nearly all metrics simultaneously. These results support the idea that reducing the image style gap between the simulated and experimental AFM images improves the accuracy and physical consistency of the predicted atomic structures.

While the distribution-level comparison above provides indirect validation on real experimental AFM images, the corresponding ground-truth atomic configurations are unavailable. As an additional validation, we evaluated the structure discovery models on experimental-style AFM images generated from simulated inputs with known reference structures. This dataset serves as a proxy for real experimental AFM data while retaining the corresponding atomic configurations, allowing direct comparison between predicted and reference structures. Individual examples are shown in Figs. S14–S17 in the SI, and quantitative comparisons are presented in Fig. S18. These additional results further support that experimental-style translation provides an effective route for improving structure prediction performance on experimental AFM images.

Discussion

With the goal of solving the inverse structure discovery problem of mapping real experimental AFM images to atomic configurations, we aim to fill the

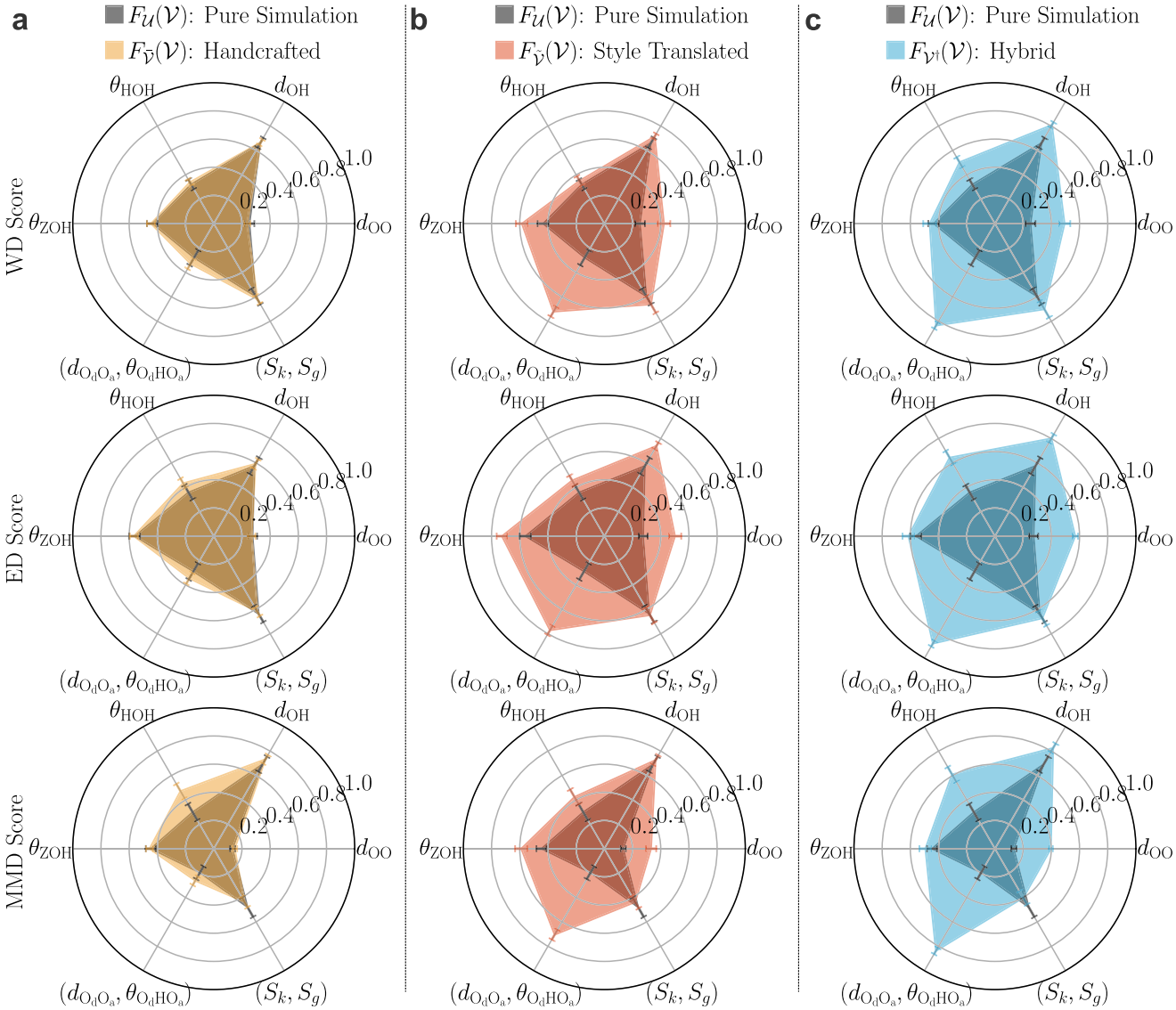


Fig. 7 | Performance evaluations based on the structural properties, using three distance metrics: Wasserstein distance, energy distance, and maximum mean discrepancy. Panels (**a**–**c**) show comparisons of models $F_{\tilde{\mathcal{V}}}$, $F_{\tilde{\mathcal{V}}}$ and $F_{\mathcal{V}^\dagger}$, respectively, each evaluated against the baseline model $F_{\mathcal{U}}$. Distances are computed between predicted distributions and reference theoretical distributions derived from top-layer water molecules.

performance gap between models trained on simulated and experimental data in the challenging scenario where the ground-truth atomic configurations are unavailable. To bridge this gap, we propose a data-driven, unpaired image-to-image style translation approach that significantly reduces the style discrepancy between simulated and experimental AFM images. Our approach relies only on unpaired samples from simulation and experiment domains, making it particularly well-suited for real-world scenarios where such paired data is impractical.

We demonstrate the effectiveness of this method using the example of water structure discovery on Au(111). By replacing the unavailable experimental training data with style-translated simulated AFM images, we show that structure discovery models can achieve significantly better performance on real experimental inputs. The style-translated dataset exposes the model to more realistic conditions, helping it focus on essential atomic features rather than noise and artefacts. These results support our hypothesis that reducing the style gap improves model performance on experimental AFM images. In addition, we propose an evaluation approach based on physically meaningful structural properties to address the lack of ground-truth atomic structures for experimental AFM images.

Although the examples presented in this work mainly focus on water molecule systems, neither the CycleGAN-based style translation nor the structure discovery model is, in principle, restricted to these systems. The discovery model was originally trained on a broad database of organic molecules and has been applied to a wide range of organic systems[13,18,48]. In addition, the style translation model was trained on diverse experimental AFM images, suggesting that some dominant differences between simulated and experimental images are largely system-independent. However, since the downstream demonstration in the present work focuses on bilayer water on Au(111), despite previous success of discovery approaches generally for a wide class of systems[13,18,20,] the broader applicability of the style translation augmentation to other molecular systems remains to be systematically validated in future studies.

The same consideration applies to experimental conditions. The present workflow should not be viewed as a universal pretrained tool for arbitrary AFM experiments, because the learned style depends on the experimental training images, including tip functionalization, imaging conditions, artefacts, and instrument drift. Although our training data include CO-tip AFM images from multiple sources, applications to substantially different tips, instruments, or molecular systems may require

retraining with representative experimental data. Therefore, the main contribution of this work is the proposed workflow for adapting simulated AFM data to experimental image conditions.

Overall, our work offers a practical pathway toward closing the simulation-to-experiment gap in AFM structure discovery. In this study, structural properties are used solely for evaluation purposes, but they are not used to guide model training. This raises an important question: can such structural properties be integrated as constraints during the model training to discourage physically implausible predictions? Another challenge lies in balancing robustness and sensitivity. While generalising across noisy experimental conditions improves stability, it may also lead to a loss of sensitivity to subtle atomic features. Understanding and balancing this trade-off is essential for developing high-fidelity models. Finally, we envision future structure discovery frameworks that can provide confidence scores for each predicted atom, helping us assess the trustworthiness of the model's outputs.

Methods

Binary classifier model and training details

The simulation-real machine expert was implemented as a convolutional neural network for binary image classification. Input images were resized to 192 × 192 pixels, and augmented during training using random horizontal and vertical flips, random 90° rotations, and additive noise. The network comprised two convolutional blocks with 32 and 64 filters, respectively, each followed by ReLU activation, batch normalization, and max pooling, after which the extracted features were flattened and passed to a single sigmoid output unit to produce a binary authenticity score. The model was trained to distinguish simulated PPAFM images from experimental AFM images using binary cross-entropy loss and stochastic gradient descent (learning rate 0.001) for 20 epochs with a batch size of 8. The training set comprised 608 simulated and 509 experimental images, and the test set comprised 207 simulated and 197 experimental images; the checkpoint with the highest test accuracy was retained for analysis. Evaluation on the held-out test set yielded an overall classification accuracy of 93.3%, with 377 of 404 images correctly assigned. Of the 207 simulated PPAFM images, 188 were correctly identified, whereas 19 were classified as AFM. Of the 197 experimental AFM images, 189 were correctly identified, whereas 8 were classified as PPAFM.

CycleGAN model and training details

The goal of each generator is to produce synthetic images that resemble real images in the target domain. The adversarial loss function of $G_{\mathcal{U}}$ and its discriminator $D_{\mathcal{V}}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\mathcal{U}\to\mathcal{V}}(G_{\mathcal{U}}, D_{\mathcal{V}}) &= \tfrac{1}{n}\sum_{i=1}^{n}\log D_{\mathcal{V}}(v_i) \\ &\quad + \tfrac{1}{m}\sum_{i=1}^{m}\log(1 - D_{\mathcal{V}}(\tilde{v}_i)) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\log D_{\mathcal{V}}(v_i) \\ &\quad + \tfrac{1}{m}\sum_{i=1}^{m}\log(1 - D_{\mathcal{V}}(G_{\mathcal{U}}(u_i))).\end{aligned} \tag{3}$$

Here, m, n are the batched sizes of samples from domains $\mathcal{U}$ and $\mathcal{V}$; $\tilde{v}_i = G_{\mathcal{U}}(u_i)$ is the synthetic output from simulation-to-experiment generator. The generator $G_{\mathcal{U}}$ is optimised to minimise the loss, while the discriminator $D_{\mathcal{V}}$ is trained to maximise it. Similarly, the adversarial loss for $G_{\mathcal{V}}$ and $D_{\mathcal{U}}$ is denoted as $\mathcal{L}_{\mathcal{V}\to\mathcal{U}}(G_{\mathcal{V}}, D_{\mathcal{U}})$. To further reduce the space of possible mapping and ensure that translation from one domain and back returns the original image, i.e., $u \to \tilde{v} \to \hat{u} \approx u$ and $v \to \tilde{u} \to \hat{v} \approx v$, CycleGAN introduces a cycle consistency loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{c}}(G_{\mathcal{U}}, G_{\mathcal{V}}) &= \tfrac{1}{m}\sum_{i=0}^{m}\| \hat{u}_i - u_i\|_1 \\ &\quad + \tfrac{1}{n}\sum_{i=0}^{n}\| \hat{v}_i - v_i\|_1,\end{aligned} \tag{4}$$

where $\|\cdot\|_1$ is the L^1 norm. To further constrain the generators and prevent unnecessary transformation of images that already resemble the target domain, we introduce an identity loss:

$$\begin{aligned}\mathcal{L}_{\mathrm{i}}(G_{\mathcal{U}}, G_{\mathcal{V}}) &= \tfrac{1}{n}\sum_{i=0}^{n}\left\|G_{\mathcal{U}}(v_i) - v_i\right\|_1 \\ &\quad + \tfrac{1}{m}\sum_{i=0}^{m}\left\|G_{\mathcal{V}}(u_i) - u_i\right\|_1.\end{aligned} \tag{5}$$

This loss encourages each generator to behave as an identity mapping when provided with inputs from the target domain. In other words, it requires the generator to recognise whether the input image already matches the target style; if so, no translation is needed; otherwise, a style translation is applied. This ensures that translations are applied only when needed, helping avoid over-modification. The total loss function used to train CycleGAN is:

$$\begin{aligned}\mathcal{L}\left(G_{\mathcal{U}}, G_{\mathcal{V}}, D_{\mathcal{V}}, D_{\mathcal{U}}\right) &= \mathcal{L}_{\mathcal{U}\to\mathcal{V}}(G_{\mathcal{U}}, D_{\mathcal{V}}) + \mathcal{L}_{\mathcal{V}\to\mathcal{U}}(G_{\mathcal{V}}, D_{\mathcal{U}}) \\ &\quad + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}}(G_{\mathcal{U}}, G_{\mathcal{V}}) + \lambda_{\mathrm{i}}\mathcal{L}_{\mathrm{i}}(G_{\mathcal{U}}, G_{\mathcal{V}}),\end{aligned} \tag{6}$$

where λ_{c} and λ_{i} are the weights controlling the contributions of cycle consistency and identity loss, respectively. The optimisation objective is then:

$$G_{\mathcal{U}}^{*}, G_{\mathcal{V}}^{*} = \arg\min_{G_{\mathcal{U}}, G_{\mathcal{V}}}\max_{D_{\mathcal{V}}, D_{\mathcal{U}}}\mathcal{L}\left(G_{\mathcal{U}}, G_{\mathcal{V}}, D_{\mathcal{V}}, D_{\mathcal{U}}\right). \tag{7}$$

ResNet-based generators with residual blocks are used to perform style translations. Each generator takes a single-channel input and outputs a single-channel image. The architecture consists of: (a) an initial 7 × 7 convolutional layer followed by instance normalisation and ReLU activation; (b) two downsampling layers (3 × 3 convolutions with stride 2); (c) 6 residual blocks, each containing two 3 × 3 convolutions with instance normalisation and skip connections; (d) two upsampling layers using transposed convolutions; and (e) a final 7 × 7 convolutional layer followed by a Tanh activation function. The number of base convolutional filters is set to 16. For each discriminator, a PatchGAN architecture is implemented as a CNN with two layers. The input is a single-channel AFM image, and the base number of filters is set to 16. The discriminator consists of a series of 4 × 4 convolutional layers with a stride of 2, followed by instance normalisation and LeakyReLU activation. This configuration allows the discriminator to operate on overlapping image patches, making local decisions about authenticity, which encourages the generator to produce realistic fine-grained textures. The final layer outputs a single-channel prediction map.

We trained the CycleGAN model using PyTorch[49]. The model was trained on greyscale AFM images resized to 192 × 192 pixels. We used the default settings of the CycleGAN framework and tuned the key loss weights, including the cycle-consistency loss weight and the identity loss weight. Training was performed for 200 epochs using the Adam optimiser with a learning rate of 0.0002.

Structure discovery model and training details

The structure discovery model follows our previous work closely[18,19]: there is a CNN that first predicts the positions of the atoms from the AFM image stack, and a graph neural network (GNN) that uses the predicted positions along with the AFM images to construct a labelled molecule graph. The training details can be found in the ref. 19. In this work, additional datasets containing experimental-style images were used in the training.

Metrics for distribution comparison

Wasserstein distance[34,50] (WD) is a measure of the dissimilarity between two distributions. It quantifies the minimum "effort" required to transform one distribution into another, where the effort is measured by the amount of probability mass that must be transported and the distance it must be moved. Mathematically, the Wasserstein distance between two distributions

X, Y is defined as follows:

$$\begin{aligned} \mathrm{WD}(X,Y) &= \inf_{\pi\in\Pi(X,Y)} \int_{\mathbb{R}\times\mathbb{R}} \| x-y \| \, \mathrm{d}\pi(x,y) \\ &= \inf_{\pi\in\Pi(X,Y)} \mathbb{E}_{(x,y)\sim\pi}\left[\| x-y \|\right]. \end{aligned} \tag{8}$$

Here, $\Pi(X, Y)$ denotes the set of all joint distributions $\pi(x, y)$ whose marginals are respectively X and Y. The joint distribution $\pi(x, y)$ specifies a transport plan indicating how much mass should be transported from x to y. The Wasserstein distance is then the minimum cost of the transport plan.

Fréchet Inception Distance (FID) was first introduced to evaluate the performance of GAN models. The FID between two image distributions X and Y is computed as follows. Collect image samples $x_1, \ldots, x_m, y_1, \ldots, y_n$ from X, Y, respectively. Encode all samples x_i and y_i by computing the activations $\mathrm{A}(x_i)$ and $\mathrm{A}(y_i)$ of the final layer of the pretrained Inception network[38]. Compute the sample means μ_1, μ_2 and the sample covariance matrices Σ_1, Σ_2 of the activations $\mathrm{A}(x_i)$, $\mathrm{A}(y_i)$. The FID is the WD between the two multivariate normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ [51].

$$\mathrm{FID}(X,Y) = \left\|\mu_1-\mu_2\right\|_2^2 + \mathrm{tr}(\Sigma_1) + \mathrm{tr}(\Sigma_2) - 2\cdot\mathrm{tr}\left(\sqrt{\Sigma_1\Sigma_2}\right) \tag{9}$$

Energy distance (ED) is a statistical distance used to measure the equality of distributions, whose name is derived from Newton's gravitational potential energy. The energy distance[52] between the d-dimensional independent random variables X and Y is defined as

$$\mathrm{ED}(X,Y) = 2\mathbb{E}\|X-Y\|_d - \mathbb{E}\left\|X-X'\right\|_d - \mathbb{E}\left\|Y-Y'\right\|_d, \tag{10}$$

where X' is an independent and identically distributed (iid) copy of X, Y' is an iid copy of Y, $\mathbb{E}$ is the expected value, and $\|\cdot\|$ denotes Euclidean norm.

Maximum mean discrepancy (MMD) is another distance measurement between random variables X and Y, which is defined as the distance between their embeddings in the reproducing kernel Hilbert space (RKHS)[53]. MMD quantifies the dissimilarity between two distributions by comparing their mean representations in a high-dimensional feature space. Given two probability distributions X and Y, the MMD between them is defined as follow:

$$\mathrm{MMD}(X,Y) = \left\|\mu_X-\mu_Y\right\|_{\mathcal{H}}, \tag{11}$$

where μ_X, μ_Y are the mean embeddings of X and Y in RKHS $\mathcal{H}$. We calculate the empirical estimation of MMD through

$$\begin{aligned} \mathrm{MMD}^2(X,Y) &= \left\|\frac{1}{n}\sum_{i=1}^{n}\varphi(x_i) - \frac{1}{m}\sum_{i=1}^{m}\varphi(y_i)\right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(y_i,y_j) \\ &\quad - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k(x_i,y_j), \end{aligned} \tag{12}$$

where kernel $k(x, y)$ is a function that measures the similarity between two data points x and y. We use the Gaussian kernel: $k(x,y) = \exp\left(-\|x-y\|^2/(2\sigma^2)\right)$, where σ is the bandwidth parameter.

AFM experiments

The experiments of water molecules as shown in Fig. 5 were performed on noncontact AFM/STM systems (Createc, Germany) at 5 K using qPlus sensors equipped with a tungsten (W) tip (spring constant k0 ≃ 1800 N/m, resonance frequencies of $f_0 \simeq 29.1$ kHz, and quality factor $Q \simeq 105$). All of the AFM frequency shift (Δf) images were obtained with the CO-terminated tips in constant-height mode. The CO-tip was obtained by positioning the tip over a CO molecule on the Au(111) surface at a set point of 100 mV and 10 pA, followed by increasing the current to 400 pA. The oscillation amplitude of experimental AFM imaging was 100 pm. The Au(111) single crystal was purchased from MaTeck, and the Au(111) surface was cleaned by repeated Ar^+ ion sputtering at 1 keV and annealing at about 700 K for multiple cycles. Ultrapure H_2O (Sigma-Aldrich, deuterium-depleted, 1 ppm) was used and further purified under vacuum by 3–5 freeze-and-pump cycles to remove remaining gas impurities. The water molecules were deposited in situ onto the surface held at 5 K via a dosing tube, then annealed at 77 K for 10 min.

Data availability

The training data for the style translation, machine expert, and structure discovery models are publicly available at https://doi.org/10.5281/zenodo.16828078.

Code availability

The code in this study is available at https://github.com/SINGROUP/StyleTransAugment.

Received: 9 January 2026; Accepted: 13 July 2026;
Published online: 27 July 2026

References

1. Giessibl, F. J. Advances in atomic force microscopy. *Rev. Mod. Phys.* **75**, 949–983 (2003).
2. Giessibl, F. J. Atomic force microscopy with qplus sensors. *MRS Bull.* **49**, 492–502 (2024).
3. Martin, Y., Williams, C. C. & Wickramasinghe, H. K. Atomic force microscope–force mapping and profiling on a sub 100-Å scale. *J. Appl. Phys.* **61**, 4723–4729 (1987).
4. García, R. & Pérez, R. Dynamic atomic force microscopy methods. *Surf. Sci. Rep.* **47**, 197–301 (2002).
5. Barth, C., Foster, A. S., Henry, C. R. & Shluger, A. L. Recent trends in surface characterization and chemistry with high-resolution scanning force methods. *Adv. Mater.* **23**, 477–501 (2010).
6. Gross, L. et al. Organic structure determination using atomic-resolution scanning probe microscopy. *Nat. Chem.* **2**, 821–825 (2010).
7. Albrecht, F., Pavliček, N., Herranz-Lancho, C., Ruben, M. & Repp, J. Characterization of a surface reaction by means of atomic force microscopy. *J. Am. Chem. Soc.* **137**, 7424–7428 (2015).
8. Hapala, P. et al. Mechanism of high-resolution STM/AFM imaging with functionalized tips. *Phys. Rev. B* **90**, 085421 (2014).
9. Hapala, P., Temirov, R., Tautz, F. S. & Jelínek, P. Origin of high-resolution IETS-STM images of organic molecules with functionalized tips. *Phys. Rev. Lett.* **113**, 226101 (2014).
10. Oinonen, N. et al. Advancing scanning probe microscopy simulations: A decade of development in probe-particle models. *Computer Phys. Commun.* **305**, 109341 (2024).
11. Heggemann, J., Ranawat, Y. S., Krejčí, O., Foster, A. S. & Rahe, P. Differences in molecular adsorption emanating from the (2 × 1) reconstruction of calcite(104). *J. Phys. Chem. Lett.* **14**, 1983–1989 (2023).
12. Cai, S., Jestilä, J. S., Liljeroth, P. & Foster, A. S. Direct imaging of chirality transfer induced by glycosidic bond stereochemistry in carbohydrate self-assemblies. *J. Am. Chem. Soc.* **147**, 9341–9351 (2025).
13. Alldritt, B. et al. Automated structure discovery in atomic force microscopy. *Sci. Adv.* **6**, eaay6913 (2020).
14. Tang, B. et al. Machine learning-aided atomic structure identification of interfacial ionic hydrates from AFM images. *Natl. Sci. Rev.* **10**, nwac282 (2023).
15. Tang, B. et al. Unveiling the Amorphous Ice Layer during Premelting Using AFM Integrating Machine Learning. *Phys. Rev. X* **15**, 041048 (2025).

16. Zhu, Z. et al. A deep-learning framework for the automated recognition of molecules in scanning-probe-microscopy images. *Angewandte Chemie Int. Ed.* **61**, e202213503 (2022).
17. Kurki, L., Oinonen, N. & Foster, A. S. Automated structure discovery for scanning tunneling microscopy. *ACS Nano* **18**, 11130–11138 (2024).
18. Oinonen, N., Kurki, L., Ilin, A. & Foster, A. S. Molecule graph reconstruction from atomic force microscope images with machine learning. *MRS Bull.* **47**, 895–905 (2022).
19. Priante, F. et al. Structure discovery in atomic force microscopy imaging of ice. *ACS Nano* **18**, 5546–5555 (2024).
20. González Lastre, M. et al. Molecular identification via molecular fingerprint extraction from atomic force microscopy images. *J. Cheminform.* **16**, 130 (2024).
21. Carracedo-Cosme, J., Romero-Muñiz, C., Pou, P. & Pérez, R. Molecular identification from AFM images using the IUPAC nomenclature and attribute multimodal recurrent neural networks. *ACS Appl. Mater. Interfaces* **15**, 22692–22704 (2023).
22. Carracedo-Cosme, J., Romero-Muñiz, C. & Pérez, R. A deep learning approach for molecular classification based on AFM images. *Nanomaterials* **11**, 1658 (2021).
23. Carracedo-Cosme, J. & Pérez, R. Molecular identification with atomic force microscopy and conditional generative adversarial networks. *npj Comput. Mater.* **10**, 19 (2024).
24. Khan, A., Lee, C.-H., Huang, P. Y. & Clark, B. K. Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images. *npj Comput. Mater.* **9**, 85 (2023).
25. Zhu, Z. et al. Automated generation and analysis of molecular images using generative artificial intelligence models. *J. Phys. Chem. Lett.* **15**, 1985–1992 (2024).
26. Huang, B., Li, Z. & Li, J. An artificial intelligence atomic force microscope enabled by machine learning. *Nanoscale* **10**, 21320–21326 (2018).
27. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017).
28. Goodfellow, I. J. et al. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. (eds.) *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014).
29. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (2017).
30. Liu, Q., Xu, J., Jiang, R. & Wong, W. H. Density estimation using deep generative neural networks. *Proc. Natl. Acad. Sci.* **118**, e2101344118 (2021).
31. Yao, Z., Su, J. & Yau, S.-T. Manifold fitting with CycleGAN. *Proc. Natl. Acad. Sci.* **121**, e2311436121 (2024).
32. Sandfort, V., Yan, K., Pickhardt, P. J. & Summers, R. M. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* **9**, 16884 (2019).
33. Wang, J., Wu, Q. J. & Pourpanah, F. DC-CycleGAN: Bidirectional CT-to-MR synthesis from unpaired data. *Computerized Med. Imaging Graph.* **108**, 102249 (2023).
34. Weng, L. From gan to wgan. lilianweng.github.io. https://lilianweng.github.io/posts/2017-08-20-gan/. (2017).
35. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN (2017). https://arxiv.org/abs/1701.07875. 1701.07875.
36. Herrmann, V. Wasserstein GAN and the Kantorovich-Rubinstein duality. vincentherrmann.github.io. https://vincentherrmann.github.io/blog/wasserstein/. (2017).
37. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium (2018). 1706.08500.
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision (2015). 1512.00567.
39. Russakovsky, O. et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Computer Vis. (IJCV)* **115**, 211–252 (2015).
40. Hong, J. et al. Imaging surface structure and premelting of ice ih with atomic resolution. *Nature* **630**, 375–380 (2024).
41. Shen, Y. R. & Ostroverkhov, V. Sum-frequency vibrational spectroscopy on water interfaces: Polar orientation of water molecules at interfaces. *Chem. Rev.* **106**, 1140–1154 (2006).
42. Tang, F. et al. Definition of free o–h groups of water at the air–water interface. *J. Chem. Theory Comput.* **14**, 357–364 (2017).
43. Du, X. et al. Revealing the molecular structures of α-al2o3(0001)–water interface by machine learning-based computational vibrational spectroscopy. *J. Chem. Phys.* **161**, 124702 (2024).
44. Offei-Danso, A., Hassanali, A. & Rodriguez, A. High-dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning. *J. Chem. Theory Comput.* **18**, 3136–3150 (2022).
45. Donkor, E. D., Laio, A. & Hassanali, A. Do machine-learning atomic descriptors and order parameters tell the same story? The case of liquid water. *J. Chem. Theory Comput.* **19**, 4596–4605 (2023).
46. Chau, P.-L. & Hardwick, A. J. A new order parameter for tetrahedral configurations. *Mol. Phys.* **93**, 511–518 (1998).
47. Duboué-Dijon, E. & Laage, D. Characterization of the local structure in liquid water by various order parameters. *J. Phys. Chem. B* **119**, 8406–8418 (2015).
48. Oinonen, N. et al. Electrostatic discovery atomic force microscopy. *ACS Nano* **16**, 89–97 (2022).
49. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8026-8037 (Curran Associates Inc., Red Hook, NY, USA, 2019).
50. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN (2017). 1701.07875.
51. Mathiasen, A. & Hvilshøj, F. Backpropagating through fréchet inception distance (2021). 2009.14075.
52. Székely, G. J. & Rizzo, M. L. Energy statistics: A class of statistics based on distances. *J. Stat. Plan. Inference* **143**, 1249–1272 (2013).
53. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–773 (2012).

Acknowledgements

This work was supported by the World Premier International Research Centre Initiative (WPI), MEXT, Japan, and by the Research Council of Finland (Projects 347319 and 346824). The authors acknowledge the computational resources provided by the Aalto Science-IT Project and CSC, Helsinki. The authors thank Peter Liljeroth, Benjamin Alldritt, and Shuning Cai for their efforts in acquiring and providing the experimental images used for style translation training. J.H. thanks Johannes Haataja for inspiring discussions and Nan Cao for valuable suggestions on data visualisations.

Author contributions

J.H. performed the style translations, conducted performance evaluations, and wrote the initial manuscript. N.O. built the structure discovery model and carried out the PPM simulations. F.P. generated the original simulation datasets for training the structure discovery model. F.F.C. analysed the results and contributed to the distributional distance calculations. L.K. contributed to data preprocessing and to the comparisons with the handcrafted perturbations. C.X. conducted the experiments for the AFM images used to train the style translation models. A.S.F. conceptualised the study and supervised the project. All authors discussed the results and added comments to the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-026-02243-2.

Correspondence and requests for materials should be addressed to Adam S. Foster.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026