

Supplementary Information for

Improving atomic force microscopy structure

discovery via style-translation

Jie Huang,[†] Niko Oinonen,[†] Fabio Priante,[¶] Filippo Federici Canova,[†] Lauri Kurki,[†] Chen Xu,[†] and Adam S. Foster^{*,†,§}

[†]*Department of Applied Physics, Aalto University, Helsinki, FI-02150, Finland*

[‡]*Nanolayers Research Computing Ltd., 51 New Way Road, London, NW9 6PL, United Kingdom*

[¶]*Department of Chemistry and Materials Science, Aalto University, Helsinki, FI-02150, Finland*

[§]*WPI Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kakuma-machi, Kanazawa, 920-1192, Japan*

E-mail: adam.foster@aalto.fi

Contents

Supplementary result 1: Simulations for the predicted structures	S3
Supplementary result 2: Distributional comparisons	S8
Supplementary result 3: Structure predictions from rotated AFM images	S16
Supplementary result 4: Table of distributional similarity scores	S19
Supplementary result 5: Structure predictions from experimental-style images	S21
Supplementary result 6: Evaluations of models on experimental-style images	S26
Supplementary result 7: Feature similarity and slice-to-slice comparisons	S28
Supplementary result 8: Reverse path workflow explorations	S35

Supplementary result 1: Simulations for the predicted structures

Figures S1 to S4 present the predicted atomic configurations and their corresponding PPM-simulated AFM images, obtained from four different structure discovery models.

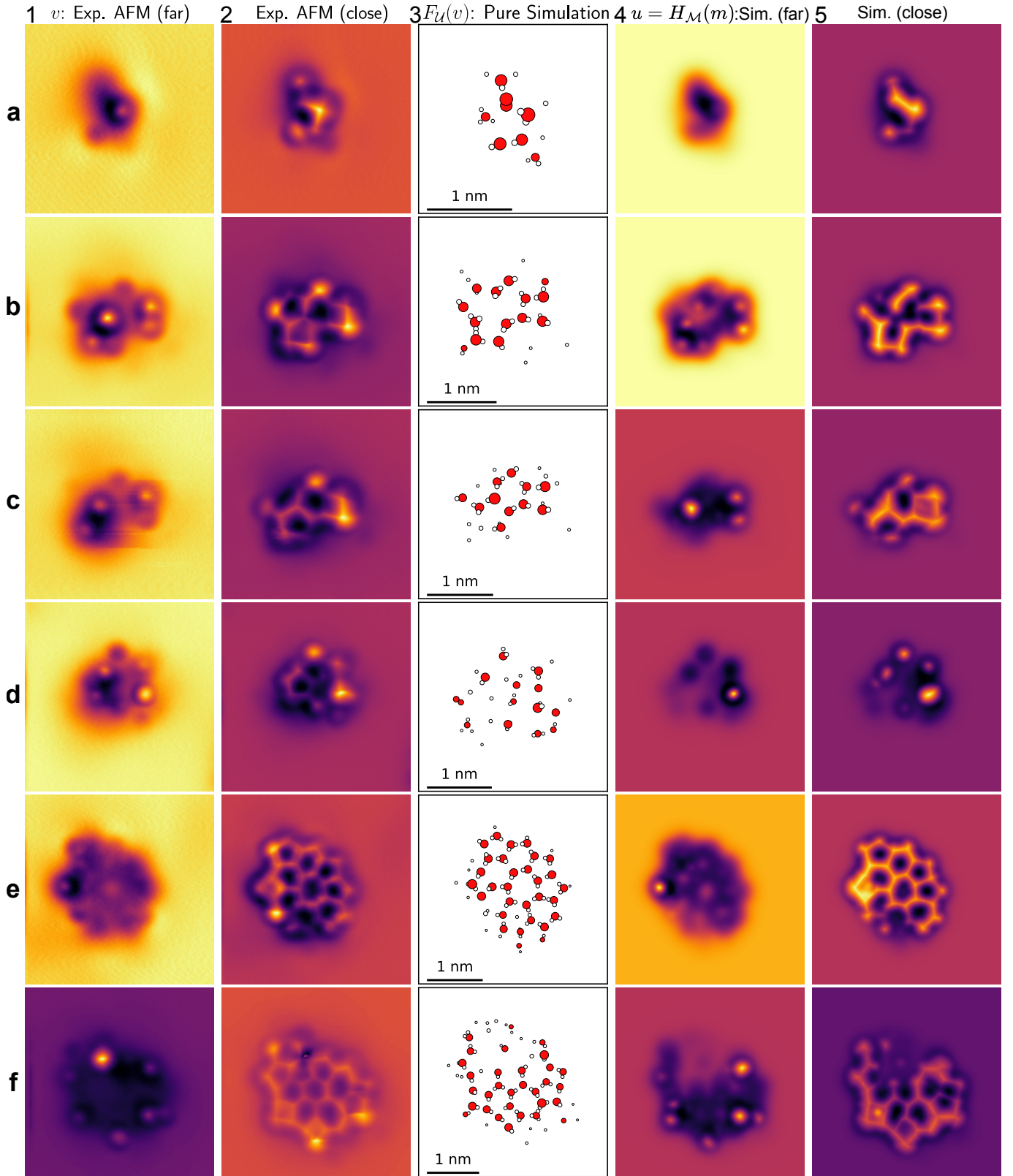


Figure S1: Structure predictions (3), and corresponding PPM-simulated AFM images (4-5) for experimental AFM inputs, using model F_U trained on pure simulated AFM data. a-f correspond to different samples.

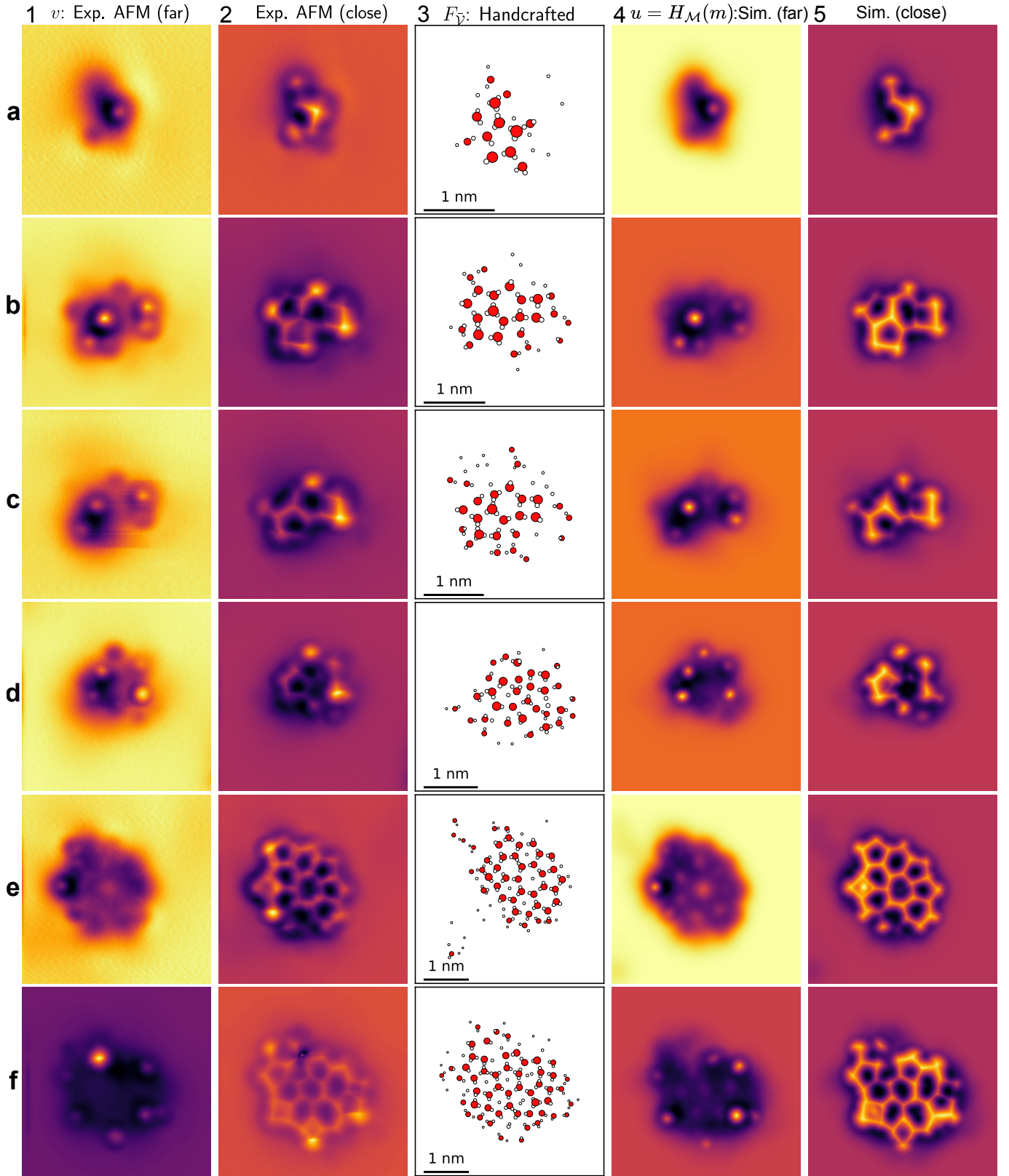


Figure S2: Structure predictions (3), and corresponding PPM-simulated AFM images (4-5) for experimental AFM inputs, using model $F_{\bar{v}}$ trained on images $\bar{\mathcal{V}}$ with handcrafted perturbations. a-f correspond to different samples.

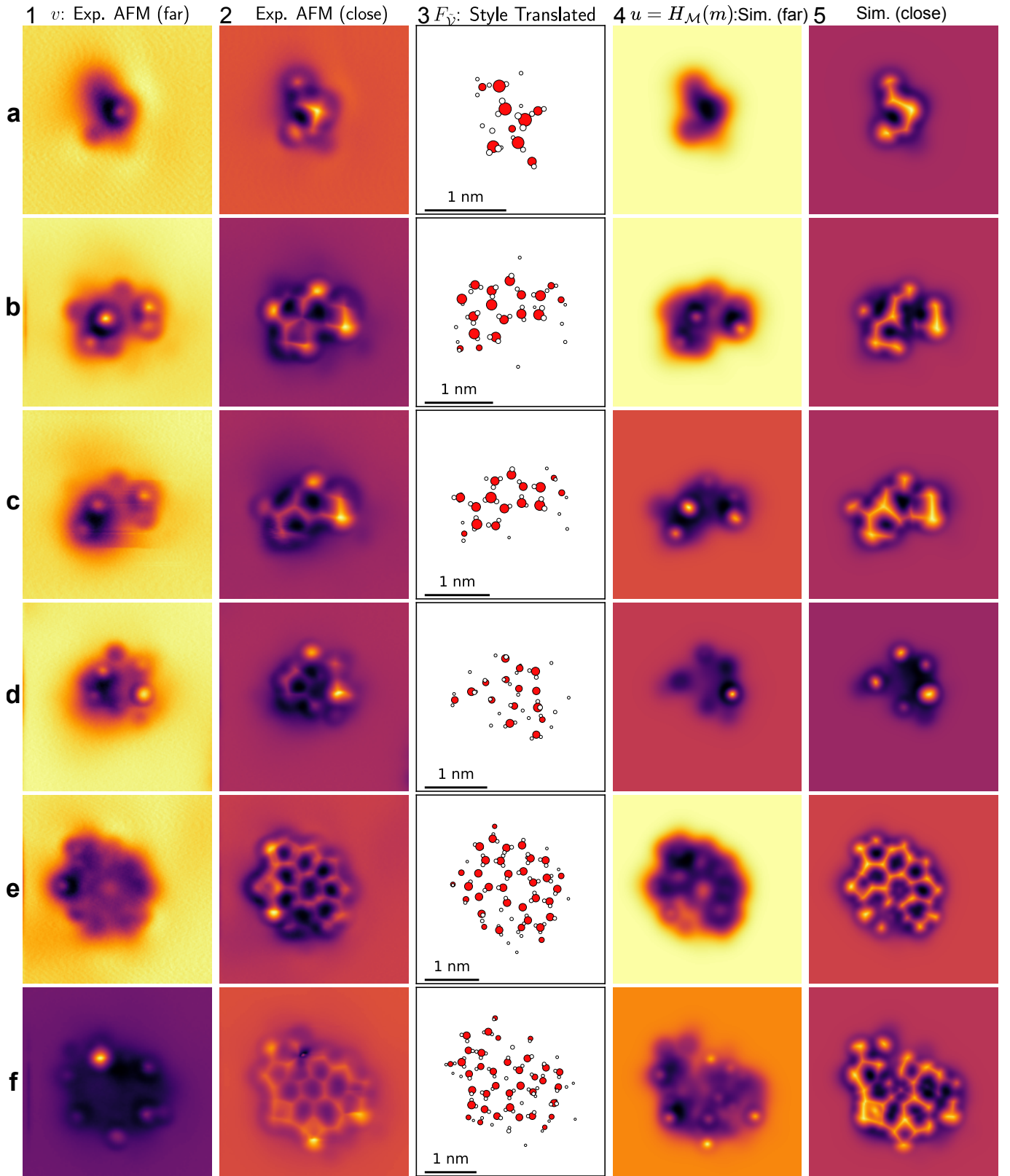


Figure S3: Structure predictions (3), and corresponding PPM-simulated AFM images (4-5) for experimental AFM inputs, using model $F_{\hat{V}}$ trained on style-translated data with $\lambda_c = 20$, $\lambda_i = 1$. a-f correspond to different samples.

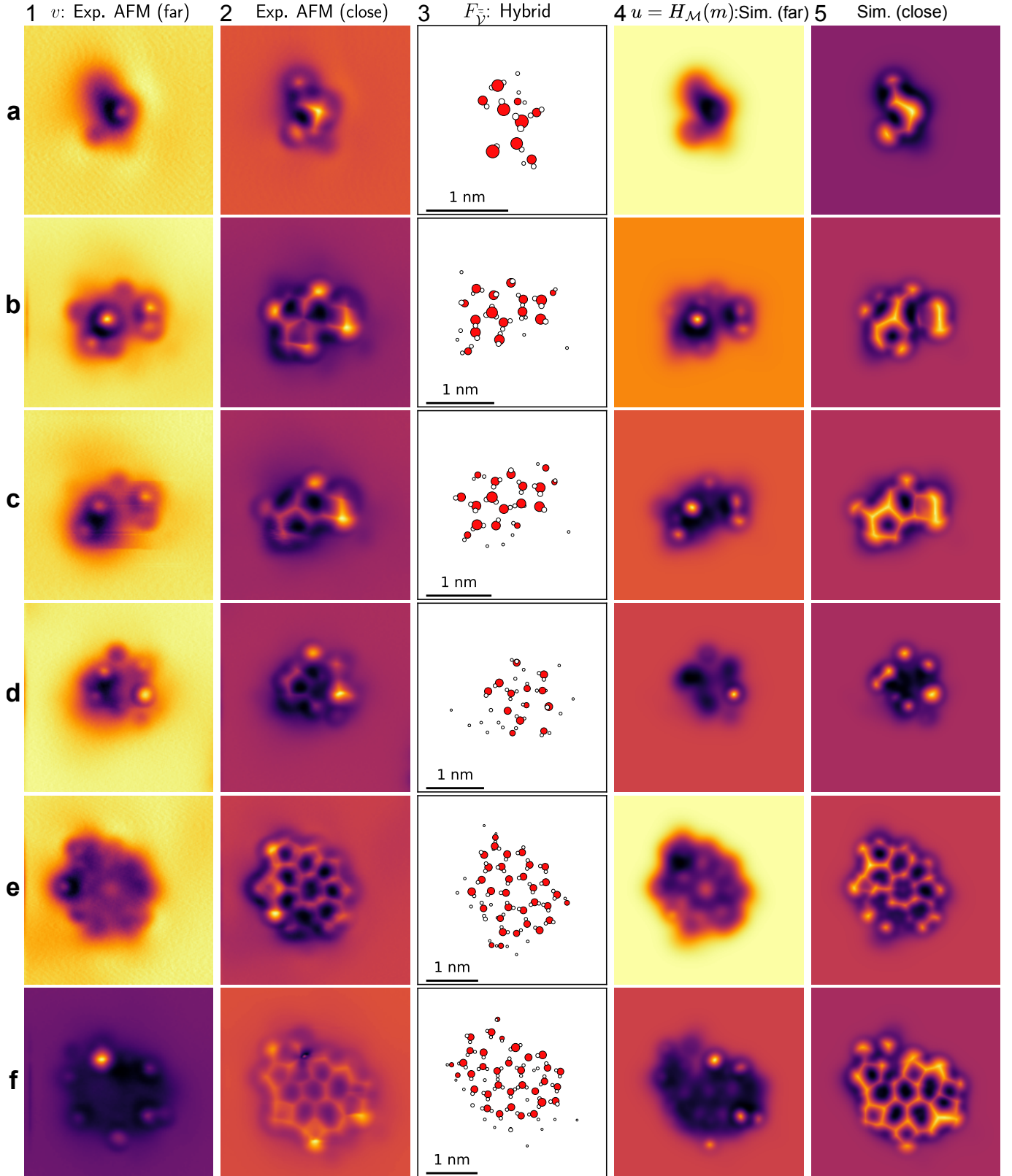


Figure S4: Structure predictions (3), and corresponding PPM-simulated AFM images (4-5) for experimental AFM inputs, using model $F_{\mathcal{V}^{\dagger}}$ trained on images $\mathcal{V}^{\dagger}$ with hybrid modifications combined with handcrafted perturbations and style translations with $\lambda_c = 20$, $\lambda_i = 1$. **a-f** correspond to different samples.

Supplementary result 2: Distributional comparisons

Figures S5 through S10 show the detailed comparisons of six local structural property distributions between predicted configurations and reference structures of water molecules on Au(111). In these figures, the distributions of predicted structures are ordered based on the Wasserstein distance to the theoretical distribution of top-layer water molecules $\mathcal{U}_{\text{top}}$, taking into account that the top layer of molecules is easier to predict by a structure discovery model compared to the underlying molecules. The theoretical distributions were computed from 74 water configurations sampled from the configuration set $\mathcal{M}$, whereas each predicted distribution was computed from 32 predicted configurations.

All structure discovery models share the same network architecture and training parameters. Two types of models are compared here: models $F_{\overline{\mathcal{V}}}$ trained on images $\overline{\mathcal{V}}$ with handcrafted perturbations and models $F_{\mathcal{V}^\dagger}$ trained on hybrid modification images $\mathcal{V}^\dagger$ that combine with handcrafted perturbations and style translations. For models $F_{\mathcal{V}^\dagger}$, we selected and show two sets of modules with parameters $\lambda_c, \lambda_i = 10, 10$ and $\lambda_c, \lambda_i = 20, 1$. In addition, we also show the distributions of the smallest and largest distance $\text{WD}(\cdot, \mathcal{U}_{\text{top}})$ in all our computational experiments. Since training dynamics can lead to variability in model performance, we train 10 independent replicas per configuration. The results shown in Fig. 7 represent the averaged performance over these replicas.

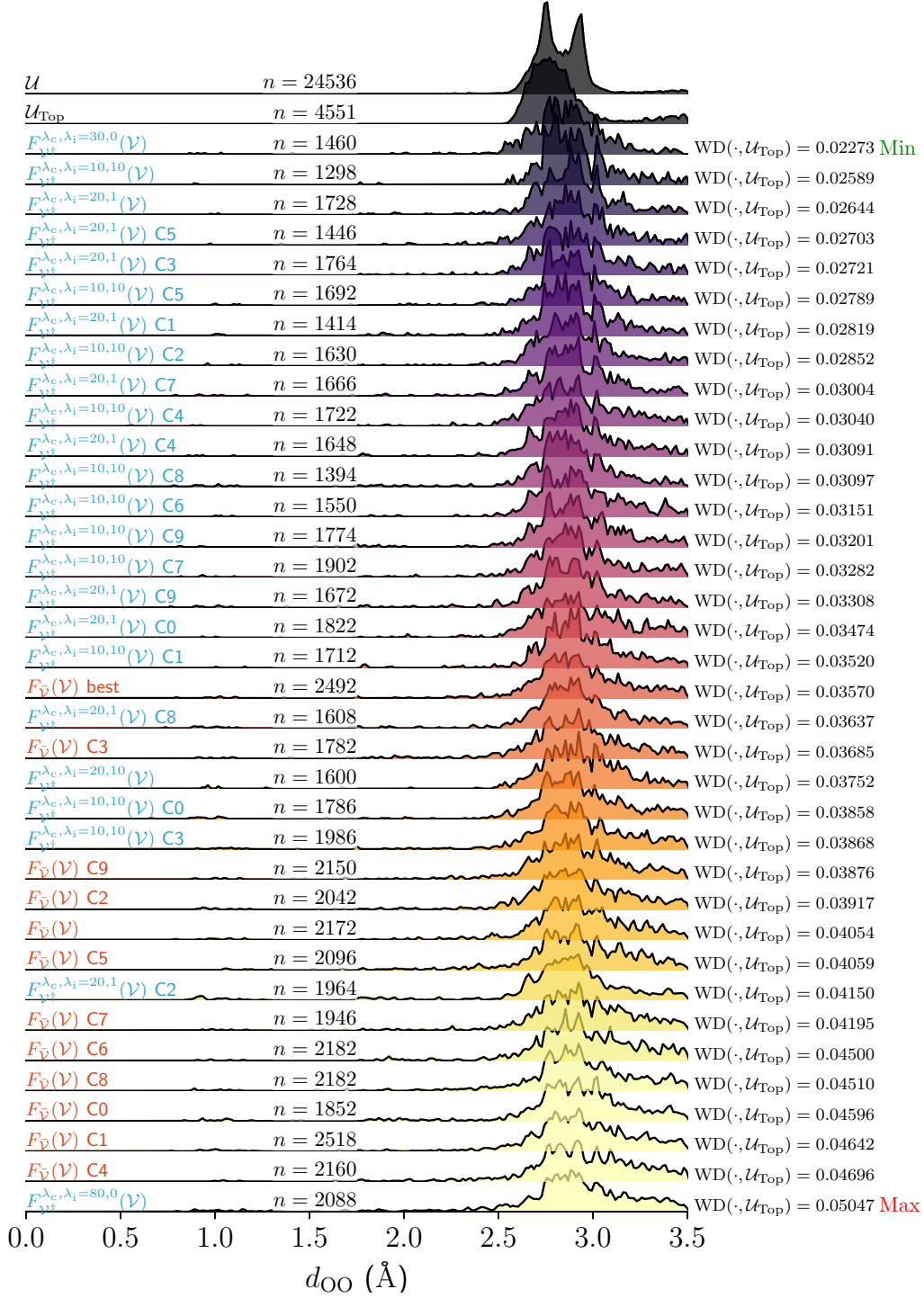


Figure S5: d_{OO} distributional comparisons between theoretical and predicted structures across different structure discovery models and datasets.

Figure S5 shows the distributional comparisons on d_{OO} . The valid range starts from 2.5 Å, and peaks around 2.75 Å according to $\mathcal{U}_{\text{top}}$. Hence, d_{OO} in the range $[0, 2.5]$ Å is considered unphysical. In general, models trained on $\mathcal{V}^\dagger$ outperform those trained on $\bar{\mathcal{V}}$. Similar trends are observed for d_{OH} (Fig. S6) and θ_{HOH} (Fig. S7).

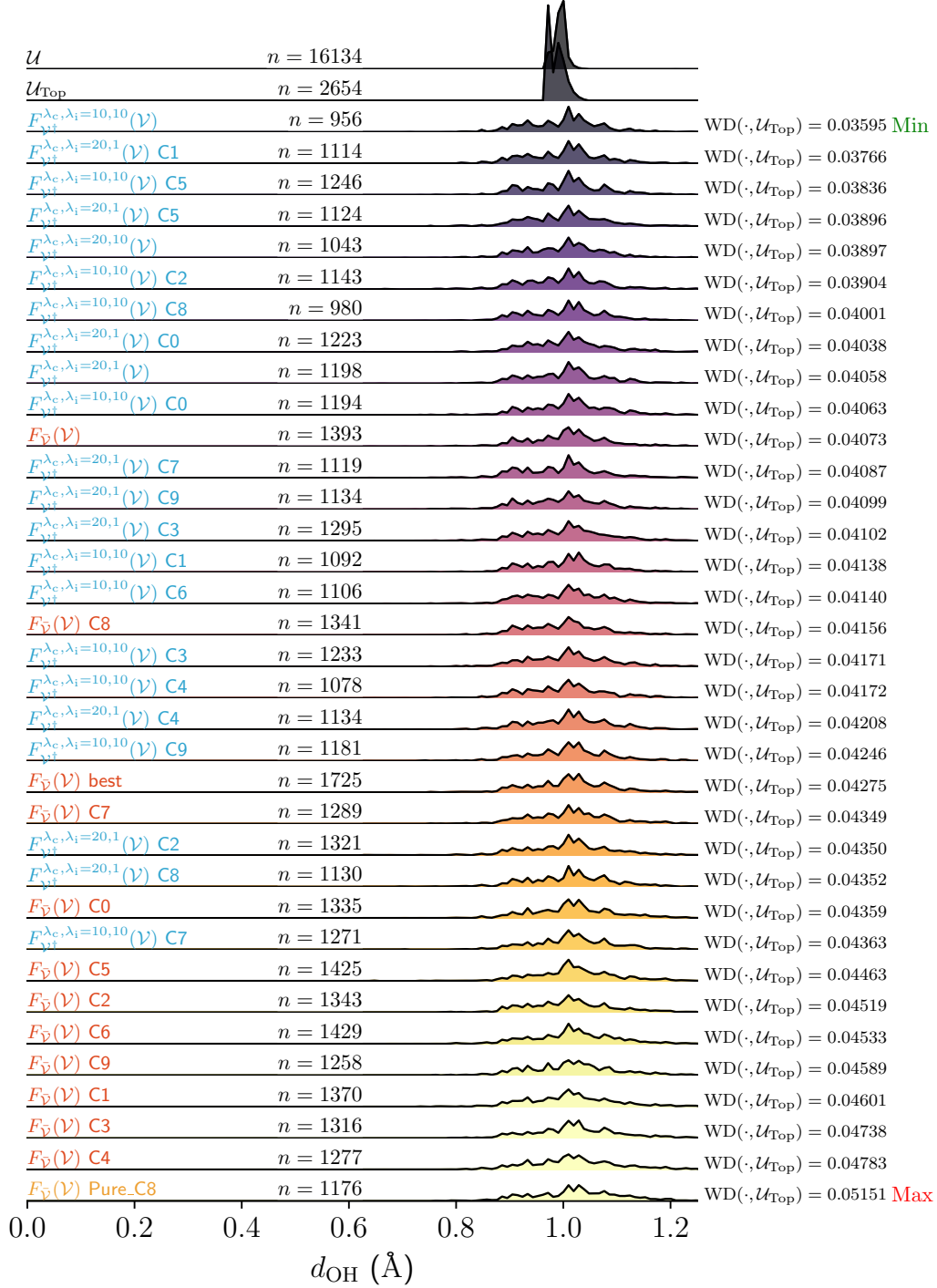


Figure S6: Distributional comparisons of d_{OH} between theoretical and predicted structures.

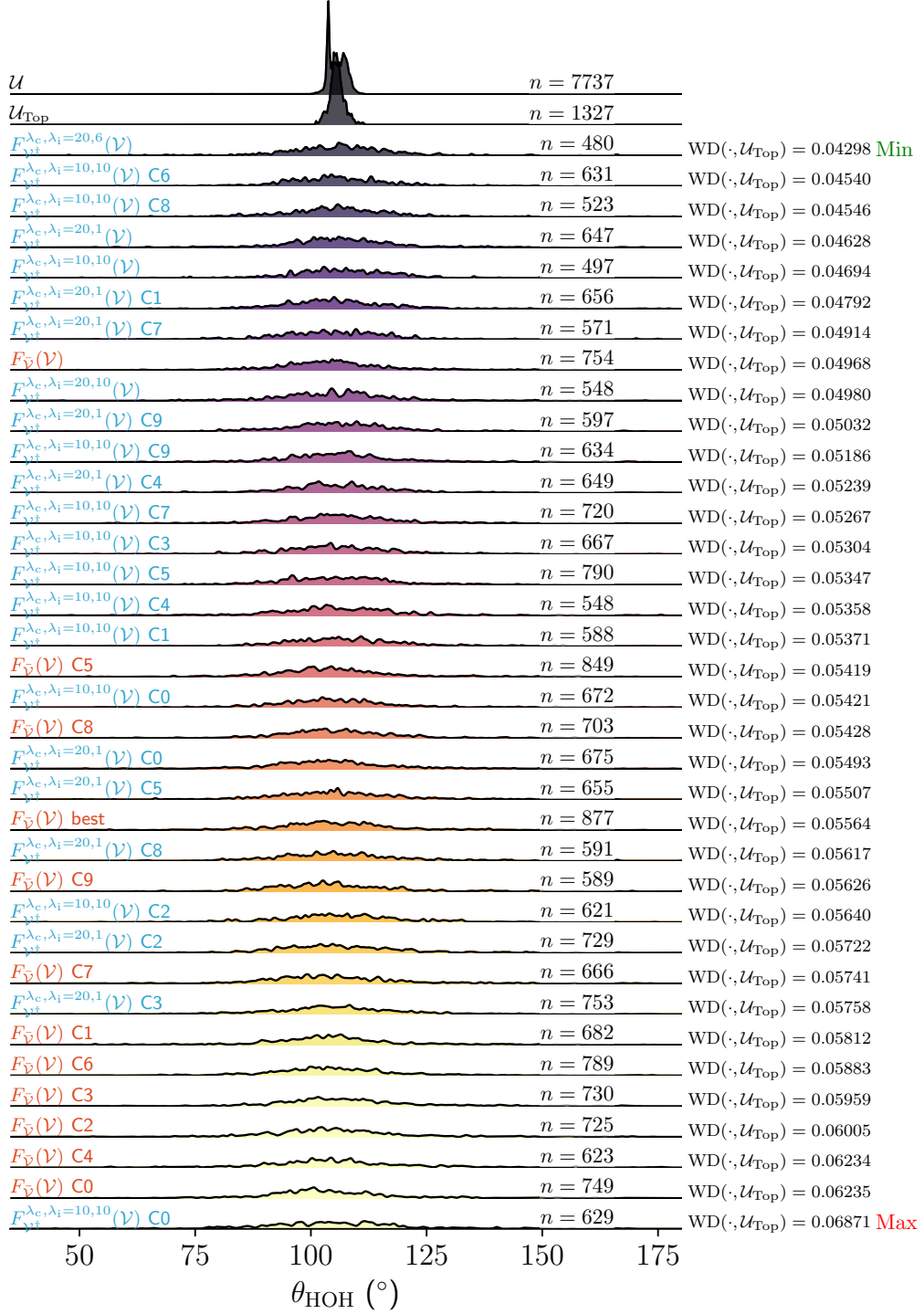


Figure S7: Distributional comparisons of θ_{HOH} between theoretical and predicted structures.

Figure S8 shows the distribution comparisons of the angle between the free OH bond and the surface normal. It’s worth noting that for the arising of the peaks around 170° , since the water molecules on the bottom layer are missing from the prediction, the OH bond pointing down becomes ‘free’, then contributing to the peak, while this is not the case in calculations from the theoretical structures. This metric sets a high standard, as it requires the model to correctly predict the molecules under the surface. When we put more experimental features into the images in the training data, we generalise a model to adapt to more real features, but it may also lose some sensitivity, making it harder for the underlying molecules to be predicted, since the subtle features below might be viewed as noise and then be ignored. Models trained on original simulated data tend to capture more of these subtle features, possibly due to their greater sensitivity to weak signals. This trade-off, between generalisation and sensitivity, is also evident in Fig. 5, where models trained on $\mathcal{U}$ recover more low-lying hydrogen atoms. Hence, style-translated training improves generalisation but may sacrifice sensitivity, suggesting that future models must carefully balance these two aspects. Confidence scoring for individual atomic predictions would also be beneficial for better interpretation.

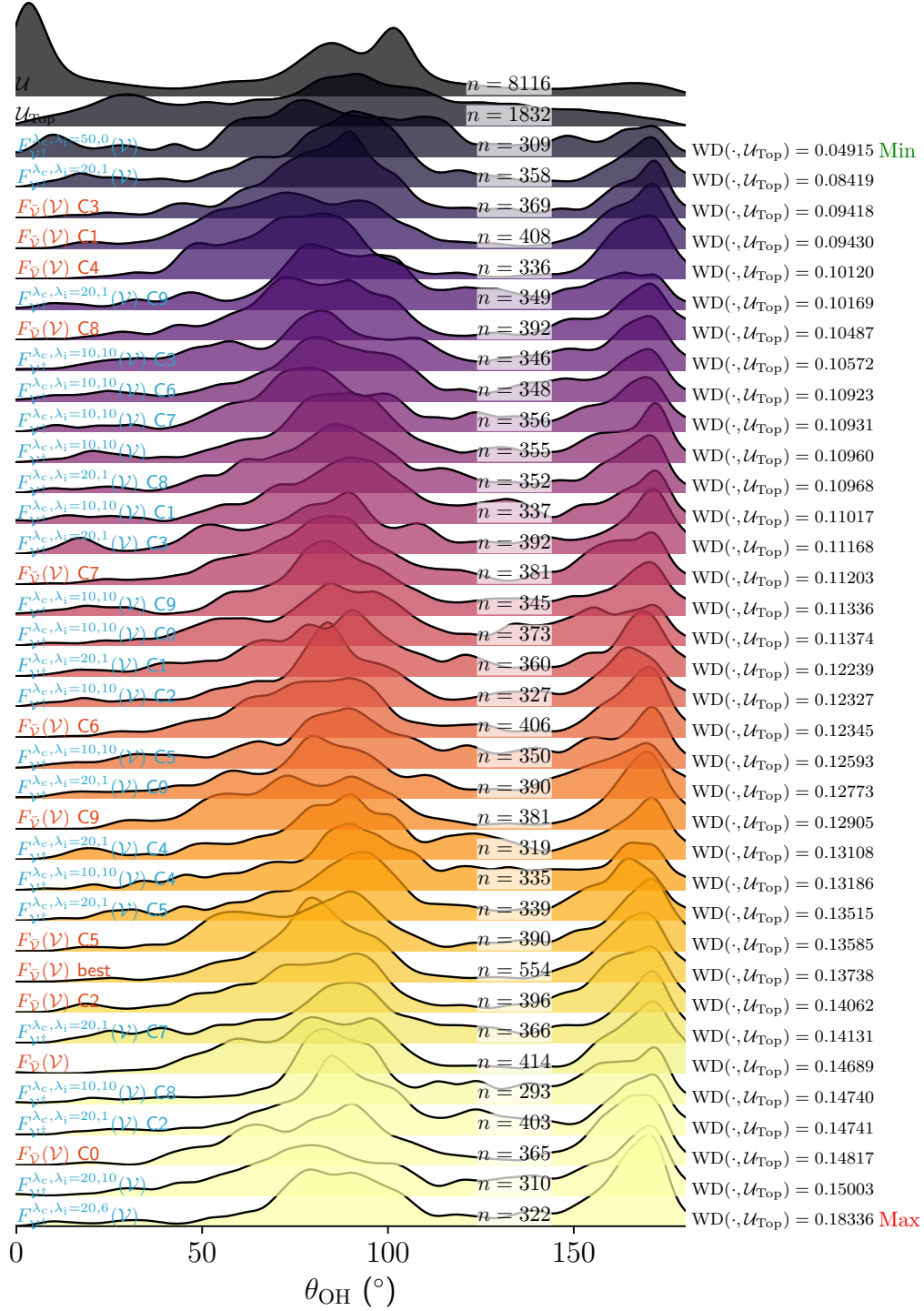


Figure S8: Distributional comparisons of the angle θ_{ZOH} between the free OH bond and the surface normal (z direction).

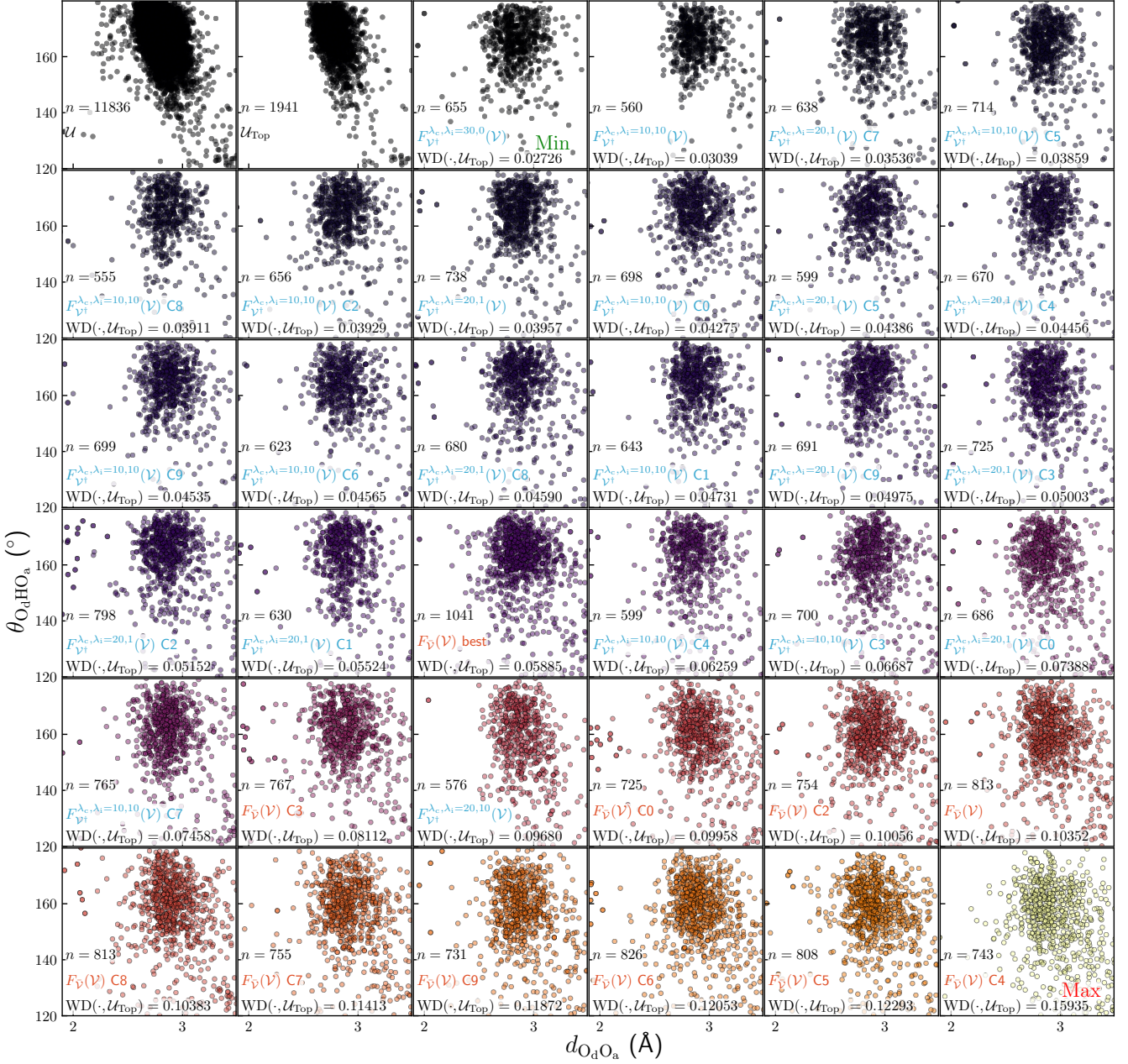


Figure S9: Comparisons of joint distributions of the donor–acceptor oxygen distance $d_{\text{O}_d\text{O}_a}$ and the hydrogen bond angle $\theta_{\text{O}_d\text{HO}_a}$.

Figure S9 shows the joint distributions of the donor–acceptor oxygen distance $d_{\text{O}_d\text{O}_a}$ and the hydrogen bond angle $\theta_{\text{O}_d\text{HO}_a}$. It is clear to see that the points obtained from the $F_{V^\dagger}$ models are more concentrated compared to the $F_{V^\ddagger}$ models, indicating better performance in this metric.

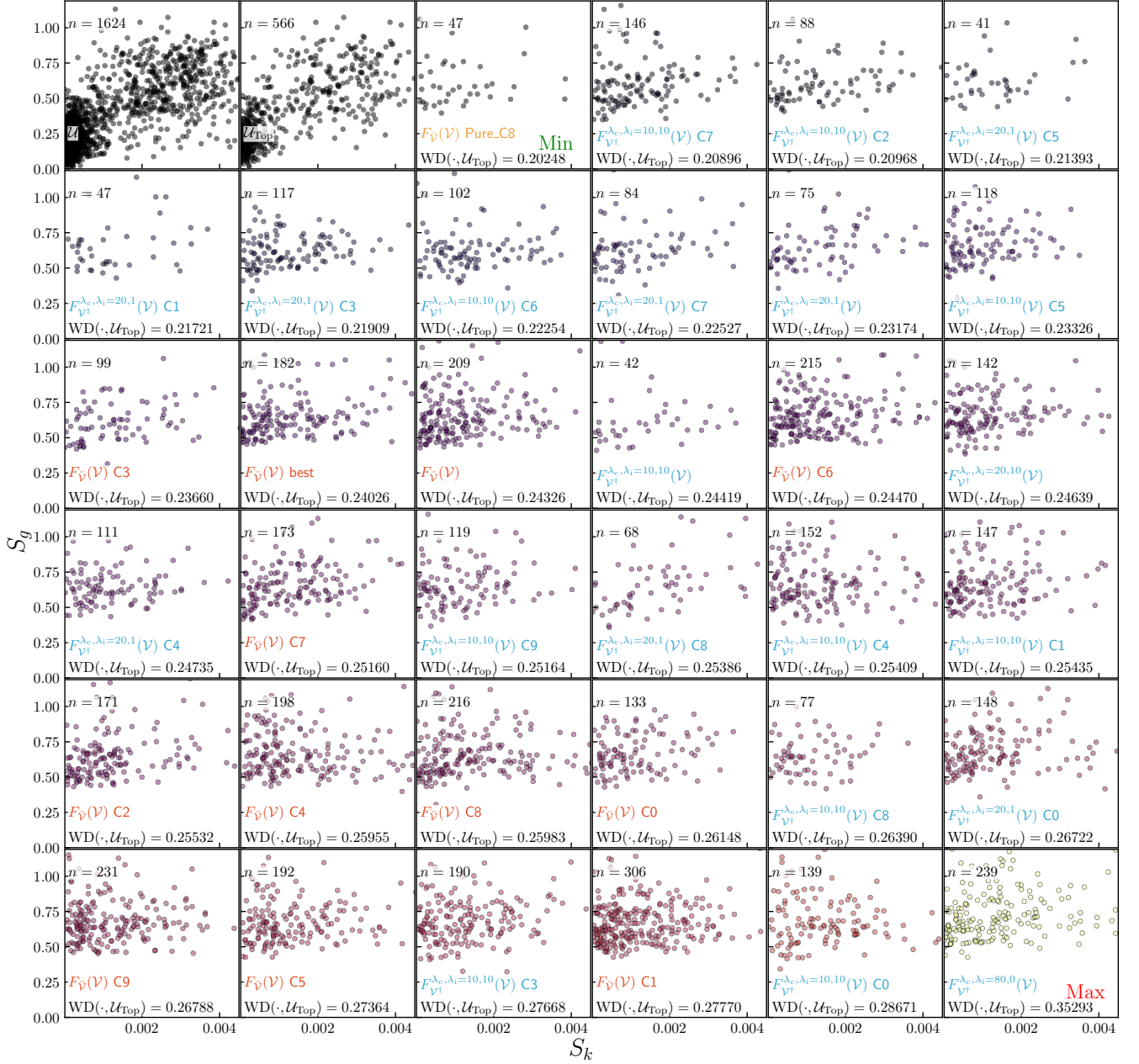


Figure S10: Comparisons of the joint distributions of translational (S_k) and orientational (S_g) tetrahedral order parameters.

Figure S10 shows the comparisons of the joint distributions of translational (S_k) and orientational (S_g) tetrahedral order parameters. Models trained on $\mathcal{V}^\dagger$ do not significantly outperform those trained on $\bar{\mathcal{V}}$. One reason for that is the generalisation increase with the sacrifice of sensitivity. Another reason is the lower accuracy of distribution estimation, which may be attributed to the limited local environments found in the predicted configurations from the real AFM experiment.

Supplementary result 3: Structure predictions from rotated AFM images

Figures S11-S13 show the configuration predictions when rotating the input AFM images.

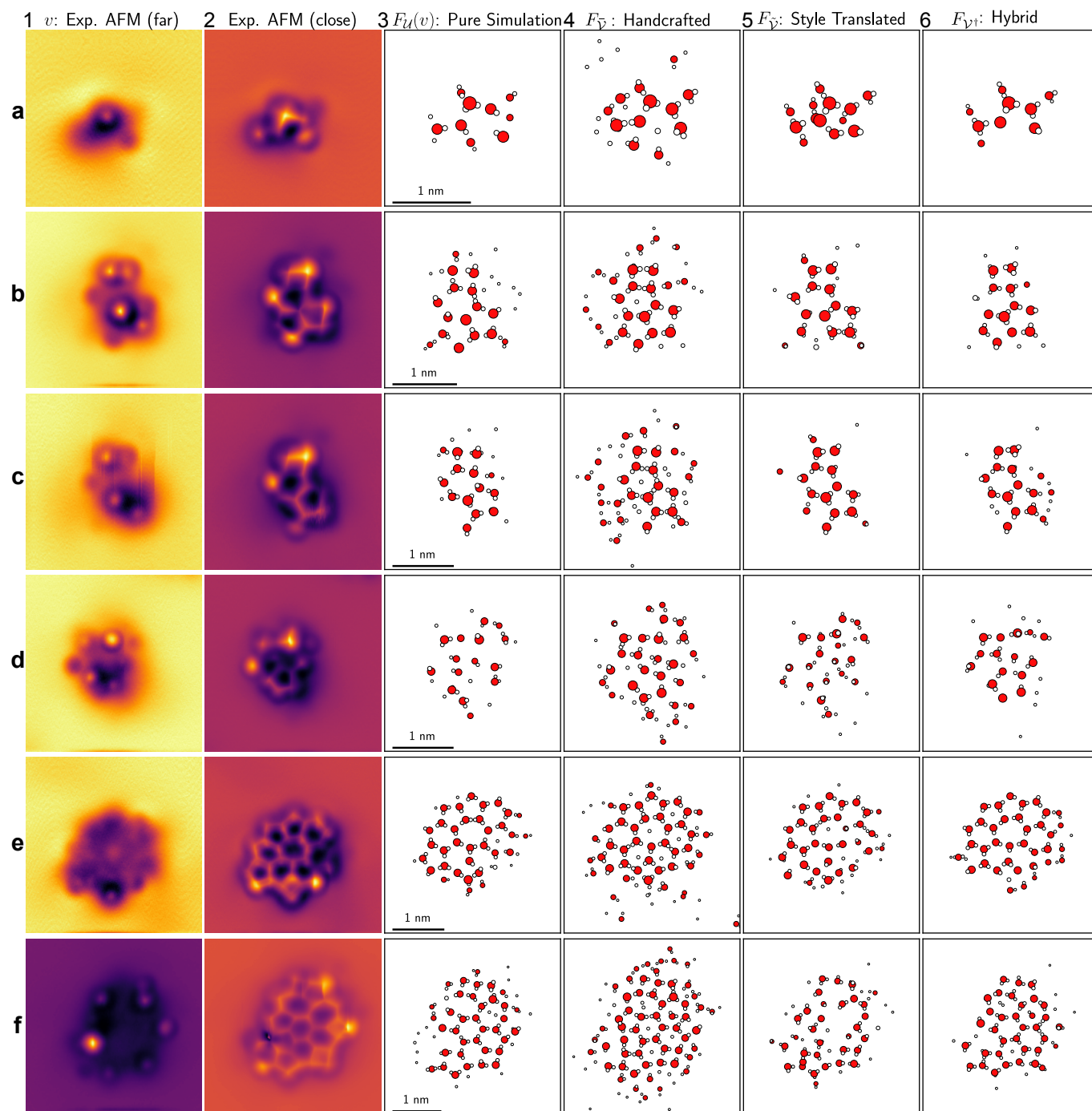


Figure S11: Atomic configuration predictions from 90°-rotated experimental AFM images using different structure discovery models.

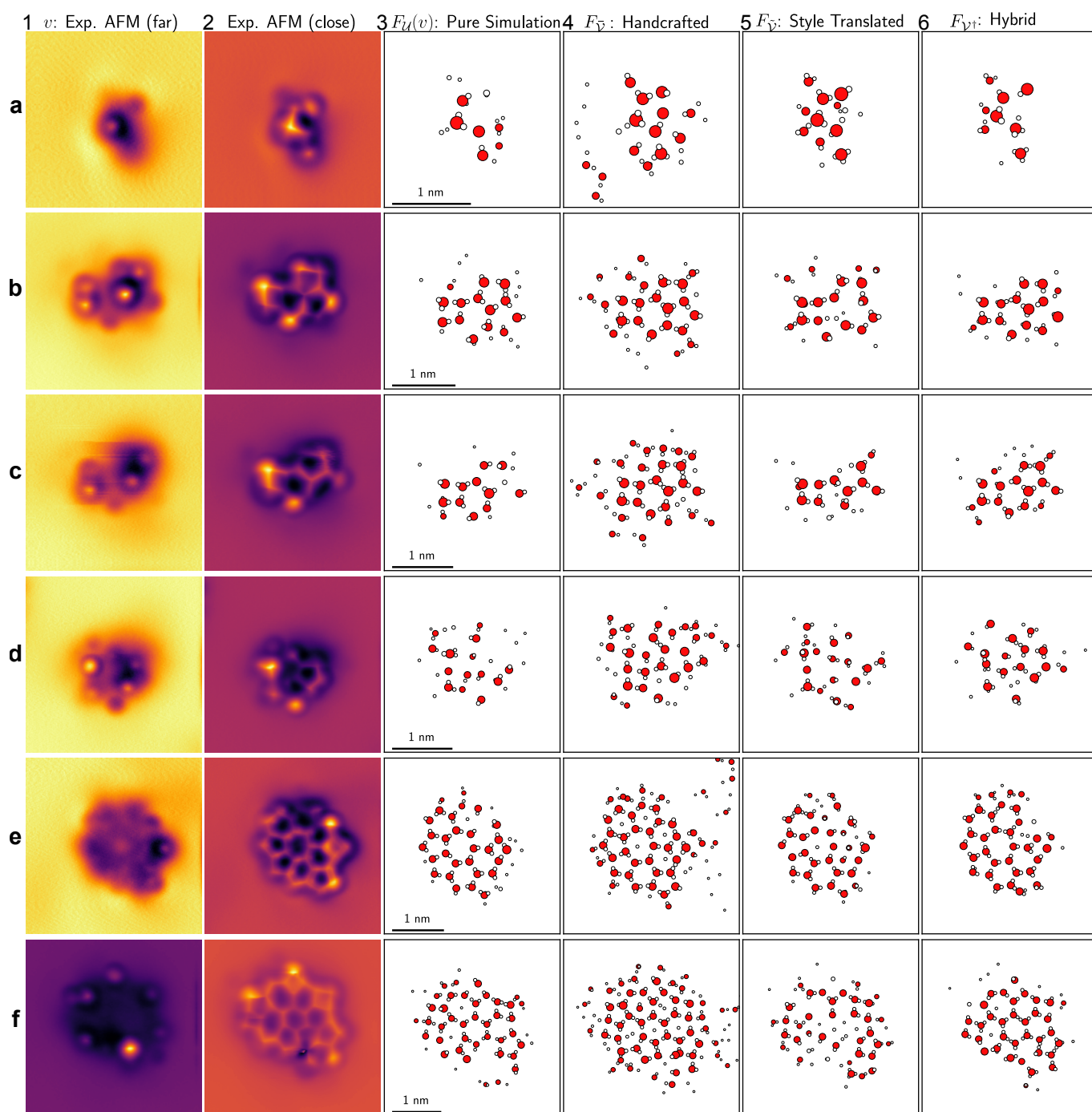


Figure S12: Atomic configuration predictions from 180°-rotated experimental AFM images using different structure discovery models.

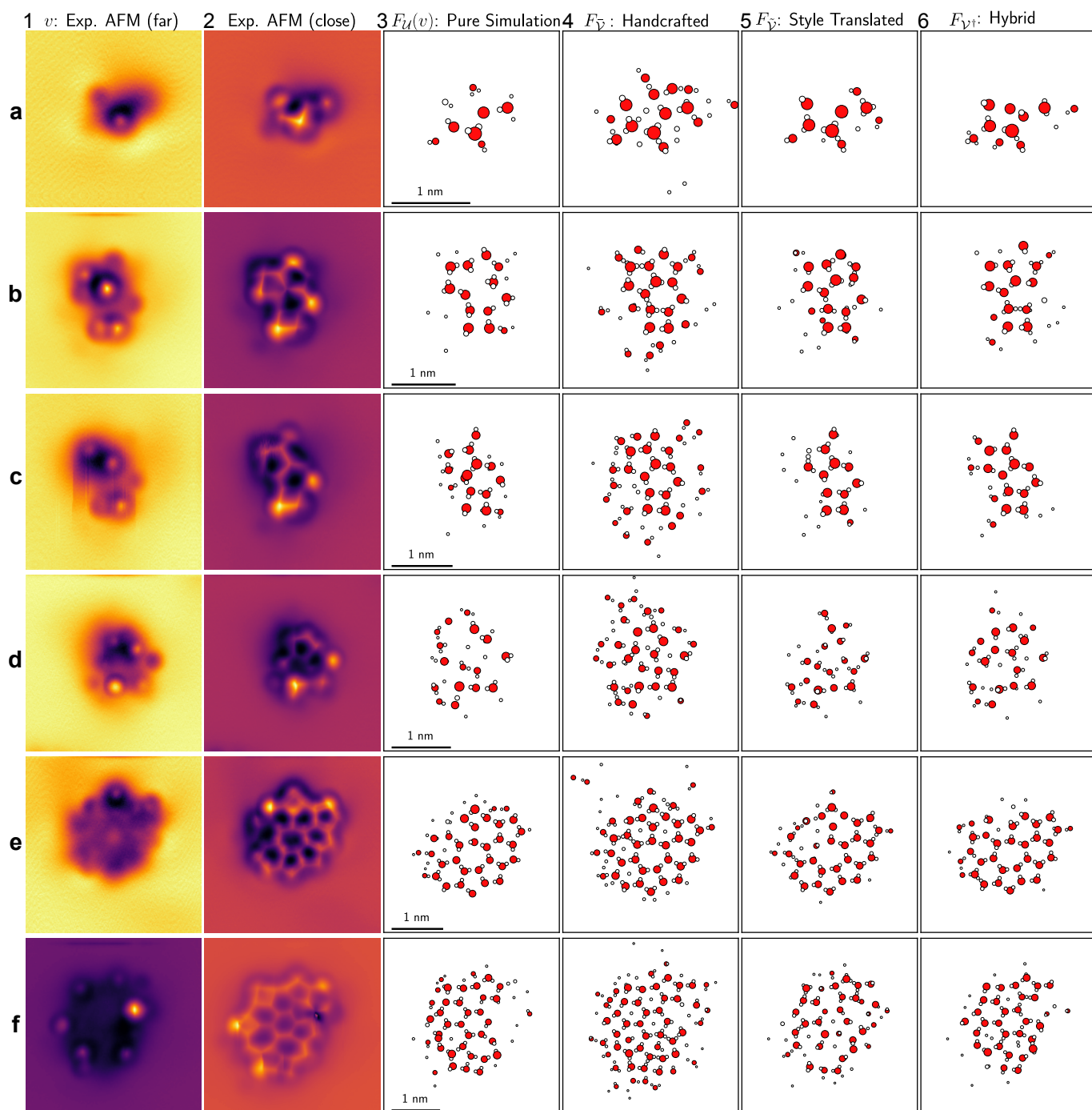


Figure S13: Atomic configuration predictions from 270°-rotated experimental AFM images using different structure discovery models.

Supplementary result 4: Table of distributional similarity scores

Table S1: WD, ED, and MMD scores for different structure discovery models, reported as mean $\pm$ standard error.

(a) Wasserstein-distance-based scores									
Model	Category	Replicas	d_{OO}	d_{OH}	θ_{HOH}	θ_{ZOH}	$(d_{O_d O_a}, \theta_{O_d HO_a})$	(S_k, S_g)	
Pure	$F_L(V)$	11	0.253171 $\pm$ 0.034706	0.662503 $\pm$ 0.040391	0.321435 $\pm$ 0.039361	0.434786 $\pm$ 0.038833	0.279645 $\pm$ 0.061220	0.597136 $\pm$ 0.063515	
Handcrafted	$F_V(V)$	11	0.208591 $\pm$ 0.027849	0.666324 $\pm$ 0.025474	0.350691 $\pm$ 0.035986	0.464916 $\pm$ 0.046306	0.339818 $\pm$ 0.047413	0.634877 $\pm$ 0.025700	
Style translated (L20L1)	$F_V(V)$	11	0.423275 $\pm$ 0.044712	0.715308 $\pm$ 0.020992	0.351592 $\pm$ 0.036873	0.585236 $\pm$ 0.042474	0.721796 $\pm$ 0.039371	0.662078 $\pm$ 0.068793	
Style translated (L10L10)	$F_V(V)$	11	0.393965 $\pm$ 0.046788	0.734885 $\pm$ 0.025011	0.480477 $\pm$ 0.051794	0.433153 $\pm$ 0.065214	0.705120 $\pm$ 0.040434	0.602500 $\pm$ 0.061902	
Hybrid (L20L1)	$F_V^+(V)$	10	0.494126 $\pm$ 0.040290	0.808963 $\pm$ 0.021906	0.503203 $\pm$ 0.040519	0.462922 $\pm$ 0.045801	0.835652 $\pm$ 0.025194	0.702949 $\pm$ 0.066932	
Hybrid (L10L10)	$F_V^+(V)$	11	0.481307 $\pm$ 0.032377	0.823615 $\pm$ 0.024620	0.540331 $\pm$ 0.035830	0.485921 $\pm$ 0.028606	0.839894 $\pm$ 0.031193	0.710749 $\pm$ 0.051066	
(b) Energy-distance-based scores									
Model	Category	Replicas	d_{OO}	d_{OH}	θ_{HOH}	θ_{ZOH}	$(d_{O_d O_a}, \theta_{O_d HO_a})$	(S_k, S_g)	
Pure	$F_L(V)$	11	0.275599 $\pm$ 0.031741	0.576532 $\pm$ 0.066782	0.359159 $\pm$ 0.058150	0.561244 $\pm$ 0.039572	0.289474 $\pm$ 0.066207	0.637434 $\pm$ 0.066332	
Handcrafted	$F_V(V)$	11	0.250668 $\pm$ 0.044479	0.600431 $\pm$ 0.042217	0.498077 $\pm$ 0.044695	0.582750 $\pm$ 0.043586	0.362176 $\pm$ 0.050520	0.589816 $\pm$ 0.032142	
Style translated (L20L1)	$F_V(V)$	11	0.501403 $\pm$ 0.044059	0.739899 $\pm$ 0.026731	0.456786 $\pm$ 0.039028	0.726704 $\pm$ 0.037023	0.771469 $\pm$ 0.037032	0.647663 $\pm$ 0.062463	
Style translated (L10L10)	$F_V(V)$	11	0.433481 $\pm$ 0.052406	0.703684 $\pm$ 0.036449	0.575560 $\pm$ 0.059645	0.564036 $\pm$ 0.066593	0.735751 $\pm$ 0.039150	0.595637 $\pm$ 0.068111	
Hybrid (L20L1)	$F_V^+(V)$	10	0.569767 $\pm$ 0.019648	0.801374 $\pm$ 0.033699	0.643962 $\pm$ 0.053991	0.613493 $\pm$ 0.041923	0.880770 $\pm$ 0.023119	0.673267 $\pm$ 0.058428	
Hybrid (L10L10)	$F_V^+(V)$	11	0.512158 $\pm$ 0.033083	0.796943 $\pm$ 0.032596	0.690493 $\pm$ 0.045397	0.641599 $\pm$ 0.029290	0.866532 $\pm$ 0.027714	0.695373 $\pm$ 0.056797	
(c) MMD-based scores									
Model	Category	Replicas	d_{OO}	d_{OH}	θ_{HOH}	θ_{ZOH}	$(d_{O_d O_a}, \theta_{O_d HO_a})$	(S_k, S_g)	
Pure	$F_L(V)$	11	0.134016 $\pm$ 0.018592	0.693685 $\pm$ 0.061712	0.299984 $\pm$ 0.065944	0.446540 $\pm$ 0.035260	0.196940 $\pm$ 0.052068	0.480083 $\pm$ 0.079647	
Handcrafted	$F_V(V)$	11	0.149388 $\pm$ 0.027545	0.760671 $\pm$ 0.036710	0.615092 $\pm$ 0.072449	0.471537 $\pm$ 0.039159	0.289909 $\pm$ 0.031952	0.384053 $\pm$ 0.038761	
Style translated (L20L1)	$F_V(V)$	11	0.331504 $\pm$ 0.036543	0.727202 $\pm$ 0.038545	0.424245 $\pm$ 0.060947	0.589496 $\pm$ 0.043429	0.704940 $\pm$ 0.040222	0.437882 $\pm$ 0.050752	
Style translated (L10L10)	$F_V(V)$	11	0.300993 $\pm$ 0.040128	0.719908 $\pm$ 0.031362	0.495177 $\pm$ 0.063986	0.454049 $\pm$ 0.068748	0.640641 $\pm$ 0.045170	0.425757 $\pm$ 0.078091	
Hybrid (L20L1)	$F_V^+(V)$	10	0.397052 $\pm$ 0.013552	0.826062 $\pm$ 0.035707	0.608789 $\pm$ 0.060362	0.493610 $\pm$ 0.042111	0.837969 $\pm$ 0.025832	0.403592 $\pm$ 0.054114	
Hybrid (L10L10)	$F_V^+(V)$	11	0.377345 $\pm$ 0.024576	0.825804 $\pm$ 0.035850	0.672892 $\pm$ 0.063560	0.515562 $\pm$ 0.029586	0.799642 $\pm$ 0.033428	0.472028 $\pm$ 0.057574	

Supplementary result 5: Structure predictions from experimental-style images

Figures S14-S17 show four examples of atomic configuration predictions from style-translated AFM images using different structure discovery models. For the experimental data, a direct sample-by-sample comparison is not possible because the corresponding atomic labels are unavailable. We therefore used an indirect method based on the comparison of local structure distributions. Here, as supplementary validation of the structure discovery models, we present structural predictions for style-translated images with known labels obtained from DFT calculations, using them as proxies for real experimental data.

Specifically, the original simulated AFM image u , shown in Figures S14-S17a, is first generated using PPM. This image is then passed through the forward style translator to obtain the experimental-style image $\tilde{v}$, which serves as a proxy for a real experimental AFM image with a known underlying atomic structure. The translated image $\tilde{v}$ is subsequently used as input to different structure discovery models. The corresponding predictions are shown in Figures S14-S17c, together with the atomic reference structure m .

From the comparisons, we observe that the models trained on experimental style data, namely $F_{\tilde{\mathcal{V}}}$, $F_{\mathcal{V}^+}$, produce structures that are closer to the reference m than the models only training on the simulated images, namely $F_{\mathcal{U}}$, $F_{\overline{\mathcal{V}}}$. This result is expected to some extent, since the test images follow a distribution similar to that of the training data for the models trained on experimental-style images. Nevertheless, it provides an additional perspective to verify the model performance. The observed improvement across different augmentation strategies is also consistent with the distributional metrics: handcrafted perturbations are beneficial, but they are not sufficiently robust to capture the complexity of experimental noise. In contrast, experimental style translation provides an effective route toward improving structure prediction performance for AFM images obtained from real experiments.

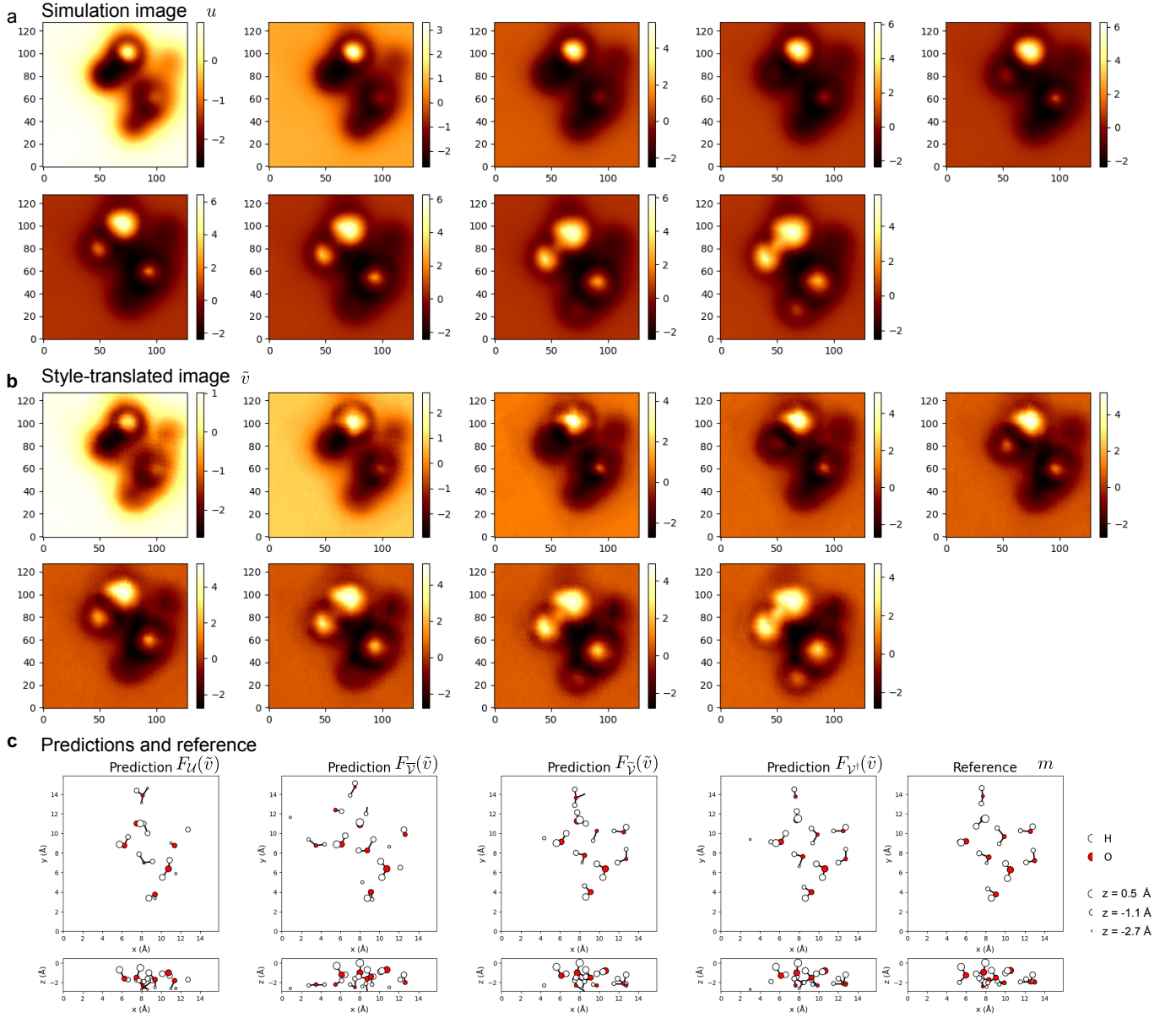


Figure S14: **Atomic configuration prediction 1 from style-translated AFM images using different structure discovery models.** **a** Simulated AFM image u generated by the PPM. **b** Experimental-style AFM image $\tilde{v}$ obtained by translating the simulated image u with the forward generator G_U . **c** Atomic structures predicted from $\tilde{v}$ using different structure discovery models trained on four types of images: pure simulated AFM images $\mathcal{U}$, handcrafted perturbed images $\bar{\mathcal{V}}$, style-translated images $\tilde{\mathcal{V}}$, and hybrid images $\mathcal{V}^\dagger$, where both style translation and handcrafted perturbation are applied. The corresponding prediction models are denoted by F_U , $F_{\bar{\mathcal{V}}}$, $F_{\tilde{\mathcal{V}}}$, and $F_{\mathcal{V}^\dagger}$, respectively. The predicted structures are compared with the reference structure m .

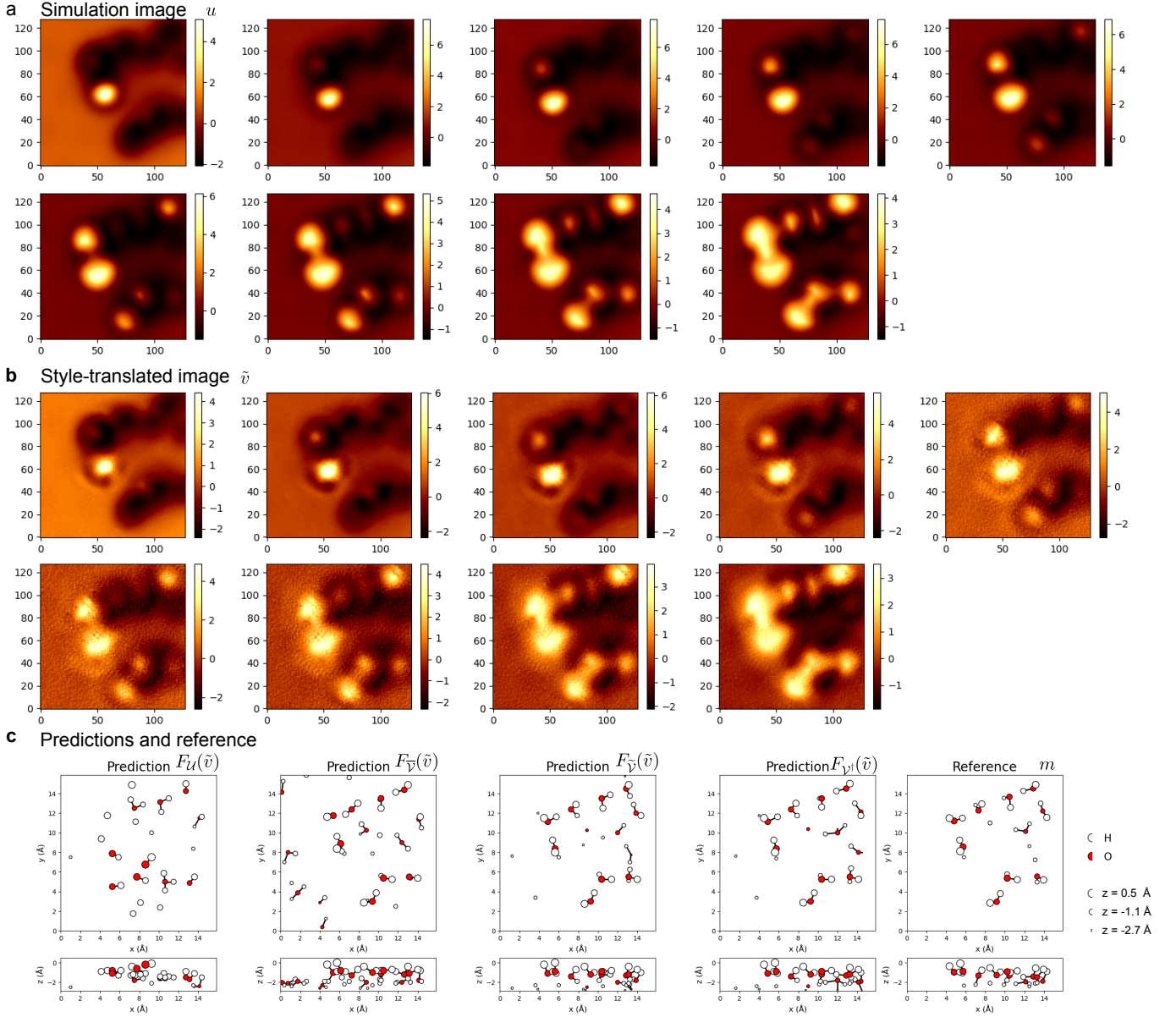


Figure S15: **Atomic configuration prediction 2 from style-translated AFM images using different structure discovery models.** **a** Simulated AFM image u generated by the PPM. **b** Experimental-style AFM image $\tilde{v}$ obtained by translating the simulated image u with the forward generator G_U . **c** Atomic structures predicted from $\tilde{v}$ using different structure discovery models trained on four types of images: pure simulated AFM images $\mathcal{U}$, handcrafted perturbed images $\bar{\mathcal{V}}$, style-translated images $\tilde{\mathcal{V}}$, and hybrid images $\tilde{\mathcal{V}}^\dagger$, where both style translation and handcrafted perturbation are applied. The corresponding prediction models are denoted by F_U , $F_{\bar{V}}$, $F_{\tilde{V}}$, and $F_{\tilde{V}^\dagger}$, respectively. The predicted structures are compared with the reference structure m .

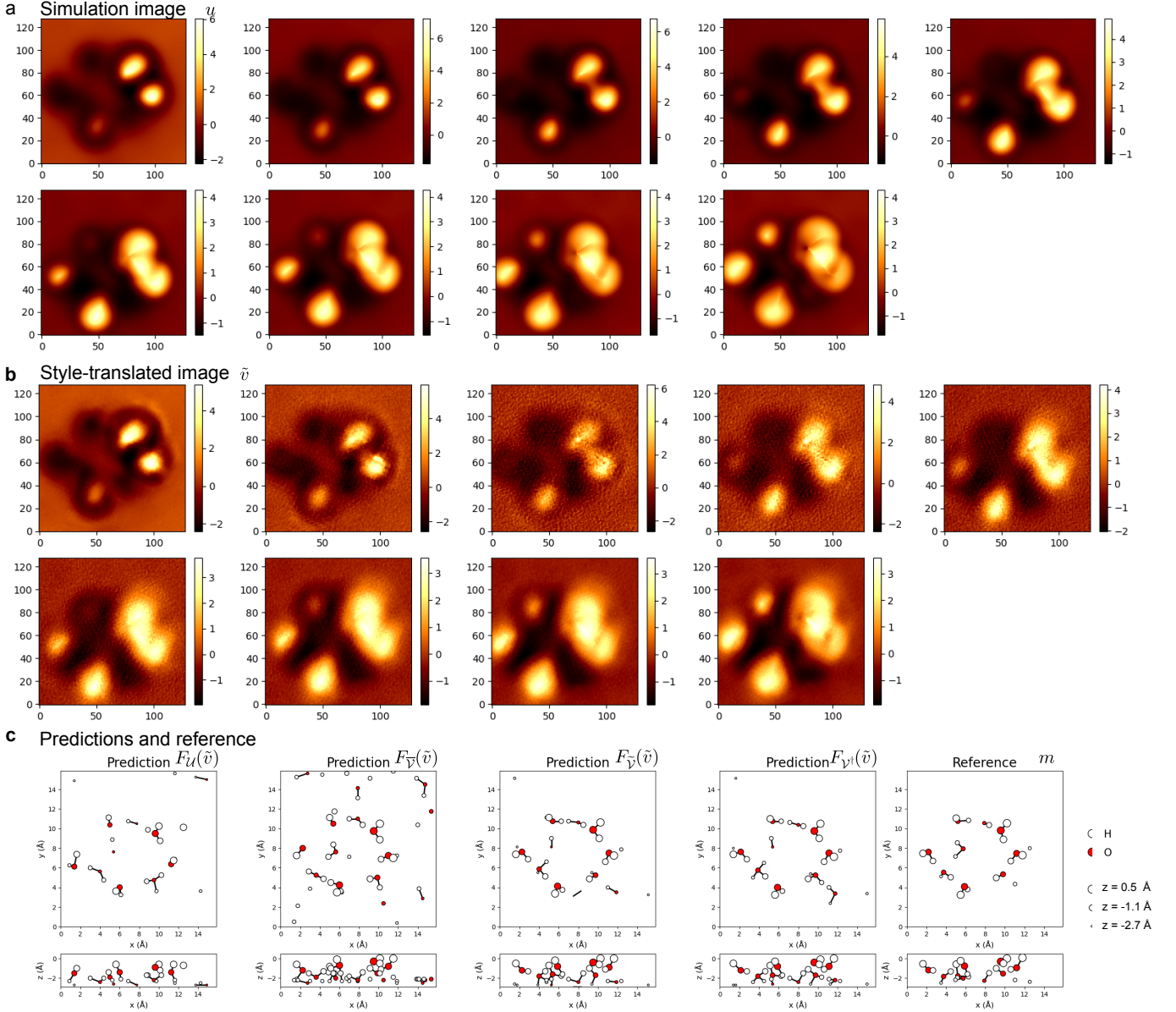


Figure S16: Atomic configuration prediction 3 from style-translated AFM images using different structure discovery models. **a** Simulated AFM image u generated by the PPM. **b** Experimental-style AFM image $\tilde{v}$ obtained by translating the simulated image u with the forward generator G_U . **c** Atomic structures predicted from $\tilde{v}$ using different structure discovery models trained on four types of images: pure simulated AFM images $\mathcal{U}$, handcrafted perturbed images $\bar{\mathcal{V}}$, style-translated images $\tilde{\mathcal{V}}$, and hybrid images $\tilde{\mathcal{V}}^\dagger$, where both style translation and handcrafted perturbation are applied. The corresponding prediction models are denoted by F_U , $F_{\bar{V}}$, $F_{\tilde{V}}$, and $F_{\tilde{V}^\dagger}$, respectively. The predicted structures are compared with the reference structure m .

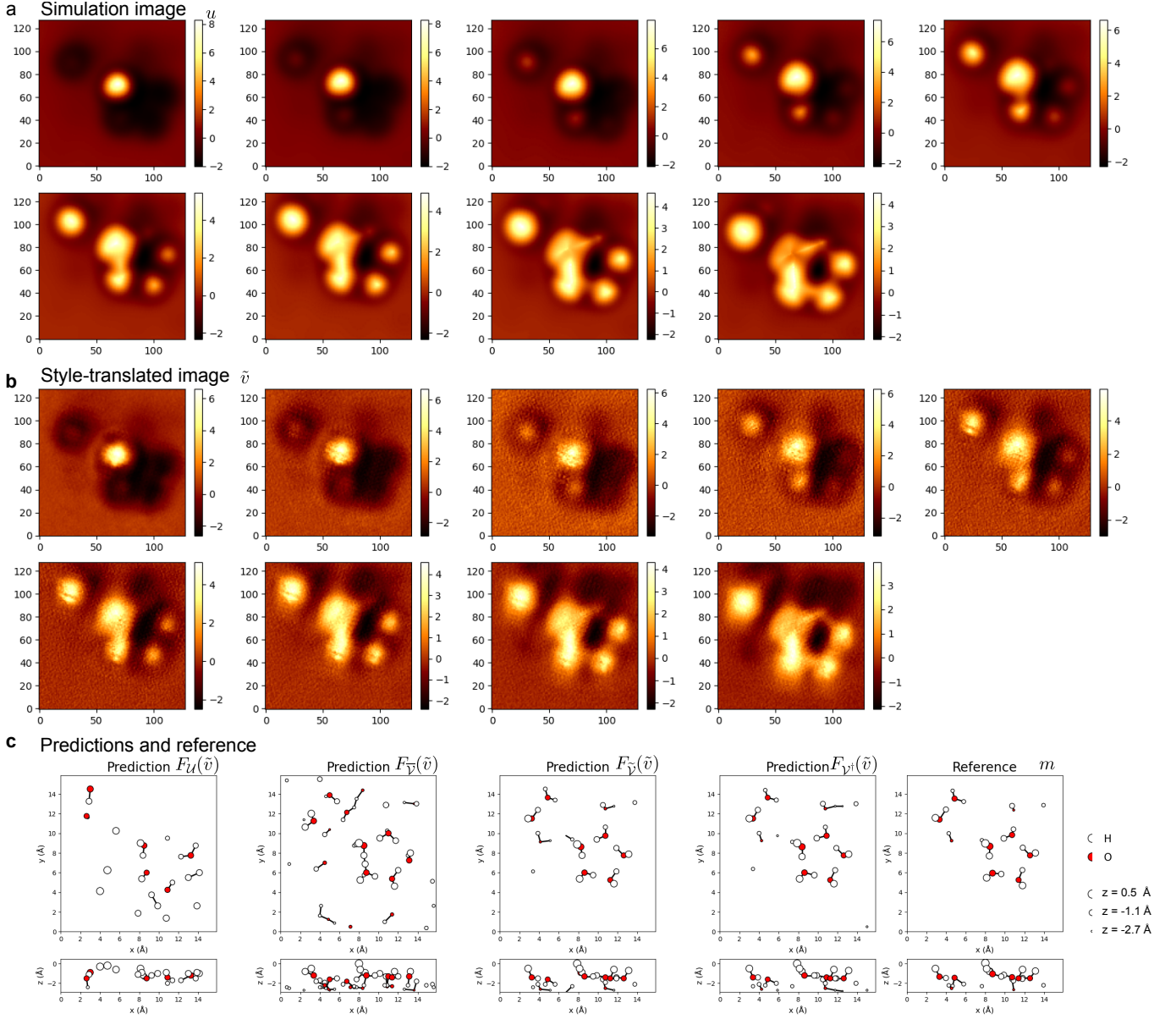


Figure S17: **Atomic configuration prediction 4 from style-translated AFM images using different structure discovery models.** **a** Simulated AFM image u generated by the PPM. **b** Experimental-style AFM image $\tilde{v}$ obtained by translating the simulated image u with the forward generator G_U . **c** Atomic structures predicted from $\tilde{v}$ using different structure discovery models trained on four types of images: pure simulated AFM images $\mathcal{U}$, handcrafted perturbed images $\bar{\mathcal{V}}$, style-translated images $\tilde{\mathcal{V}}$, and hybrid images $\mathcal{V}^\dagger$, where both style translation and handcrafted perturbation are applied. The corresponding prediction models are denoted by F_U , $F_{\bar{\mathcal{V}}}$, $F_{\tilde{\mathcal{V}}}$, and $F_{\mathcal{V}^\dagger}$, respectively. The predicted structures are compared with the reference structure m .

Supplementary result 6: Evaluations of models on experimental-style images

To evaluate model performance, we use two geometric metrics, the Hausdorff distance and the matching distance, to quantify how closely the predicted atomic positions match the reference structure. Let $\mathcal{P} = \{p_i\}$ be the set of predicted atomic positions and $\mathcal{R} = \{r_j\}$ the set of reference atomic positions.

Hausdorff distance. The symmetric Hausdorff distance is defined as

$$d_H(\mathcal{P}, \mathcal{R}) = \max \left\{ \sup_{p \in \mathcal{P}} \inf_{r \in \mathcal{R}} \|p - r\|, \sup_{r \in \mathcal{R}} \inf_{p \in \mathcal{P}} \|r - p\| \right\}. \quad (\text{S1})$$

This metric captures the worst-case positional disagreement between the two structures.

Matching distance. For each reference atom $r_j \in \mathcal{R}$, we identify the nearest predicted atom within a distance threshold $\tau = 0.35 \text{ \AA}$. Let $\mathcal{M} \subseteq \mathcal{R}$ denote the set of reference atoms for which at least one predicted atom falls within τ , and let $\pi(j)$ denote the index of the closest such predicted atom for $r_j \in \mathcal{M}$. The matching distance is then

$$d_M(\mathcal{P}, \mathcal{R}) = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \|r_j - p_{\pi(j)}\|. \quad (\text{S2})$$

Unmatched reference atoms (those with no predicted neighbor within τ) are excluded from the average. This metric reflects the mean local alignment quality of matched atoms. For both d_H and d_M , smaller values indicate better agreement with the reference structure.

Empirical cumulative distribution function. To visualize the full error distribution across all test samples, we use the empirical cumulative distribution function (ECDF). For a metric X and threshold x ,

$$G(x) = \frac{1}{N} \sum_{k=1}^N \mathbf{1}(X_k \leq x), \quad (\text{S3})$$

where N is the number of samples and $\mathbf{1}(\cdot)$ is the indicator function. $G(d_H)$ and $G(d_M)$ are shown in panels (c) and (d), respectively. Curves shifted to the left indicate that a larger fraction of samples achieve low error, corresponding to better prediction performance.

Discussion of Fig. S23. Overall, Figure S18 shows that the style-translated models, $F_{\tilde{\mathcal{V}}}$ and $F_{\mathcal{V}^\dagger}$, achieve the best geometric accuracy, as indicated by lower d_H and d_M distributions. In contrast, F_U and $F_{\tilde{\mathcal{V}}}$ are less competitive. Specifically, $F_{\tilde{\mathcal{V}}}$ yields the lowest Hausdorff-distance errors, whereas $F_{\mathcal{V}^\dagger}$ gives the lowest matching-distance errors.

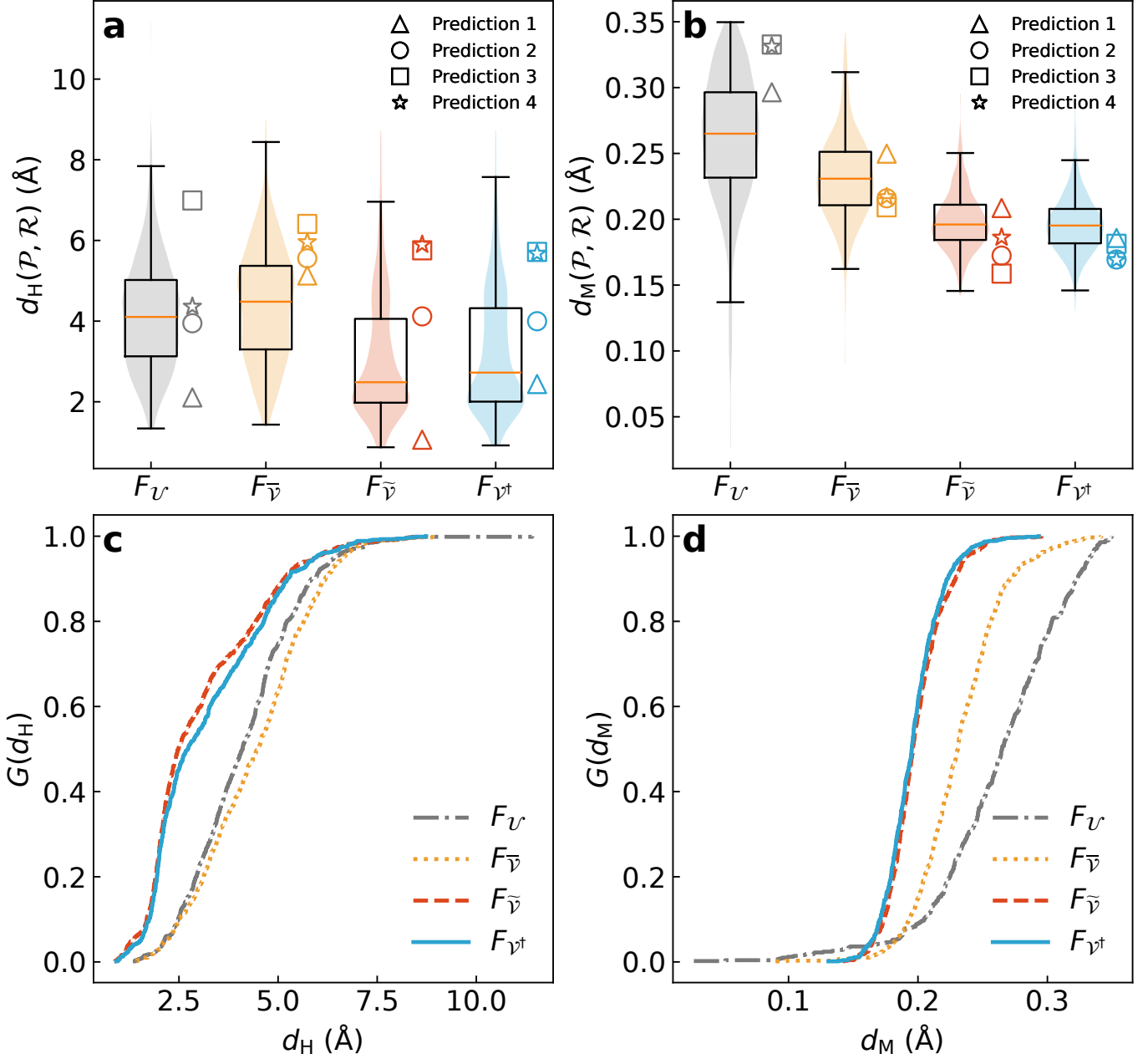


Figure S18: **Quantitative evaluation of atomic structure prediction on experimental style data using four structure discovery models.** **a,b** Distribution comparison of per-sample geometry errors across shared samples ($n=720$): Hausdorff distance (d_H) (**a**) and matching distance (d_M) (**b**), shown as violin-box plots for (F_U), ($F_{\tilde{\mathcal{V}}}$), ($F_{\tilde{\mathcal{V}}}$), and ($F_{\mathcal{V}^\dagger}$). **c,d** Empirical cumulative distribution functions of (d_H) (**c**) and (d_M) (**d**) for the same four models; left-shifted curves indicate lower error. Unfilled markers overlaid on the violin plots denote four selected examples (Prediction 1–4, shown in Figure S14–S17, respectively.), and marker edge colors follow the corresponding model color. All distances are reported in (Å), with lower values indicating better agreement with the reference structure.

Supplementary result 7: Feature similarity and slice-to-slice comparisons

Let u_i denote the original AFM slice at index i , and $\tilde{v}_i$ the corresponding style-translated slice. All slices are converted to grayscale and normalized to $[0, 1]$ before metric computation.

Slice-wise structural similarity. For each paired slice $(u_i, \tilde{v}_i)$, we compute structural similarity index measure (SSIM) using Gaussian local statistics ($\sigma = 1.5$):

$$\text{SSIM}_{\text{map}}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (\text{S4})$$

with

$$C_1 = (k_1 L)^2, \quad C_2 = (k_2 L)^2, \quad k_1 = 0.01, \quad k_2 = 0.03, \quad (\text{S5})$$

where L is the intensity range of the current image pair:

$$L = \max(\max(x, y) - \min(x, y), 10^{-8}). \quad (\text{S6})$$

The reported scalar SSIM is the mean over all pixels:

$$\text{SSIM}(u_i, \tilde{v}_i) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \text{SSIM}_{\text{map}}(u_i, \tilde{v}_i)(p). \quad (\text{S7})$$

Feature masks. For each slice I , we build a binary feature mask automatically using the same blob detector shown in panel **b** in Figs. S19–S22:

1. Convert I to 8-bit grayscale.
2. Enhance dark and bright structures by subtracting a Gaussian-smoothed background.
3. Detect dark blobs and bright blobs.
4. Keep the strongest detections, define the region inside each detected circle, and **take the union of these circular regions.**

This procedure produces one binary mask per slice, denoted $M(I)$. Accordingly, in the Dice computation we set

$$M(u_i^f) = M(u_i), \quad M(\tilde{v}_i^f) = M(\tilde{v}_i),$$

so $\text{Dice}(u_i^f, \tilde{v}_i^f)$ measures the overlap between automatically detected feature regions in the original and translated slices.

Feature-overlap Dice score. Let $M(u_i^f)$ and $M(\tilde{v}_i^f)$ be binary feature masks for original and translated slices (here obtained from dark/bright blob detections). The feature similarity is measured by

$$\text{Dice}(u_i^f, \tilde{v}_i^f) = \frac{2 |M(u_i^f) \cap M(\tilde{v}_i^f)|}{|M(u_i^f)| + |M(\tilde{v}_i^f)| + \varepsilon}, \quad \varepsilon = 10^{-12}. \quad (\text{S8})$$

Adjacent-slice correlation. To quantify slice-to-slice continuity within each domain, we compute Pearson correlation between neighboring slices after flattening:

$$C_{\text{orig}}(i, i+1) = \rho(\text{vec}(u_i), \text{vec}(u_{i+1})), \quad (\text{S9})$$

$$C_{\text{trans}}(i, i+1) = \rho(\text{vec}(\tilde{v}_i), \text{vec}(\tilde{v}_{i+1})). \quad (\text{S10})$$

This is the $C(i, i+1)$ quantity shown in the correlation panels. Here, $\rho(\cdot, \cdot)$ denotes the Pearson correlation coefficient between two vectors:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^N (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^N (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^N (y_k - \bar{y})^2}}.$$

Its value lies in $[-1, 1]$: $\rho \approx 1$ indicates strong positive similarity, $\rho \approx 0$ indicates weak linear relation, and $\rho \approx -1$ indicates strong inverse relation.

Discussion of Figs. S19–S22. Figures S19–S22 show per-sample behavior. In panel **c**, $\text{SSIM}(u_i, \tilde{v}_i)$ and feature Dice share the same trend along the height. $\text{Dice}(u_i^f, \tilde{v}_i^f)$ is more variable than SSIM, because Dice is sensitive to local feature displacement, size changes, and missing/extra detected blobs. Panel **d** compares $C_{\text{orig}}(i, i+1)$ and $C_{\text{trans}}(i, i+1)$: original data remain close to unity across adjacent pairs, while translated data show index-dependent dips, then partial recovery.

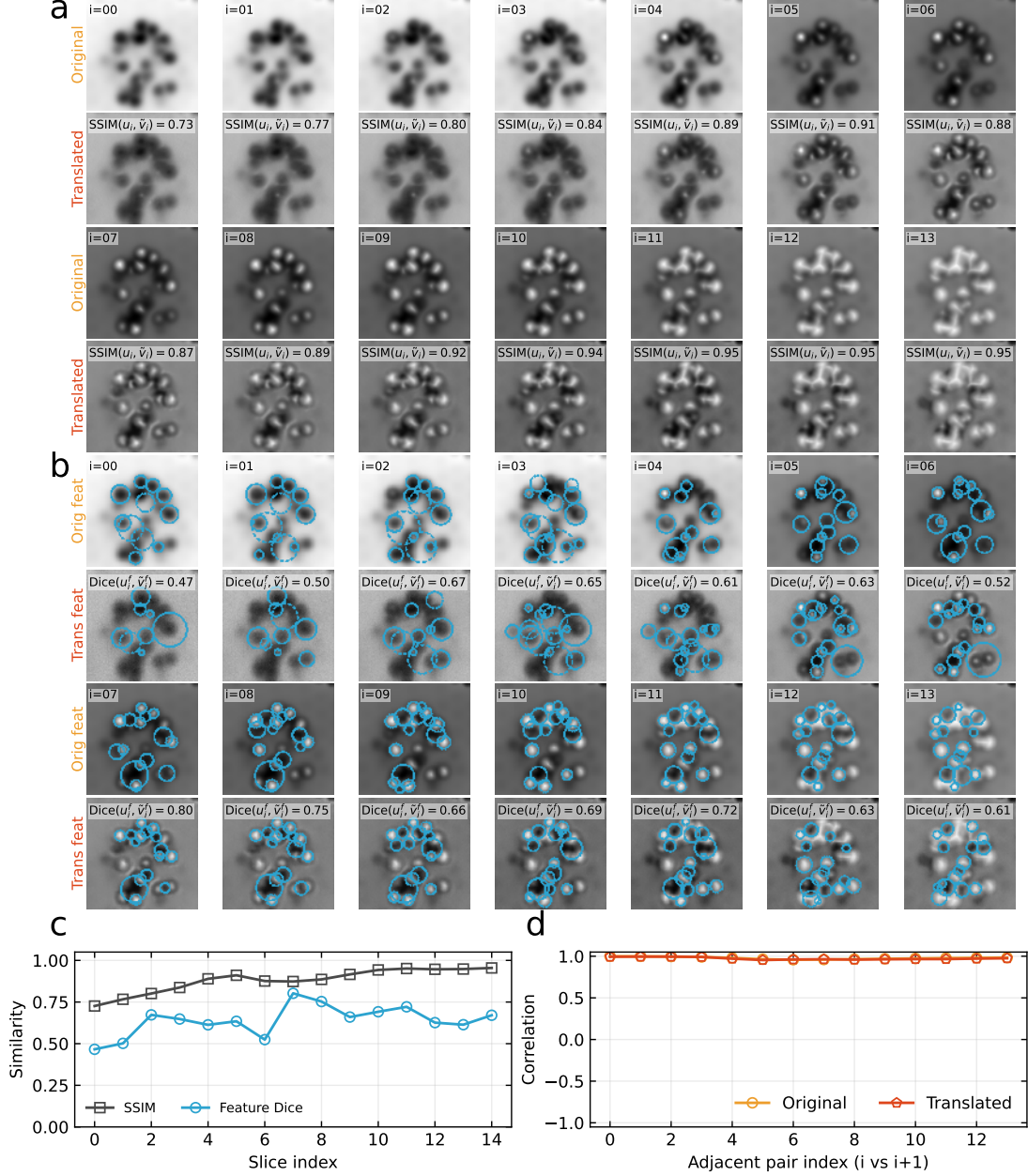


Figure S19: **Similarity and slice-to-slice comparisons for sample 1.** **a** Original and style-translated 2D image slices. The SSIM values between original and style-translated images are shown as text. **b** Dark and bright blob feature detections for original and style-translated 2D image slices. The Dice values between the regions in these circles are shown as text. **c** Slice-to-slice similarity comparisons. **d** Adjacent pair correlation comparison between original and style-translated images.

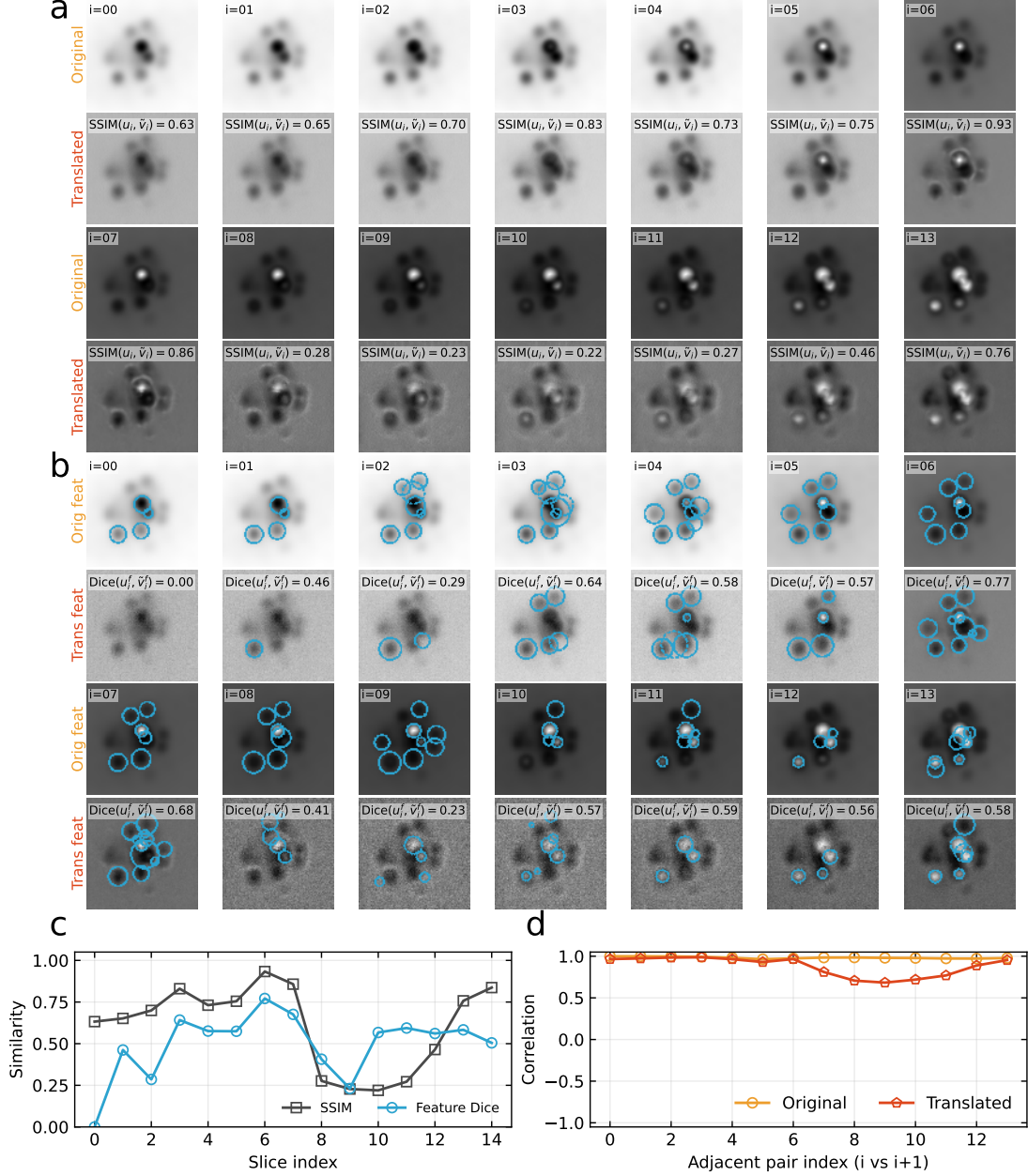


Figure S20: **Similarity and slice-to-slice comparisons for sample 2.** **a** Original and style-translated 2D image slices. The SSIM values between original and style-translated images are shown as text. **b** Dark and bright blob feature detections for original and style-translated 2D image slices. The Dice values between the regions in these circles are shown as text. **c** Slice-to-slice similarity comparisons. **d** Adjacent pair correlation comparison between original and style-translated images.

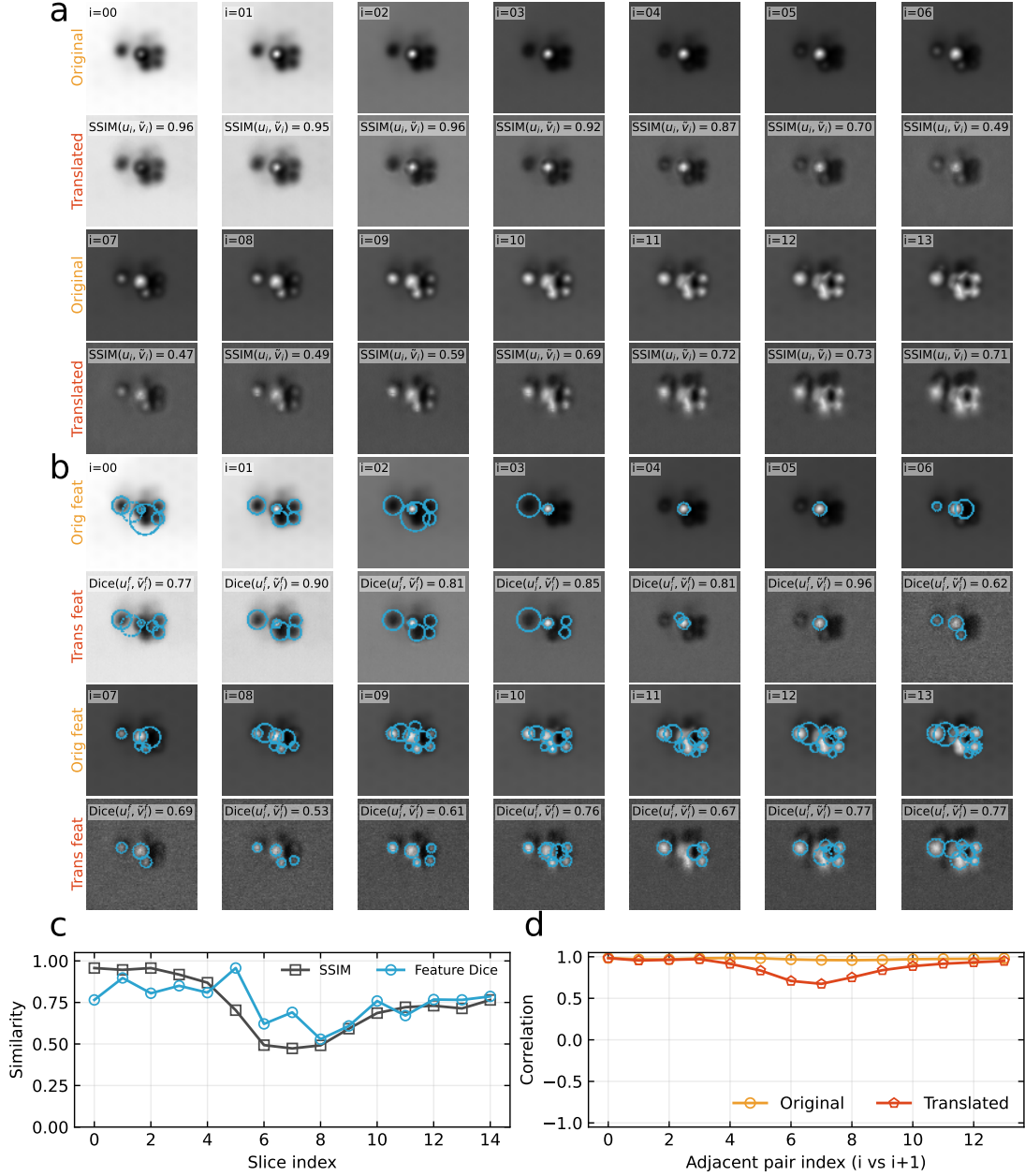


Figure S21: **Similarity and slice-to-slice comparisons for sample 2.** **a** Original and style-translated 2D image slices. The SSIM values between original and style-translated images are shown as text. **b** Dark and bright blob feature detections for original and style-translated 2D image slices. The Dice values between the regions in these circles are shown as text. **c** Slice-to-slice similarity comparisons. **d** Adjacent pair correlation comparison between original and style-translated images.

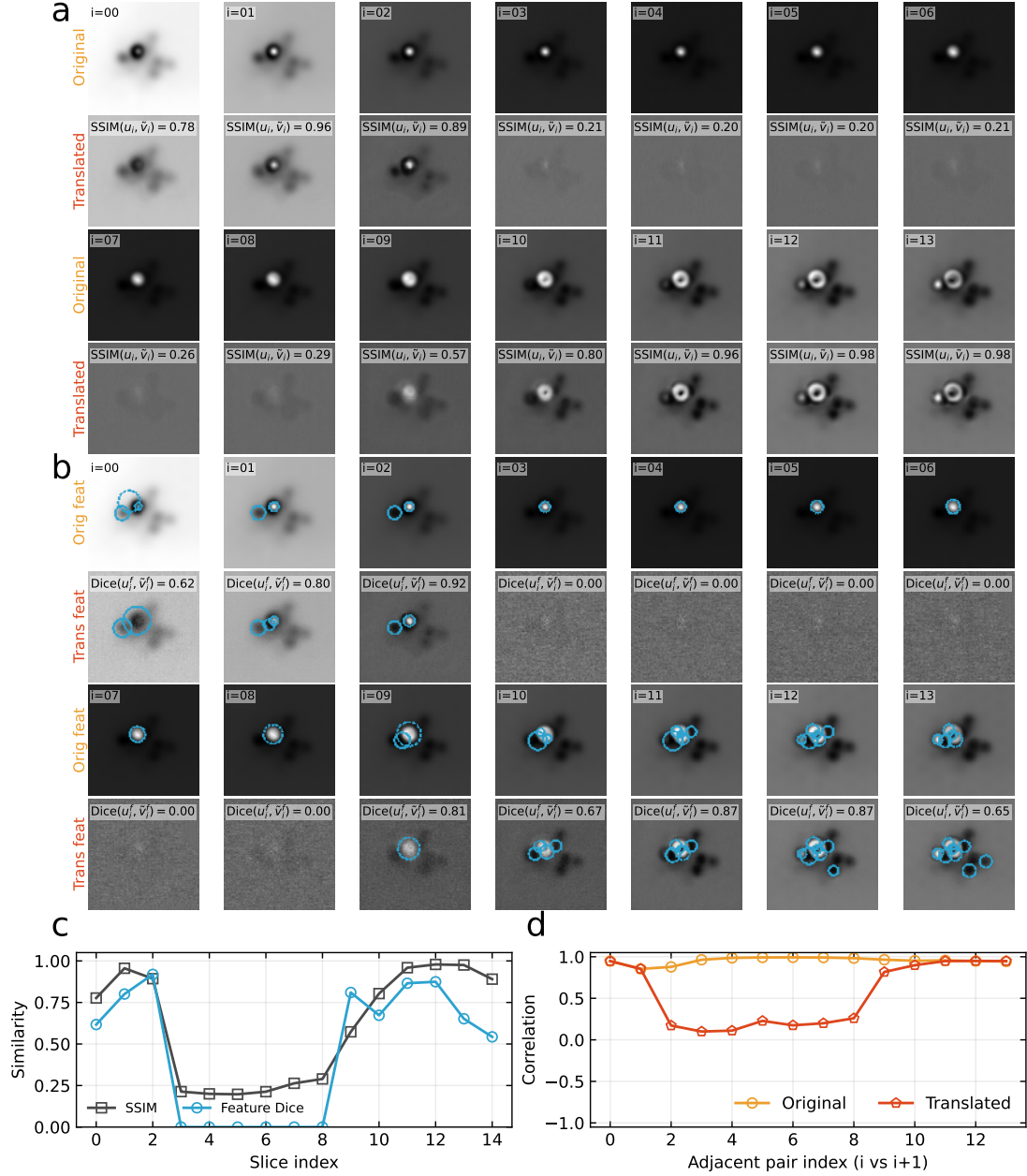


Figure S22: **Similarity and slice-to-slice comparisons for sample 2.** **a** Original and style-translated 2D image slices. The SSIM values between original and style-translated images are shown as text. **b** Dark and bright blob feature detections for original and style-translated 2D image slices. The Dice values between the regions in these circles are shown as text. **c** Slice-to-slice similarity comparisons. **d** Adjacent pair correlation comparison between original and style-translated images.

Discussion of Fig. S23. Figure S23 shows dataset-wide feature similarity and slice-to-slice consistency on the validation set. Panel **a** shows that SSIM values are mostly high, indicating that the overall image structure is largely preserved after style translation, while the broader feature-Dice distribution reflects stronger sensitivity to local feature changes. Panel **b** shows that the original simulated stacks have nearly perfect adjacent-slice correlations, whereas the translated stacks remain positively correlated but exhibit reduced consistency at intermediate slice indices.

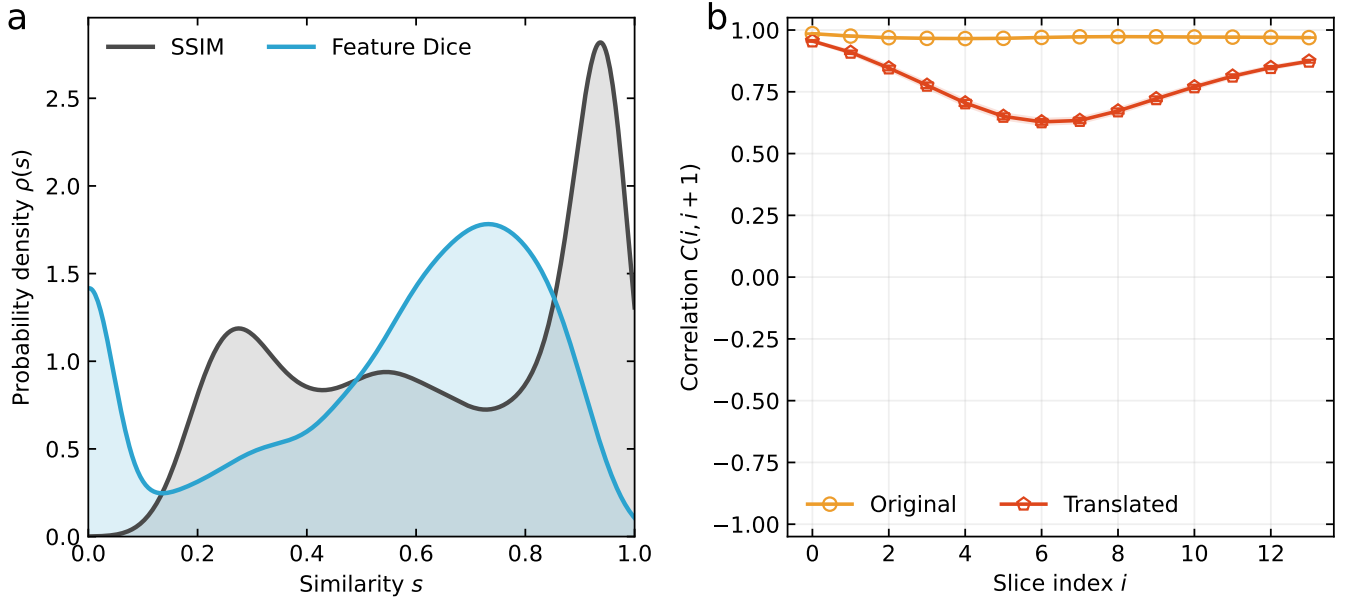


Figure S23: **Dataset wide similarity and slice-to-slice comparisons for validation set.** **a** SSIM and feature dice distributions on the validation set. **b** Slice-to-slice adjacent correlations for original and style-translated data.

Supplementary result 8: Reverse path workflow explorations

Discussion of Figs. S24–S35. We applied the experimental-to-simulation translator to the experimental inputs and show the original images, translated simulated-style images, and difference maps in Figs. S24–S35, where we demonstrated two reverse translators with different hyperparameters: $\lambda_c, \lambda_i = 20, 1$ and $\lambda_c, \lambda_i = 50, 0.5$.

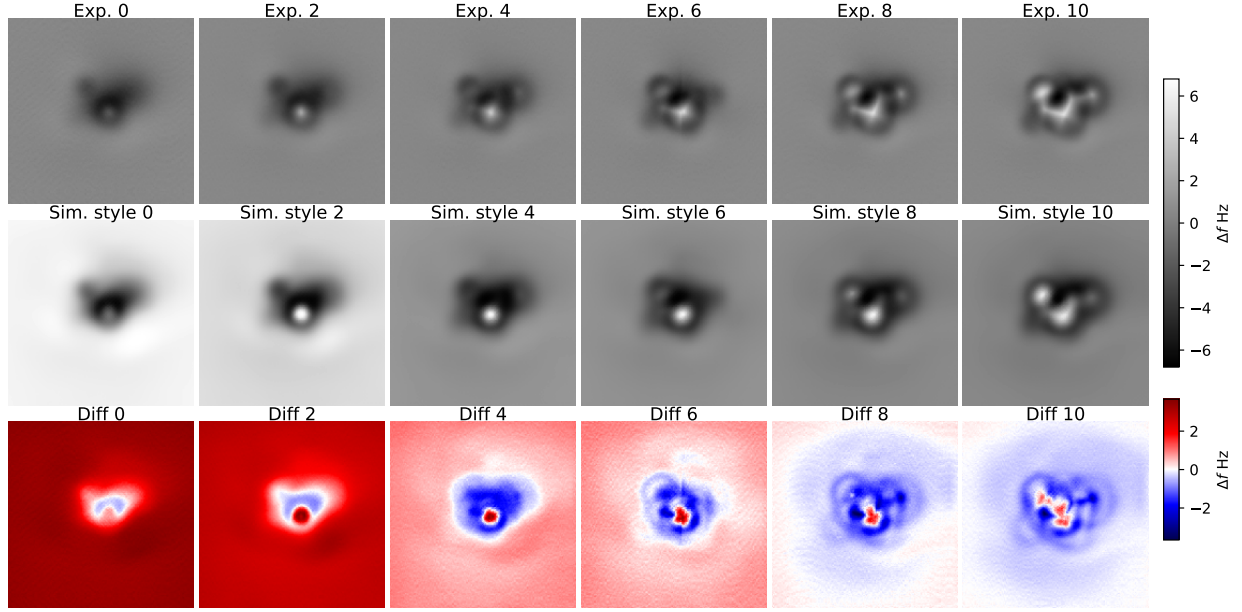


Figure S24: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

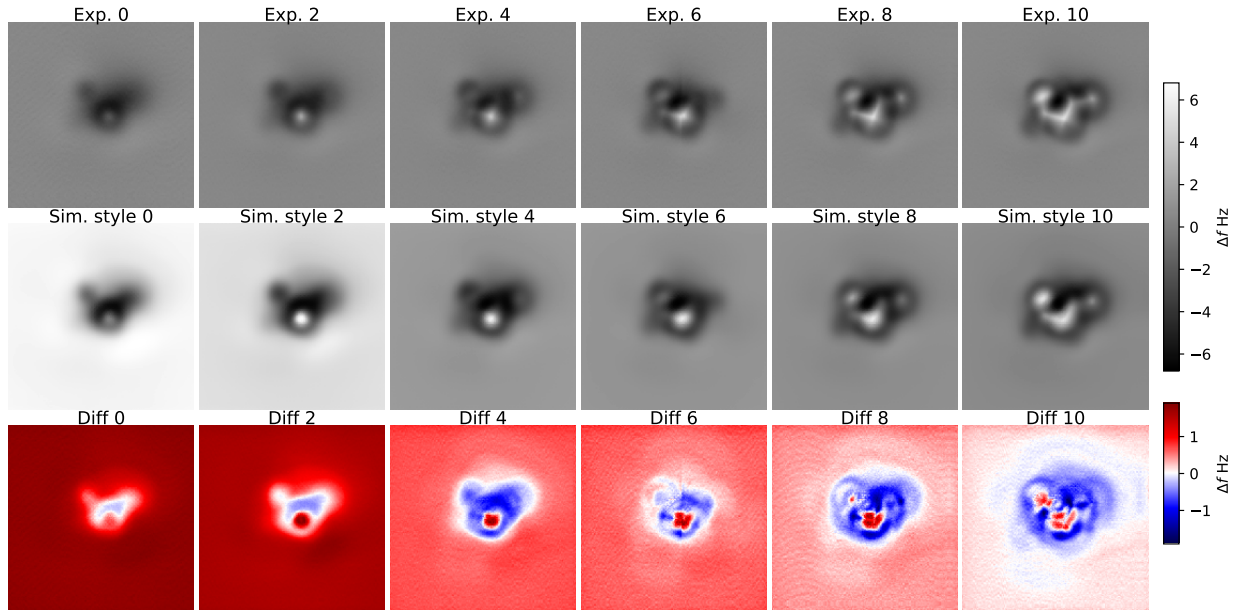


Figure S25: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

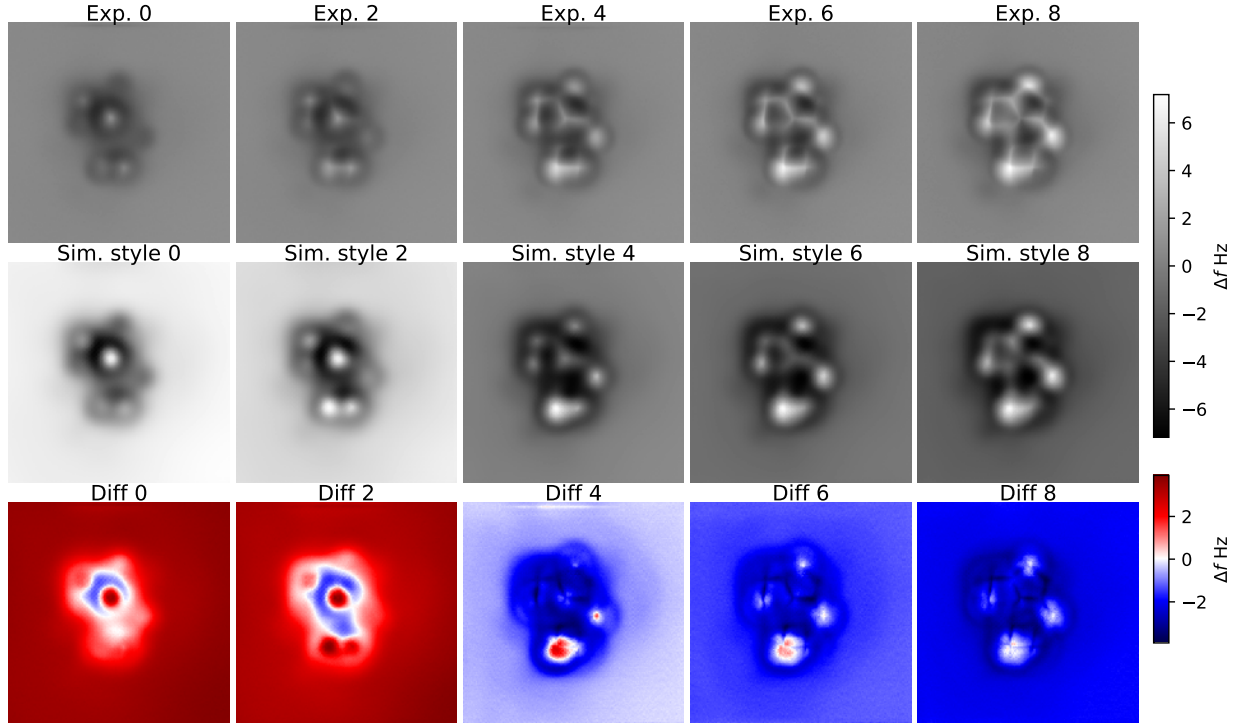


Figure S26: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

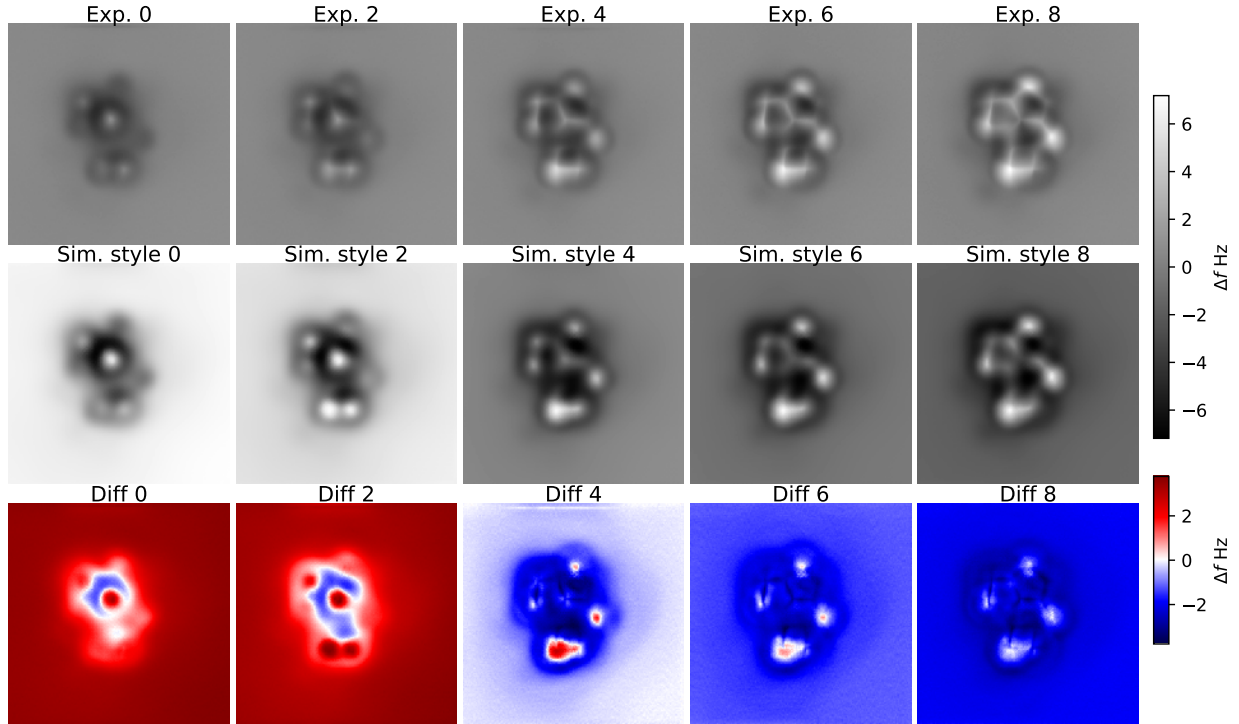


Figure S27: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

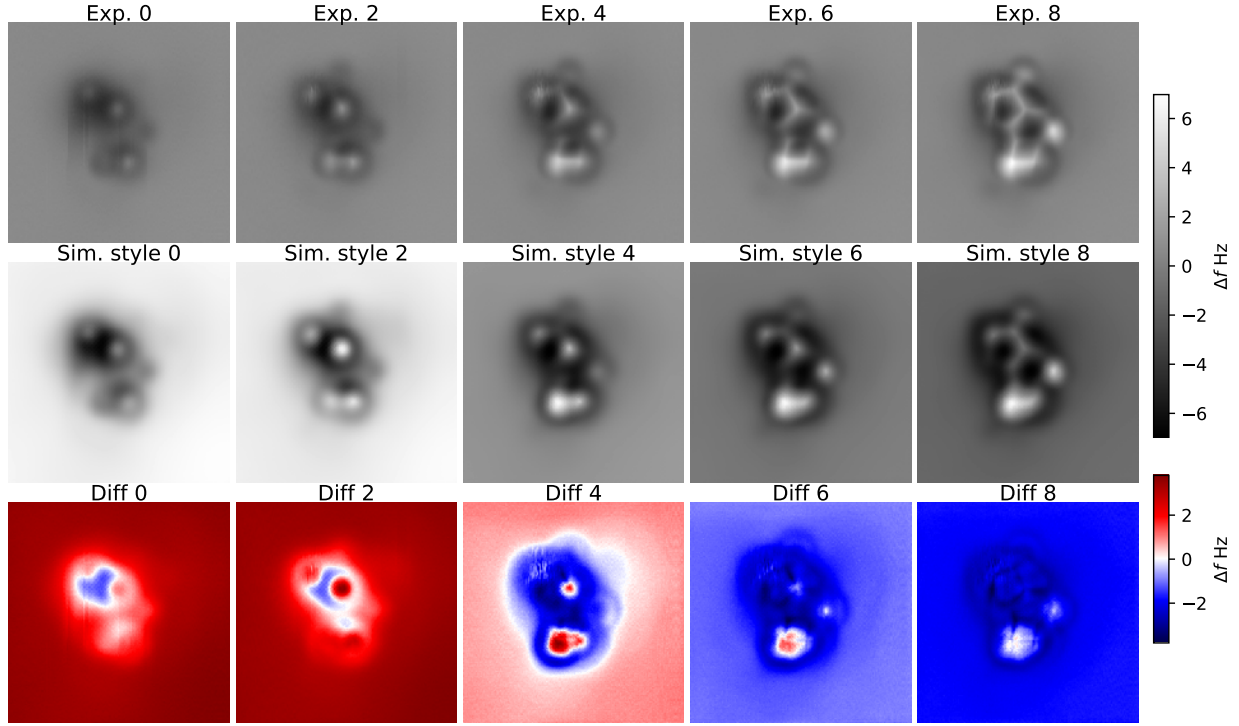


Figure S28: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

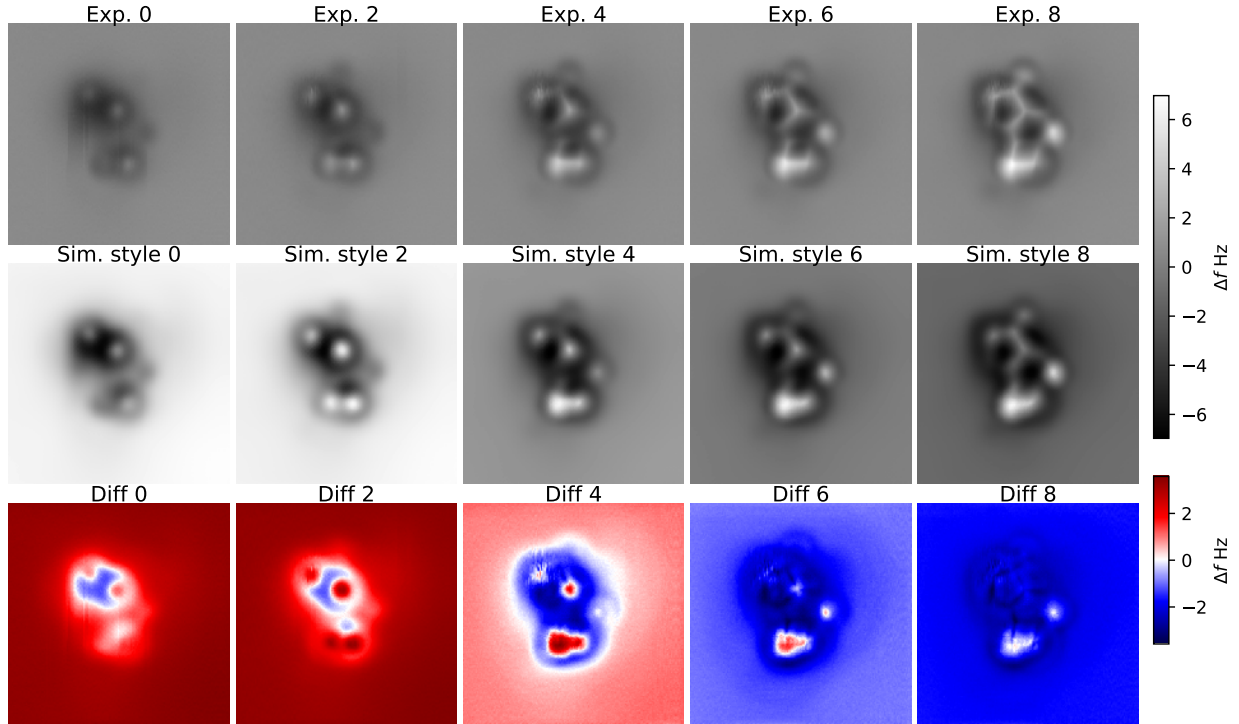


Figure S29: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

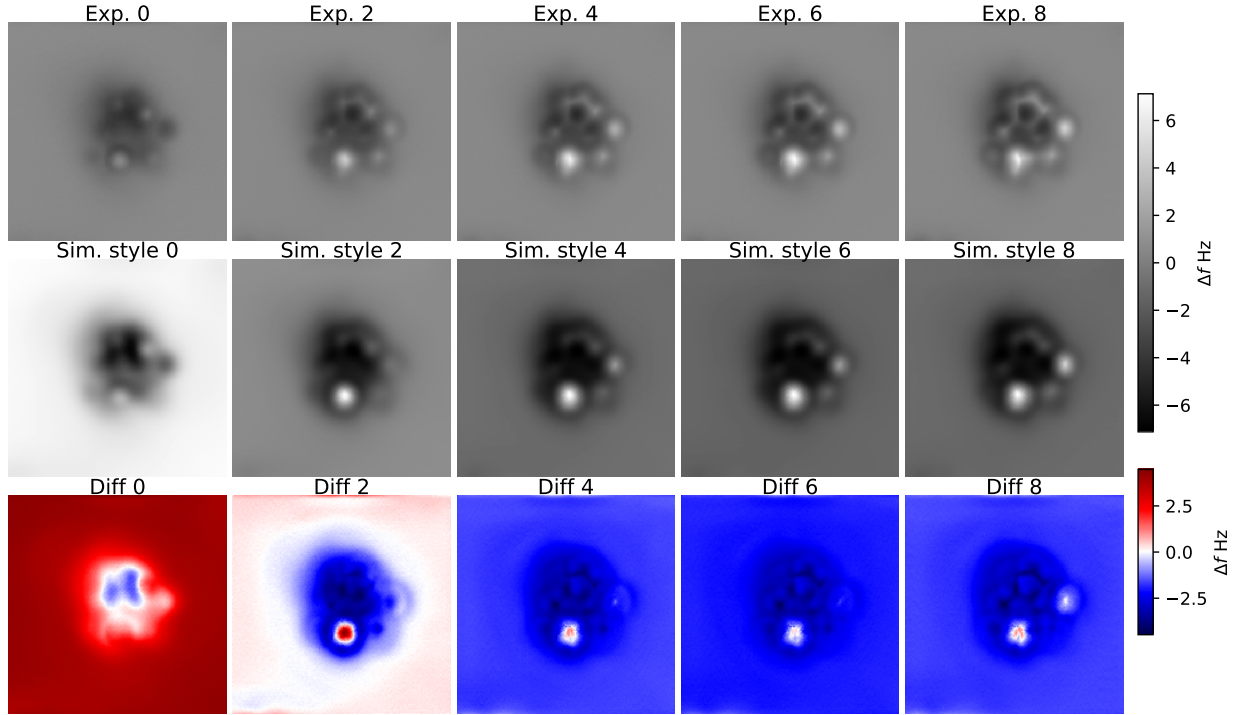


Figure S30: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

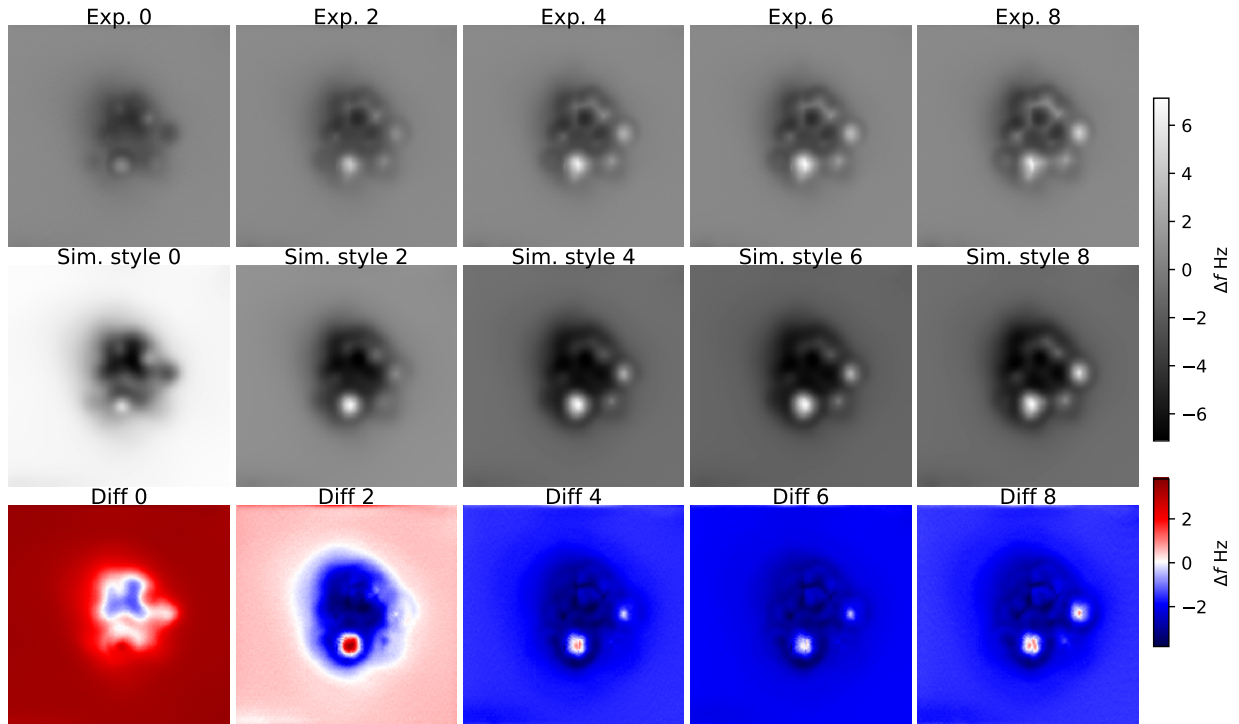


Figure S31: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

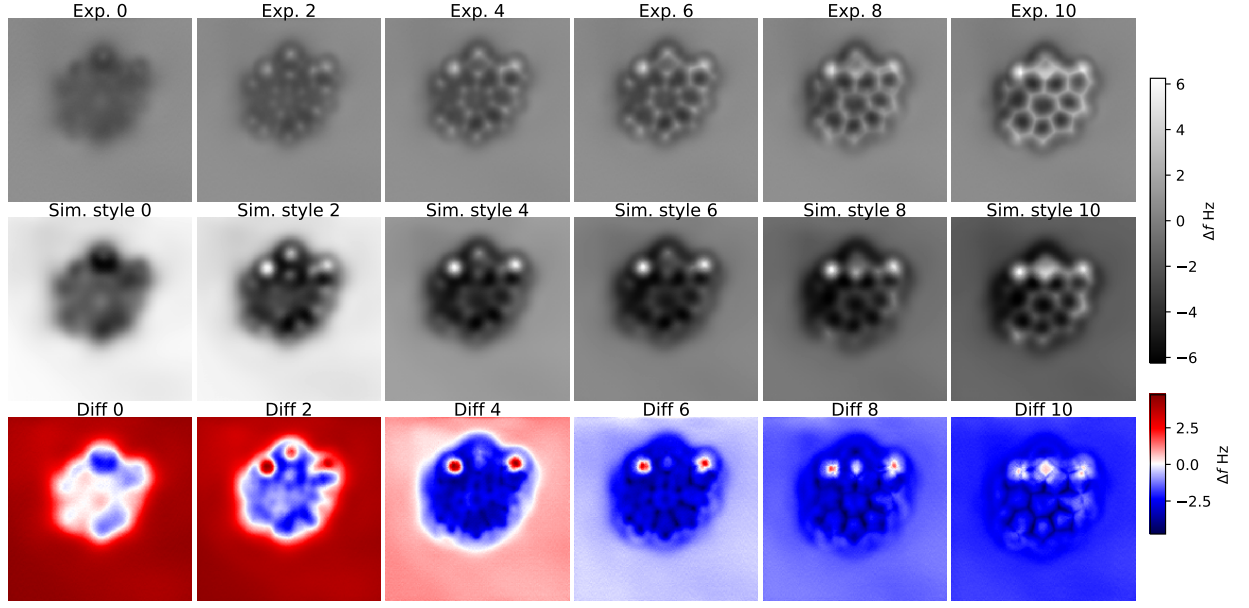


Figure S32: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

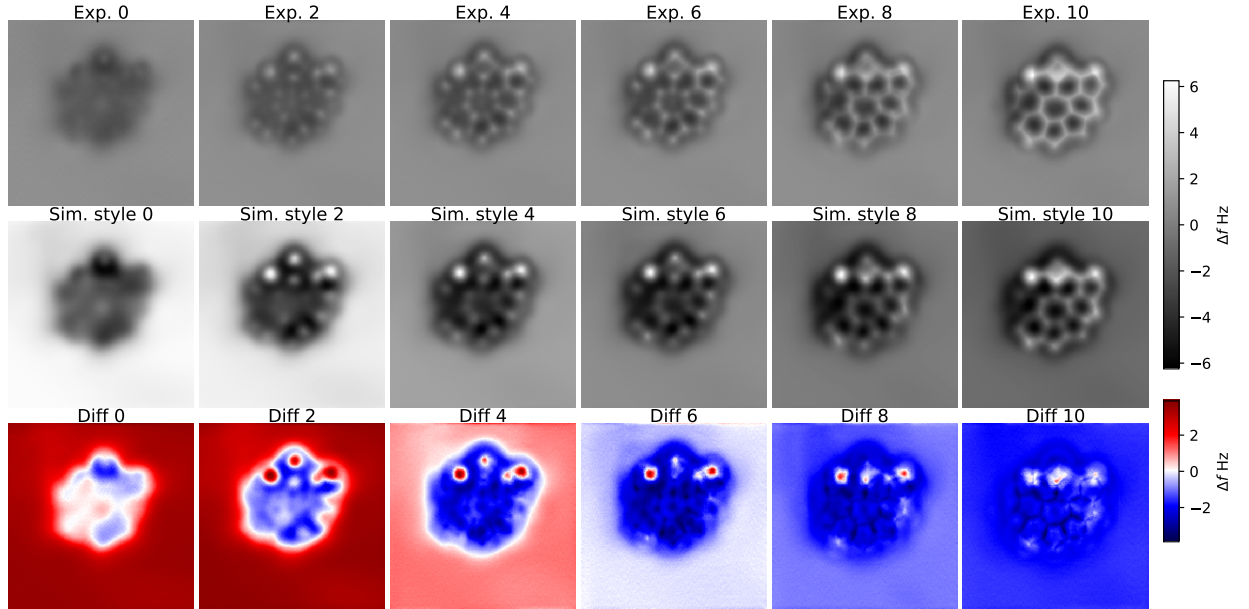


Figure S33: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

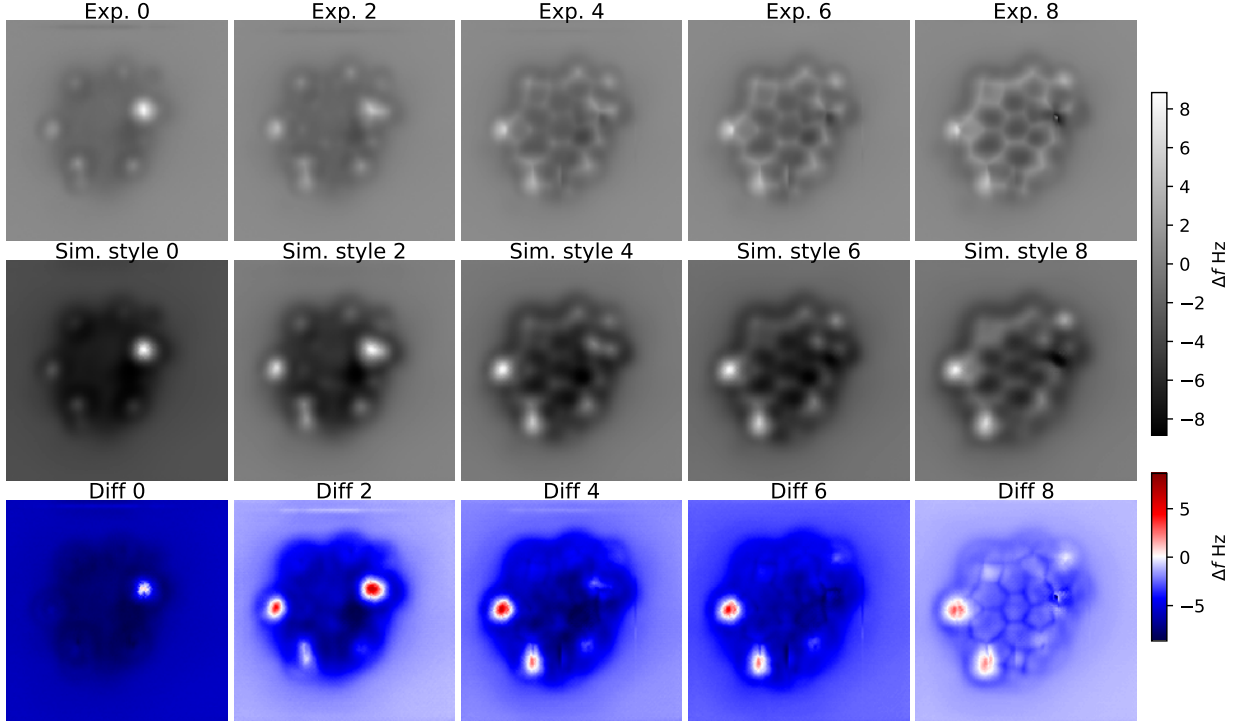


Figure S34: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (20, 1)$

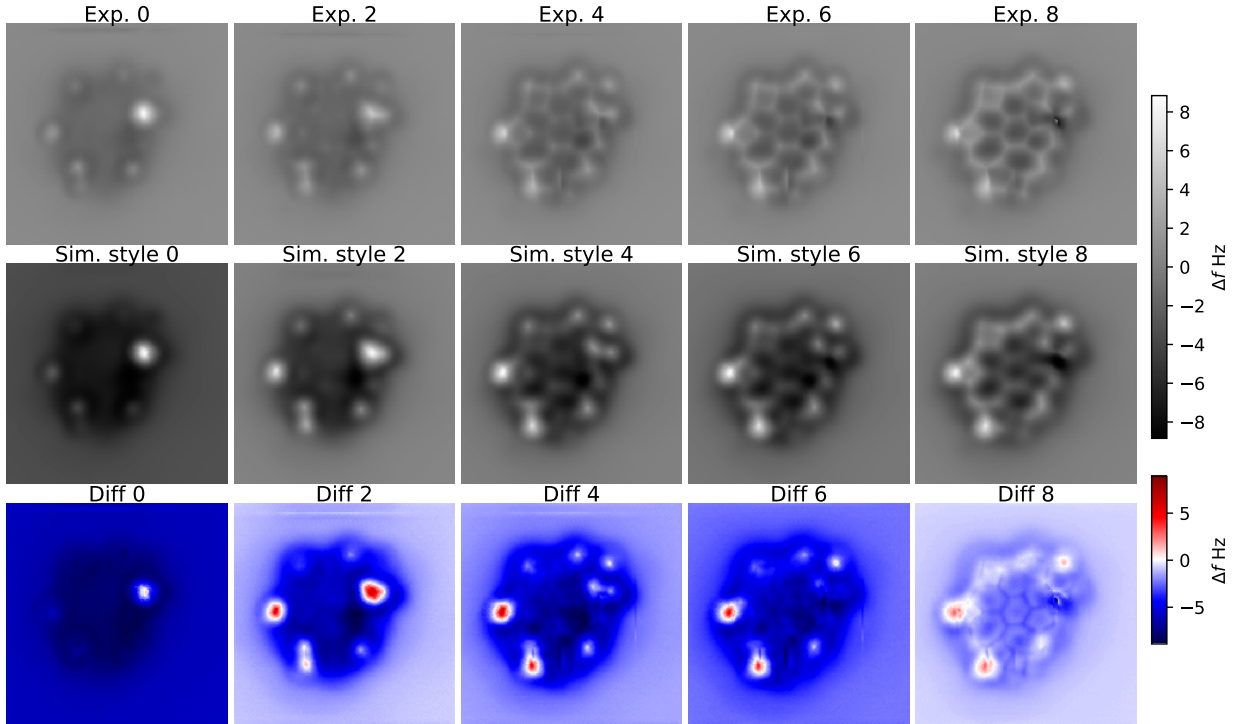


Figure S35: Original experimental images v and the corresponding simulation style images $\tilde{u}$, as well as the difference between the two kinds of images. $(\lambda_c, \lambda_i) = (50, 0.5)$

Discussion of Figs. S36–S39. The translated images $\tilde{u}$ in Figs. S24–S35 were fed into the structure discovery model $F_{\tilde{V}}$, trained on simulation data with handcrafted perturbations. Figures S36–S39 compares the resulting predictions (column 5–6) with those (column 3–4) from the forward-path workflow. The reverse-path predictions are less reliable, especially for larger clusters, where more atoms are missing. This may indicate that the experimental-to-simulation translator removes not only noise and background variations, but also image details important for structure recovery. We also observed that the predictions are sensitive to the choice of reverse translator, even when the translated images look visually similar.

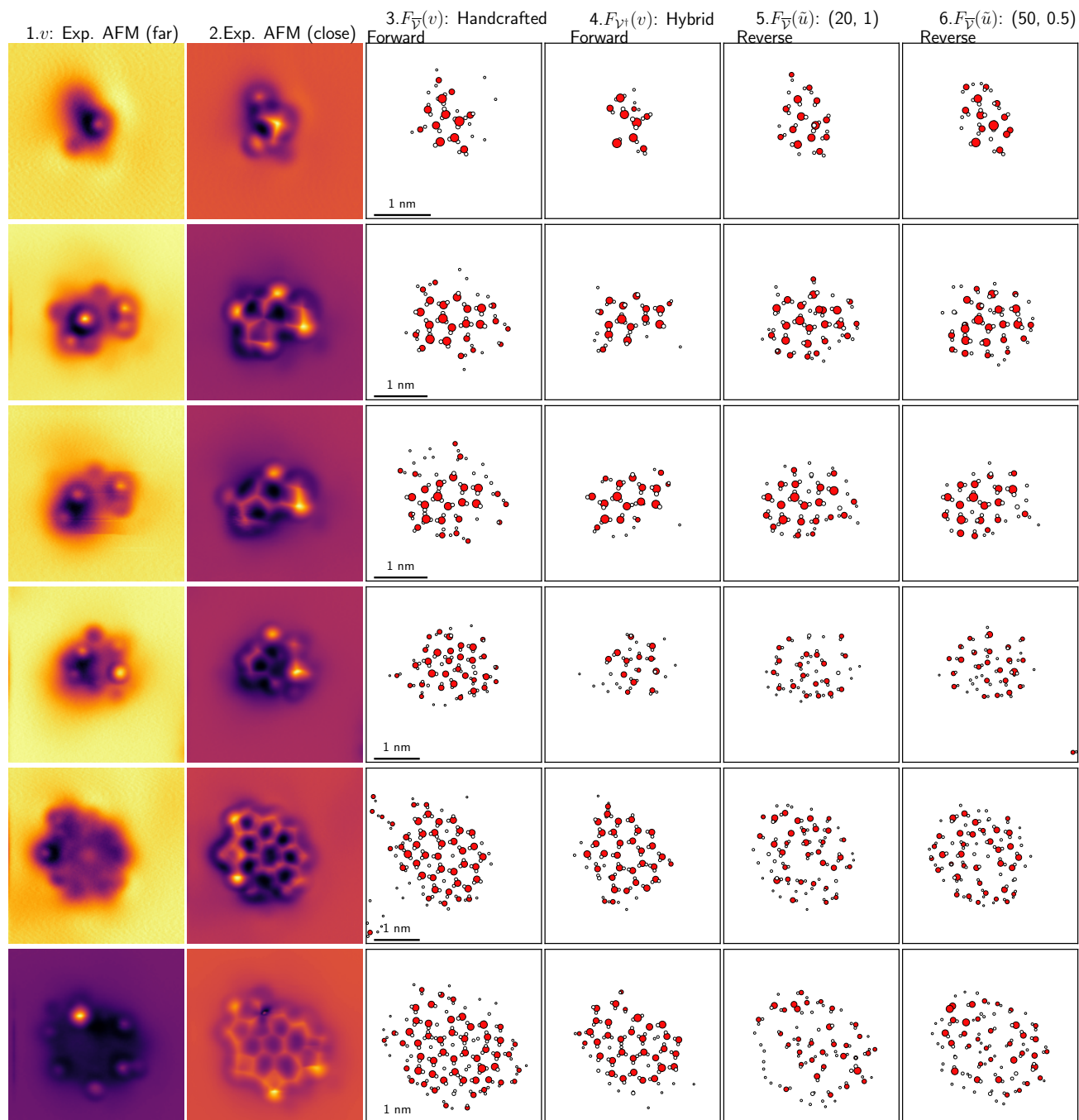


Figure S36: Comparisons between forward and backward workflow (0° rotated)..

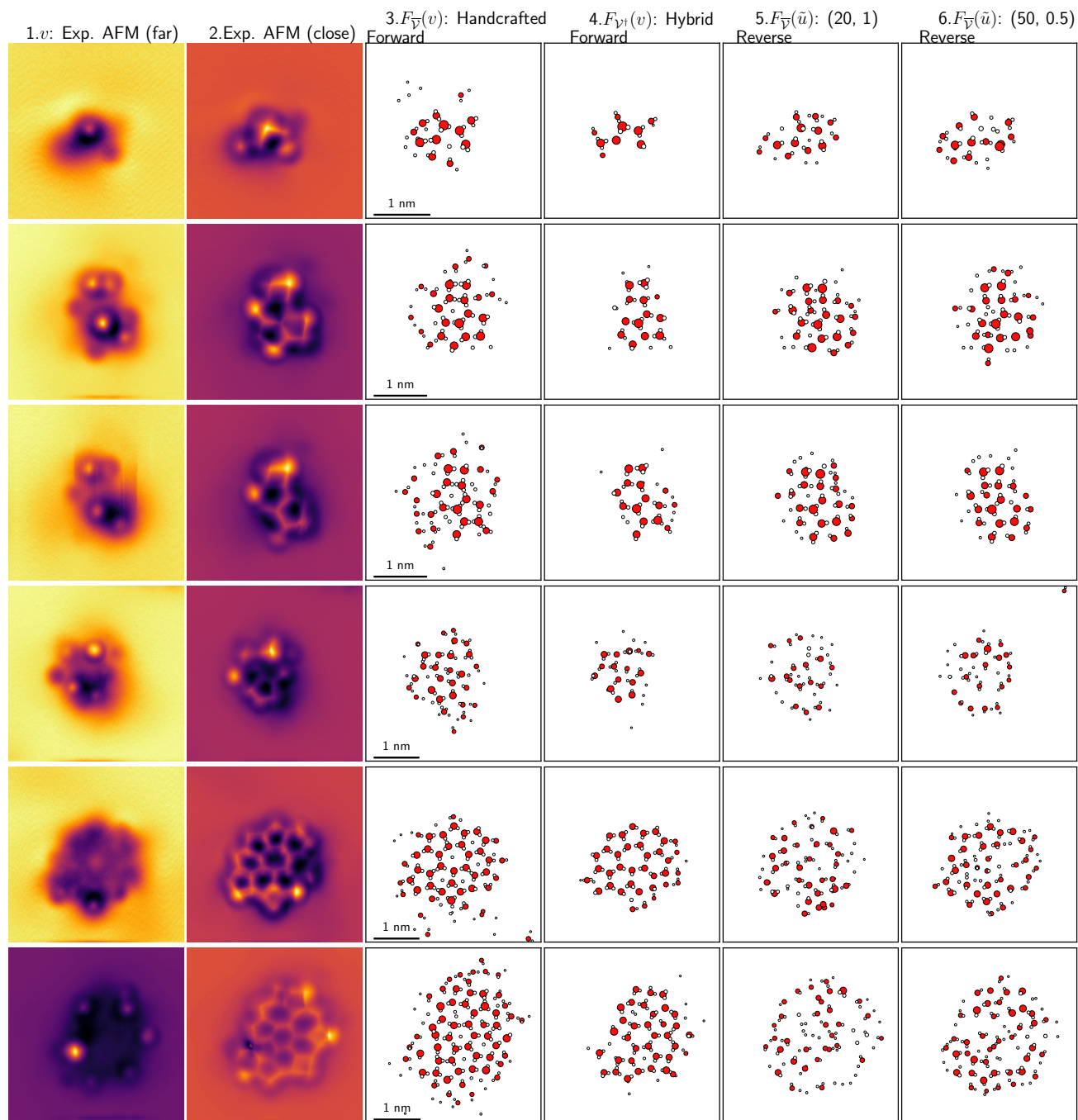


Figure S37: Comparisons between forward and backward workflow (90° rotated).

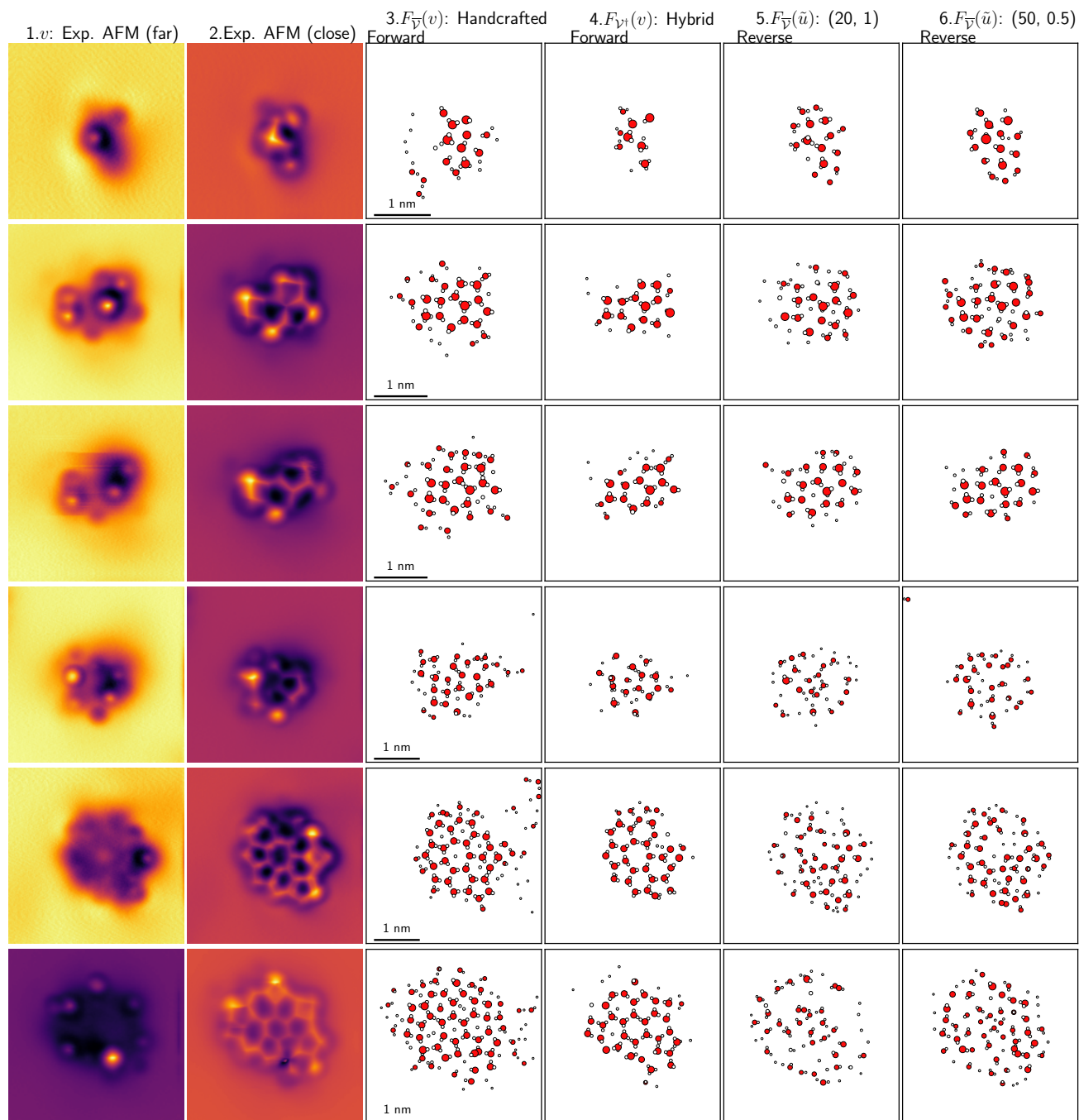


Figure S38: Comparisons between forward and backward workflow (180° rotated).

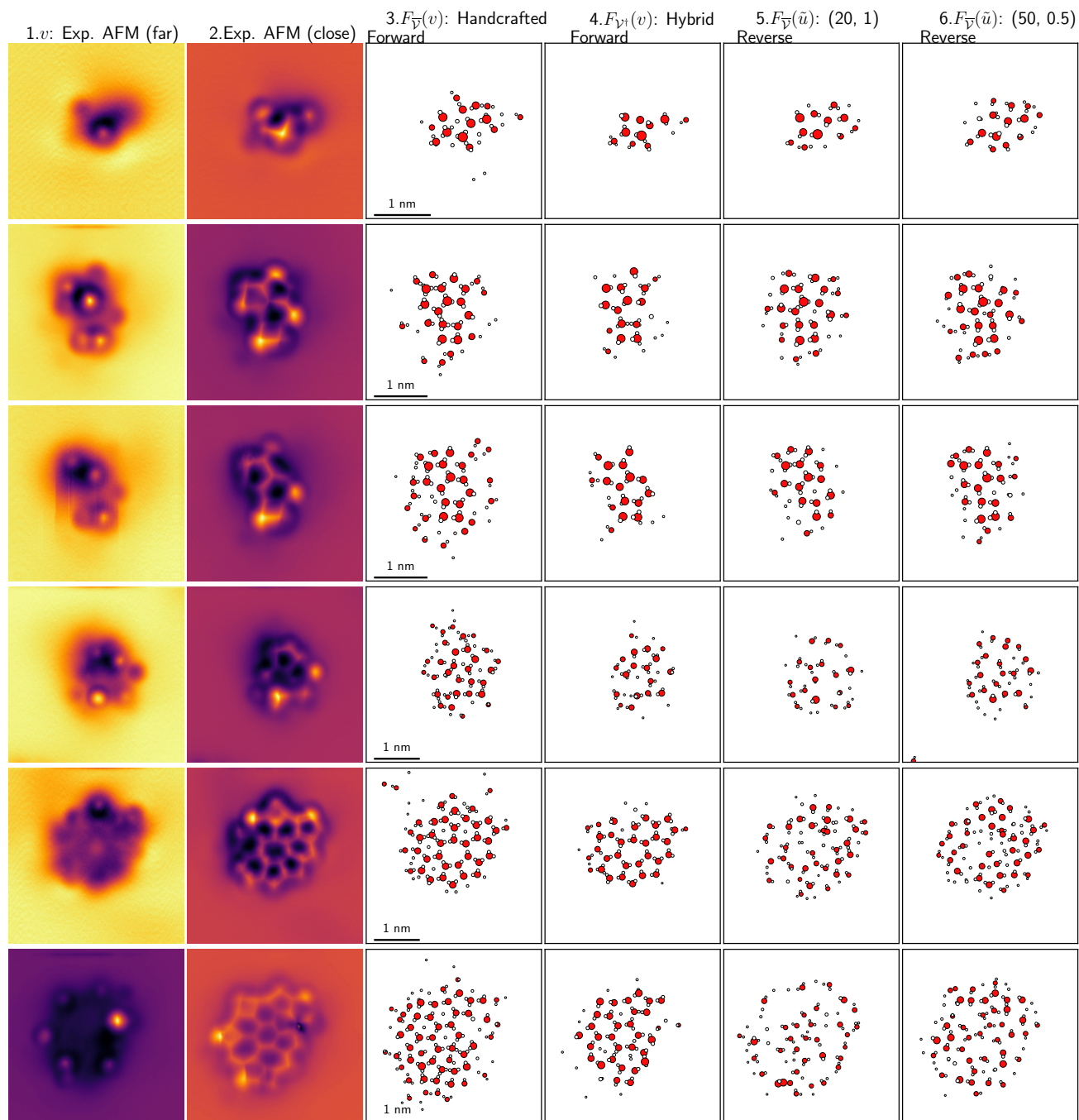


Figure S39: Comparisons between forward and backward workflow (270° rotated).