Observation-Aware Knowledge Transfer from Atomistic Models to Scientific Images

Anonymous authors
Paper under double-blind review

Abstract

Pretrained universal machine-learning force fields (uMLFFs) provide structural knowledge for scientific image models, but transfer from three-dimensional atomistic representations to two-dimensional observations is not guaranteed to improve the image task. We study training-time transfer from a frozen atomistic encoder while retaining image-only inference, using a local spatial objective that associates projected atom positions with image features. In a DenseNet AtomVision comparison, invariant MLFF supervision increases native accuracy from 79.43% to 80.09% but increases the measured rotation loss from 65.49% to 72.01%. In a matched fixed-tip STM study, an SO(2) image backbone has a lower rotation loss than a CNN backbone, while invariant MLFF supervision does not provide a consistent additional reduction. These results support conditional structural knowledge transfer and show why native task quality and rotation loss must be evaluated separately.

1 Introduction

Scientific microscopy images can support atom localization, lattice classification, molecular-graph reconstruction, and atomic-structure reconstruction from scanning-tunneling-microscopy observations (Choudhary et al., 2023; Lin et al., 2021; Oinonen et al., 2022; Priante et al., 2024; Zhu et al., 2025). These tasks are inverse problems: the input is a two-dimensional image or slice rather than a direct three-dimensional observation, and probe effects, finite support, occlusion, acquisition artifacts, and pixel sampling can make different structures produce similar observations. We therefore ask whether three-dimensional structural features learned outside the image can supplement structural information missing from an image predictor.

Pretrained universal machine-learning force fields (uMLFFs) learn chemical environments and local geometry from energies and forces across diverse atomic configurations (Kovács et al., 2025; Batatia et al., 2025). Models such as MACE therefore provide a possible source of reusable structural features, and Hessian distillation shows that information learned by an atomistic teacher can be transferred to another atomistic model (Batatia et al., 2022; Amin et al., 2025). The source and target tasks nevertheless differ: the teacher observes atom types and three-dimensional coordinates, whereas the image model observes pixel intensities. The transfer must therefore be tested with paired training data while keeping the deployed predictor dependent on the image input.

High-quality microscopy annotations and reliable image–structure correspondence are difficult to obtain at scale, while noise, tip effects, and acquisition artifacts further complicate image interpretation (Millan-Solsona et al., 2026). To determine whether structural supervision itself is useful, comparisons must hold the image backbone, data split, initialization, and training budget fixed within each task. This separates the contribution of three-dimensional structural features from changes in the image model or training recipe.

Equivariant atomistic models such as MACE organize hidden features according to irreducible representations of the three-dimensional rotation group SO(3), with non-scalar components transforming with the atomic coordinate frame (Batatia et al., 2022). Image

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

features instead transform under planar rotations in SO(2). A direct feature match can therefore compare components expressed in incompatible frames or discard the directional information that the teacher provides. The transfer requires an observation frame and a spatial correspondence between projected atoms and image features.

Steerable $E(2)$-equivariant CNNs construct image features with prescribed transformation laws under planar rotations (Weiler & Cesa, 2019). Cross-domain equivariant embeddings and induced-representation methods connect two-dimensional observations with three-dimensional rotation-aware features (Esteves et al., 2019; Howell et al., 2023). These works provide the geometric ingredients for relating image and atomistic representations.

We formulate an observation-aware route for transferring pretrained atomistic knowledge to scientific image models. We express frozen teacher features in the imaging frame and use projected atom positions together with image-feature locations to match local teacher and student features. We first examine invariant transfer in existing image predictors and then compare invariant and geometric supervision on SO(2) backbones for STM reconstruction and AtomVision classification.

Our main contributions are as follows:

- Atomistic knowledge transfer for scientific images. We investigate pretrained uMLFF representations as training-time supervision for scientific image predictors, using matched comparisons to assess their value while retaining image-only inference.
- Observation-aware geometric supervision. We formulate a spatial readout that uses projected atom positions and image-feature locations to connect local atomistic features with image features in the imaging frame.
- Controlled evaluation of transfer benefits and limitations. We establish matched comparisons on STM reconstruction and AtomVision classification that separate task performance, rotation behavior, initialization, and auxiliary-loss effects.

2 Background: Equivariant Atomistic and Image Representations

An atomistic image-learning problem associates a scientific image I with an atomic configuration X and a task target Y. The image is defined on a two-dimensional observation domain, whereas the atomistic representation is defined on a three-dimensional configuration.

For an in-plane rotation $R_\alpha \in \mathrm{SO}(2)$, define the image action as

$$(\rho_\alpha^{\mathrm{img}} I)(u) = I(R_\alpha^{-1} u), \tag{1}$$

where u is measured relative to the image centre. The desired relation for a task predictor is

$$f_\eta\big(\rho_\alpha^{\mathrm{img}} I\big) = \rho_\alpha^{\mathrm{out}} f_\eta(I). \tag{2}$$

For classification, $\rho_\alpha^{\mathrm{out}}$ is the identity; for spatial reconstruction, it acts on the predicted coordinates or fields.

Equivariant atomistic models such as MACE organize hidden features according to irreducible representations of the three-dimensional rotation group SO(3). A feature of degree ℓ at atom a and channel c is denoted by $h_{a,c}^{(\ell)}(X) \in \mathbb{R}^{2\ell+1}$. Under a three-dimensional rotation $Q \in \mathrm{SO}(3)$,

$$h_{a,c}^{(\ell)}(QX) = D^{(\ell)}(Q) h_{a,c}^{(\ell)}(X), \tag{3}$$

where $D^{(\ell)}(Q)$ is the corresponding irreducible representation matrix (Batatia et al., 2022).

An image feature field of planar frequency n is written as $z_{n,c} : \Omega \to \mathbb{R}^{d_n}$, with $d_0 = 1$ and $d_n = 2$ for $n > 0$. Under an in-plane rotation,

$$z'_{n,c}(u) = \rho_n(\alpha) z_{n,c}(R_\alpha^{-1} u), \qquad \rho_0(\alpha) = 1, \quad \rho_n(\alpha) = R_{n\alpha} \ (n > 0). \tag{4}$$

The standard steerable-kernel constraint associated with this action is recorded in Appendix A.

3 Limitations of Existing Studies and Research Gap

Scientific image predictors are task-specific. Scientific image models have demonstrated that microscopy observations can support classification, localization, and structure reconstruction. However, these predictors are generally designed for a particular imaging modality, dataset, and output representation. Existing task performance therefore does not establish a reusable mechanism for incorporating three-dimensional chemical knowledge into an image predictor.

Transferability and attribution of MLFF knowledge. Pretrained MLFFs provide a possible source of structural knowledge, but accuracy on energy or force prediction does not guarantee improvement on a scientific-image task. Transfer depends on teacher coverage, source–target correspondence, and compatibility between the atomistic representation and the image observation. The effect of structural supervision must therefore be separated from image-model initialization, training history, and auxiliary-loss design.

The SO(2)–SO(3) observation gap. Atomistic representations and image representations follow different transformation laws: the former are defined under three-dimensional rotations, while the latter are defined under planar image rotations. A direct feature match can therefore compare components in incompatible frames or discard directional information. Invariant transfer and a steerable image backbone each address only part of this problem; neither alone specifies the spatial correspondence needed for an observation-aware interaction.

Research questions and positioning. These limitations define the questions studied in this work. First, can pretrained atomistic features improve image prediction under matched within-task comparisons? Second, does an observation-aware interaction between atomistic and image representations provide an advantage over invariant transfer? We answer these questions by separating native task quality, controlled rotation performance, prediction consistency, initialization, and auxiliary-loss effects.

4 Symmetry-Aware Cross-Modal Transfer

4.1 Problem Setup and Training Path

We use a frozen atomistic encoder as a training-time teacher for an image predictor. The image model receives a scientific image I and retains its original supervised target Y. The teacher receives the paired atomic configuration X and produces structural features used by the local transfer objective. The deployed prediction path remains the image model and its original output head.

4.2 Teacher Target and Local Spatial Transfer

The primary transfer objective is local spatial supervision. For each visible atom, the teacher target is associated with the image feature sampled at the atom's projected position. The local readout maps the student feature to the corresponding teacher target while preserving the representation type required by the observation frame.

The transferred teacher descriptor concatenates scalar channels with vector norms:

$$\phi_{\mathrm{inv},a}(X) = \mathrm{concat}\left(h_a^{(0)}(X), \{\|h_{a,c}^{(1)}(X)\|_2\}_{c=1}^{128}\right). \tag{5}$$

The descriptor is invariant to the teacher coordinate frame but does not retain vector directions. The complete local objective, projection dimensions, normalization procedure, and task-specific implementation details are given in Appendix B.

4.3 Observation-Aware Spatial Interaction

To connect atomistic and image representations, we express structural quantities in the imaging frame. Let $F \in \mathrm{SO}(3)$ denote the observation frame and o its origin. A structural

position r and a teacher vector v are represented in this frame as

$$r^F = F^\top (r - o), \qquad v^F = F^\top v. \tag{6}$$

The projected position of r^F determines the image location at which the corresponding student feature is sampled. Let $\pi(r^F)$ denote this projected position. The local readout uses the relative displacement

$$d_a(u) = u - \pi(r_a^F) \tag{7}$$

between an image-feature location u and the projected position of atom a. It then maps local image features and the corresponding teacher target according to their compatible representation types.

4.4 Training Objective and Image-Only Inference

Each treatment optimizes the original task loss together with the local auxiliary objective:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{local}}. \tag{8}$$

The plain control sets the auxiliary weight to zero. The teacher parameters are frozen, and the auxiliary targets do not receive gradients. The image predictor retains its original supervised output target. At deployment, the prediction path uses the image model and its task head without requiring the paired atomic configuration.

5 Experiments

We evaluate training-time transfer from a frozen machine-learned force field on two complementary scientific-image tasks: AtomVision lattice classification and original-scale fixed-tip STM reconstruction. AtomVision measures classification under controlled in-plane rotations, whereas STM measures spatial reconstruction and transformation consistency of predicted image fields. All primary comparisons use seed 0, validation-based checkpoint selection, and image-only inference.

5.1 Experimental Setup

The main AtomVision comparison uses the DenseNet image predictor with its original rendered-image input and initialization protocol. The dataset contains 6,450 training, 1,324 validation, and 1,376 test images with composition-grouped splits. Checkpoints are selected by validation negative log likelihood with epoch 0 eligible, and the selected checkpoint is evaluated at 24 in-plane angles.

The STM bank contains 254 training, 32 validation, and 32 in-distribution test images from disjoint molecule sets. Its matched comparison uses CNN U-Net and SO(2) U-Net, each with and without frozen MACE-OFF23 supervision. All STM arms use seed 0, batch size 8, 100 epochs, FP32, 192-pixel inputs, and no online augmentation.

For classification, we define rotation loss as the mean error rate over the measured angles,

$$\mathcal{L}_{\text{rot}}^{\text{cls}} = 1 - \frac{1}{K} \sum_{k=1}^{K} \text{Acc}(f(\rho_{\theta_k} I), y). \tag{9}$$

For reconstruction, rotation loss is the mean L1 error between the prediction from a rotated image and the correspondingly rotated target. Both quantities are lower-is-better.

5.2 Main Results

The DenseNet comparison tests whether invariant MLFF supervision improves classification under the original image pipeline. Native accuracy changes from 79.43% to 80.09%, while rotation loss increases from 65.49% to 72.01%. The small native improvement is therefore accompanied by worse performance under measured rotations.

Table 1: DenseNet AtomVision comparison at validation-selected checkpoints. Native accuracy is higher-is-better; rotation loss is the mean error rate over 24 measured angles and is lower-is-better.

Supervision	Native accuracy (%)	Rotation loss (%)
midrule Image-only	79.43	65.49
Invariant MLFF bottomrule	80.09	72.01

Table 2: Original-scale fixed-tip STM reconstruction at validation-selected checkpoints. Values are L1 losses multiplied by 10^3 and are lower-is-better.

Student	Native loss	Rotation loss	Nonzero-angle consistency loss
midrule CNN U-Net	3.266	11.783	11.000
CNN U-Net + invariant MLFF	3.573	11.625	10.582
SO(2) U-Net	3.034	5.591	1.128
SO(2) U-Net + invariant MLFF bottomrule	3.129	5.745	1.172

The STM results compare the image backbone and invariant MLFF supervision under a matched reconstruction protocol. The SO(2) image-only model has a lower native error and a substantially lower rotation loss than the CNN image-only model. Under validation-based selection, invariant MLFF supervision changes the CNN rotation loss from 11.783 to 11.625 and the SO(2) rotation loss from 5.591 to 5.745.

These results separate the contribution of the image backbone from invariant structural supervision. They support conditional MLFF transfer and a lower measured rotation loss for the SO(2) backbone under the stated protocol, but do not establish universal MLFF gains, exact continuous equivariance, or a validated general SO(2)–SO(3) interaction.

6 Related Work

Scientific image analysis and structure reconstruction. Machine learning has been increasingly applied to scientific images for atom localization, material classification, and structure reconstruction. AtomVision studies machine-vision problems for atomistic images, including classification and image-based structural analysis (Choudhary et al., 2023). TEMImageNet and AtomSegNet focus on atom segmentation, localization, denoising, and related microscopy-image processing tasks (Lin et al., 2021). In atomic force microscopy, GraphAFM reconstructs molecular graphs from image stacks, and Ice-AFM studies structure discovery in water-cluster images (Oinonen et al., 2022; Priante et al., 2024). STM-Net and related image-to-structure models further investigate the reconstruction of atomic information from scanning-tunneling-microscopy observations (Zhu et al., 2025). These works provide task settings and image-to-structure benchmarks relevant to our study, while their outputs and imaging modalities remain distinct.

Machine-learned force fields and equivariant atomistic models. Machine-learned force fields have evolved from invariant descriptors and local-environment functions toward neural architectures with explicit geometric structure. MACE and later atomistic foundation models use equivariant representations to encode local environments across chemical and materials datasets (Batatia et al., 2022; 2025). MACE-OFF23 provides a transferable force-field model for organic molecules and is used here as a structural teacher for the STM setting (Kovács et al., 2025). Hessian distillation and DPA-2 distillation demonstrate transfer between atomistic force-field models, but their students remain atomistic predictors (Amin et al., 2025; Zhang et al., 2024).

Cross-modal alignment and knowledge distillation. CLIP aligns paired image and text encoders through contrastive learning (Radford et al., 2021). TinyCLIP studies affinity-based distillation and weight inheritance, while Proteus transfers visual foundation-model representations to smaller students without access to the original pretraining data (Wu et al.,

2023; Zhang et al., 2025). MatterChat connects frozen CHGNet or MACE encoders to a language model through a trainable bridge for structure-conditioned question answering (Tang et al., 2026). These studies motivate projected embeddings, contrastive objectives, and training-time bridges, while our teacher represents three-dimensional atomic environments and the structural modality is removed at inference.

Equivariant image models and cross-domain geometry. Steerable $E(2)$-equivariant CNNs encode known transformation laws for planar image rotations (Weiler & Cesa, 2019). Cross-Domain 3D Equivariant Image Embeddings and Image to Sphere connect image features with three-dimensional equivariant representations for image understanding and pose prediction (Esteves et al., 2019; Klee et al., 2023). Induced and restricted representations and ray-space SE(3) operators further show how observation geometry can constrain cross-domain mappings (Howell et al., 2023; Xu et al., 2023). Our setting connects a three-dimensional atomistic teacher to spatial image features through an explicit imaging frame and compatible representation types.

Positioning of this work. Our work lies at the intersection of scientific image reconstruction, equivariant atomistic modeling, and cross-modal knowledge transfer. Compared with existing image-analysis and image-to-structure models, we use a pretrained atomistic model as a training-only structural teacher. Compared with standard feature alignment, we explicitly account for the distinction between three-dimensional atomistic representations and two-dimensional image representations. Compared with purely equivariant image models, we are designed to test whether transferred atomistic information improves an image task and whether any improvement is reflected in controlled rotation behavior under the specified protocols. The comparisons are designed to distinguish atomistic knowledge from initialization, invariant alignment, and, once implemented, the proposed geometric interaction.

7 Discussion and Conclusion

The completed experiments support conditional transfer from atomistic models to scientific images. In the DenseNet AtomVision comparison, invariant MLFF supervision slightly improves native accuracy but increases rotation loss. In the matched STM comparison, the SO(2) backbone has substantially lower rotation loss than the CNN backbone, while invariant MLFF supervision does not provide a consistent additional reduction. These results show that structural supervision and image-backbone geometry must be evaluated separately. They do not establish universal MLFF gains, exact continuous equivariance, or a validated general SO(2)–SO(3) interaction.

The present evidence is limited to the stated tasks, data representations, and training protocols. AtomVision classification uses simulated images, and its JPEG and numerical-image pipelines are not interchangeable controls. Finite support and interpolation can change observable information, while graph construction and discrete decoding add further sources of variation. Most current comparisons use one training seed and do not support across-seed uncertainty estimates. The AFM numbers reported here are development-validation observations, not a completed fifty-epoch held-out evaluation. Warm initialization also inherits author pretraining that is absent from the corresponding random initialization. These limitations constrain the scope of empirical conclusions without changing the architectural transformation laws being investigated.

Geometric compatibility, observability, and task relevance are separate conditions. A structurally valid teacher feature may encode a component that cannot be inferred from a single projection, or that is not useful for the downstream label. Under squared-error supervision, an image-only predictor can at best learn the conditional mean of an uncertain target given its available image information. An invariant target is therefore not automatically inferior to a directional one. Teacher-domain coverage and overlap between foundation-model pretraining and downstream structures also require an explicit provenance audit; no disjoint-pretraining claim is made in this draft.

Pending. Finalize this abstract after the geometric operator, its analysis, and matched experiments are complete. Current numerical statements refer only to the completed comparisons in Section 5.

Pending. Complete the contribution statement only after the new operator has a precise definition and validation. Existing SO(2) classification and view-frame directional distillation are completed baselines, not evidence that the proposed general interaction has already succeeded.

Pending. Replace the provisional conclusion with the final scope of evidence after model selection and the designated test evaluations. Do not infer a successful new geometric method from the current invariant-transfer results.

AI use statement

Generative AI tools were used for literature retrieval, research and experimental-design discussions, implementation and debugging, experiment orchestration, and drafting and editing text. These tools were not treated as scientific evidence. The authors are responsible for the final claims, analyses, and citations.

References

Ishan Amin, Sanjeev Raja, and Aditi S. Krishnapriyan. Towards fast, specialized machine learning force fields: Distilling foundation models via energy hessians. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=1durmugh3I.

Ilyes Batatia, David P. Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Advances in Neural Information Processing Systems, volume 35, pp. 11423–11436, 2022. doi: 10.52202/068431-0830. URL https://papers.neurips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. The Journal of Chemical Physics, 163(18):184110, November 2025. ISSN 1089-7690. doi: 10.1063/5.0297006. URL https://doi.org/10.1063/5.0297006.

Kamal Choudhary, Ramya Gurunathan, Brian DeCost, and Adam Biacchi. AtomVision: A machine vision library for atomistic images. Journal of Chemical Information and Modeling, 63(6):1708–1722, 2023. doi: 10.1021/acs.jcim.2c01533. URL https://pubs.acs.org/doi/10.1021/acs.jcim.2c01533.

Carlos Esteves, Avneesh Sud, Zhengyi Luo, Kostas Daniilidis, and Ameesh Makadia. Cross-domain 3D equivariant image embeddings. In Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 1812–1822. PMLR, 2019. URL https://proceedings.mlr.press/v97/esteves19a.html.

Owen Howell, David Klee, Ondrej Biza, Linfeng Zhao, and Robin Walters. Equivariant single view pose prediction via induced and restricted representations. In Advances in Neural Information Processing Systems, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/93b3d975f9a2448964a906199db98a9d-Paper-Conference.pdf.

David M. Klee, Ondrej Biza, Robert Platt, and Robin Walters. Image to sphere: Learning equivariant features for efficient pose prediction. In The Eleventh International Conference on Learning Representations, 2023. URL https://arxiv.org/abs/2302.13926.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. Journal of the American Chemical Society, 147(21):17598–17611, 2025. doi: 10.1021/jacs.4c07099. URL https://pubs.acs.org/doi/10.1021/jacs.4c07099.

Ruoqian Lin, Rui Zhang, Chunyang Wang, Xiao-Qing Yang, and Huolin L. Xin. TEMImageNet training library and AtomSegNet deep-learning models for high-precision atom segmentation, localization, denoising, and deblurring of atomic-resolution images. Scientific Reports, 11(1):5386, 2021. doi: 10.1038/s41598-021-84499-w. URL https://www.nature.com/articles/s41598-021-84499-w.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Ruben Millan-Solsona, Marti Checa, Spenser R. Brown, Amber N. Bible, Bernadeta Srijanto, Laura Wiggins, Sita Sirisha Madugula, Alice L. B. Pyne, Jennifer L. Morrell-Falvey, Scott Retterer, Rama K. Vasudevan, and Liam Collins. Synthetic data-driven deep learning for label-free autonomous atomic force microscopy. Nature Communications, 17:3886, 2026. doi: 10.1038/s41467-026-70421-3. URL https://www.nature.com/articles/s41467-026-70421-3.

Niko Oinonen, Lauri Kurki, Alexander Ilin, and Adam S. Foster. Molecule graph reconstruction from atomic force microscope images with machine learning. MRS Bulletin, 47(9):895–905, July 2022. ISSN 1938-1425. doi: 10.1557/s43577-022-00324-3. URL https://doi.org/10.1557/s43577-022-00324-3.

Fabio Priante, Niko Oinonen, Ye Tian, Dong Guan, Chen Xu, Shuning Cai, Peter Liljeroth, Ying Jiang, and Adam S. Foster. Structure discovery in atomic force microscopy imaging of ice. ACS Nano, 18(7):5546–5555, February 2024. ISSN 1936-086X. doi: 10.1021/acsnano.3c10958. URL https://doi.org/10.1021/acsnano.3c10958.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Yingheng Tang, Wenbin Xu, Jie Cao, Weilu Gao, Steven Farrell, Benjamin Erichson, Michael W. Mahoney, Andy Nonaka, and Zhi Jackie Yao. A multimodal large language model for materials science. Nature Machine Intelligence, 8:588–601, 2026. doi: 10.1038/s42256-026-01214-y. URL https://www.nature.com/articles/s42256-026-01214-y.

Maurice Weiler and Gabriele Cesa. General $E(2)$-equivariant steerable CNNs. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/45d6637b718d0f24a237069fe41b0db4-Abstract.html.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 21970–21980, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wu_TinyCLIP_CLIP_Distillation_via_Affinity_Mimicking_and_Weight_Inheritance_ICCV_2023_paper.html.

Yinshuang Xu, Jiahui Lei, and Kostas Daniilidis. Equivariant light field convolution and transformer in ray space. In Advances in Neural Information Processing Systems, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/075b2875e2b671ddd74aeec0ac9f0357-Paper-Conference.pdf.

Duo Zhang, Xinzijian Liu, Xiangyu Zhang, Chengqian Zhang, Chun Cai, Hangrui Bi, Yiming Du, Xuejian Qin, Anyang Peng, Jiameng Huang, Bowen Li, Yifan Shan, Jinzhe Zeng, Yuzhi Zhang, Siyuan Liu, Yifan Li, Junhan Chang, Xinyan Wang, Shuo Zhou, Jianchuan Liu, Xiaoshan Luo, Zhenyu Wang, Wanrun Jiang, Jing Wu, Yudi Yang, Jiyuan Yang, Manyi Yang, Fu-Qiang Gong, Linshuang Zhang, Mengchao Shi, Fu-Zhi Dai, Darrin M. York, Shi Liu, Tong Zhu, Zhicheng Zhong, Jian Lv, Jun Cheng, Weile Jia, Mohan Chen, Guolin Ke, Weinan E, Linfeng Zhang, and Han Wang. DPA-2: a large atomic model as a multi-task learner. npj Computational Materials, 10:293, 2024. doi: 10.1038/s41524-024-01493-2. URL https://www.nature.com/articles/s41524-024-01493-2.

Yitian Zhang, Xu Ma, Yue Bai, Huan Wang, and Yun Fu. Accessing vision foundation models via ImageNet-1K. In The Thirteenth International Conference on Learning Representations, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/9156b0f6dfa9bbd18c79cc459ef5d61c-Paper-Conference.pdf.

Yu Zhu, Renjie Xue, Hao Ren, Yicheng Chen, Wenjie Yan, Bingzheng Wu, Sai Duan, Haiming Zhang, Lifeng Chi, and Xin Xu. Reconstructing pristine molecular orbitals from scanning tunneling microscope images via artificial intelligence approaches. JACS Au, 5(7):3163–3170, 2025. doi: 10.1021/jacsau.5c00310. URL https://pubs.acs.org/doi/10.1021/jacsau.5c00310.

A Observation-Frame Registration and Representation Matching

Let $F = [e_x, e_y, n] \in \mathrm{SO}(3)$ be a right-handed observation frame whose columns are written in world coordinates, and let o denote its origin. The coordinates of a structural position r in this frame are

$$r^F = F^\top (r - o). \tag{10}$$

For a teacher vector v, the corresponding frame coordinates are $v^F = F^\top v$.

For a displayed in-plane rotation α, the registered frame can be written as

$$F' = QF\overline{R}_{-\alpha}, \qquad \overline{R}_\alpha = \mathrm{diag}(R_\alpha, 1). \tag{11}$$

This convention determines the sign of the planar rotation and specifies how directional teacher components are expressed in the image frame.

Restricting an SO(3) irreducible representation to the imaging-related SO(2) subgroup gives

$$D^{(\ell)}\Big|_{\mathrm{SO}(2)} \cong \rho_0 \oplus \rho_1 \oplus \cdots \oplus \rho_\ell. \tag{12}$$

For $\ell = 1$, the component parallel to the imaging normal has planar type $m = 0$, while the two in-plane components form an $m = 1$ representation. For C copies of a planar vector representation and C' output copies, a same-type channel map $W \otimes I_2$ mixes copies without mixing the two spatial components:

$$(W \otimes I_2)(I_C \otimes R_\alpha) = (I_{C'} \otimes R_\alpha)(W \otimes I_2). \tag{13}$$

B Local Transfer Objective and Implementation Details

For local transfer, let s_{ba} denote the student feature sampled at the projected location of atom a in sample b, let t_{ba} denote the corresponding standardized teacher target, let n_b be the number of valid atoms, and let A be the active target-channel set. A trainable projection P is optimized with

$$\mathcal{L}_{\text{local}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{\max(n_b,1)|A|}\sum_{a=1}^{n_b}\sum_{c\in A}\left(P(s_{ba})_c - t_{ba,c}\right)^2. \tag{14}$$

Teacher-target standardization is fitted using training atoms only. Samples with no valid local atoms contribute zero local loss. Projection dimensions are selected to match the feature dimensions of each student architecture. The teacher remains frozen and the local target is detached from the gradient graph.

For the AFM local objective, Graph-AFM uses an $8 \to 128 \to 384$ projection, while Ice-AFM uses a $32 \to 128 \to 384$ projection. The invariant teacher descriptor contains scalar channels and vector norms. Task-specific feature sampling, channel masks, and normalization constants are recorded with the corresponding experimental protocol.

C Implementation and Training Budgets

The registered AFM comparisons use seed 0, full data, and one selected auxiliary objective per treatment. Graph uses 180,000 training views and 3,750 updates per epoch at global batch 48; Ice uses 16,400 views and 513 updates at global batch 32. Both target 50 downstream epochs per method. Graph warm conditions continue their own method-specific epoch-10 states; scratch conditions start from the same random base model. Ice scratch initializes PosNet randomly while retaining the frozen author GraphImgNet branches for atom classes and bonds. Thus, scratch has different scopes across tasks, and equal downstream epochs do not imply equal lifetime pretraining cost. Every completed epoch is retained as a distinct checkpoint with available optimizer, scheduler, and random-state information.

The AFM teacher is frozen MACE-MPA-0. Its transferred descriptor contains 256 scalar channels and 128 vector norms. Normalization is fitted on training atoms only. Ice teacher extraction includes the supplied gold environment as a finite cluster; gold atoms are excluded from fitting the H/O target scaler. GPU migrations preserve global batch and the original virtual-rank augmentation grouping, but are not claimed to produce bitwise-identical trajectories. Hardware and execution versions therefore remain part of the experimental record.

The completed AtomVision SO(2) comparison uses 6,450 training, 1,324 validation, and 1,376 test images with composition-grouped splits. All four arms use the same 1,681,931-parameter classification backbone, seed 0, batch 16, and 50 epochs: 322,500 sample presentations and 20,200 updates. Training uses Adam at 2×10^{-4} in FP32, with TF32 disabled and no random rotation augmentation. The local-norm and directional arms share the global invariant objective, local sites, teacher channels, and same-type adapter; their local losses retain vector length or both planar components, respectively. This completed classification study is a baseline for the proposed general SO(2)-SO(3) interaction.

D Data Splits and Evaluation Protocol

AFM model selection uses full validation task loss: 19,997 Graph views or 1,220 Ice views per epoch. Separate reconstruction panels contain 256 unique Graph validation molecules and 244 Ice views from 122 structures. These panels are development validation, not held-out test results. At 37°, scoring uses the intersection of original and rotated fields of view with a 0.25 Å boundary margin and a corresponding unrotated common-support reference; 90° retains the full-field convention. Atom matching first maximizes the number of one-to-one matches within 0.35 Å, then minimizes distance. Empty predictions remain in the error

counts. Graph class scores concern the author's five element groups, not exact elemental identification.

The completed AtomVision SO(2) study selects checkpoints by validation negative log likelihood, including epoch 0, and separately evaluates the fixed epoch-50 endpoint. Each selected checkpoint is tested on all 1,376 test images at 24 angles. Rotation means average orientations of the same test set; they are not independent training replicates. Inference receives images only. Current seed-0 studies do not estimate variability across training seeds, and no unmeasured dense-angle AFM results are inferred from the three validation angles.

E Additional Quantitative Results

The completed TEMImageNet transfer experiment trains the original AtomSegNet CNN from scratch for 100 epochs per condition at seed 0 (Lin et al., 2021). At validation-selected checkpoints, plain, weak, and strong transfer give test MSEs of 0.00378529, 0.00375981, and 0.00408545, respectively. Localization F1 is 97.36%, 96.92%, and 96.85%. Weak transfer therefore slightly improves MSE while worsening localization. Image metrics include all test samples, whereas localization uses verified structure registration. The teacher operates on finite atom clouds with nonperiodic boundaries; this experiment does not establish periodic-bulk transfer or a consistent benefit across metrics.

A separate historical three-seed $sp \rightarrow s$ reconstruction study is retained without adding seeds. Validation selected the original epoch-0 checkpoint in every condition. At the fixed 100-epoch endpoint, mean in-distribution error was 0.00561708 for plain continuation, 0.00569749 for weak transfer, and 0.00615079 for strong transfer. These endpoint means must not be substituted for validation-selected outcomes or pooled with the current single-seed studies.

Table 4 records a fixed epoch-40 snapshot of the Ice warm continuation, separately from final model selection. All four methods completed the stated validation measurements. Density MSE uses 1,220 validation views, whereas reconstruction metrics use 244 views from 122 validation structures. At this endpoint, the plain control has the lowest MSE and the highest unrotated atom and bond F1; relational transfer has slightly higher 90° atom F1. These mixed results do not establish a stable advantage or the final ranking of the 50-epoch study.

Table 3: Fixed epoch-40 Ice validation snapshot. All methods have the same downstream epoch budget; this table does not report final selected checkpoints or held-out tests. F1 values are percentages.

Supervision	Density MSE	Atom F1 (0°)	Bond F1 (0°)	Atom F1 (90°)
None	0.030353	92.55	90.84	92.06
Relational invariant	0.030675	92.39	90.75	92.19
Local invariant	0.030710	92.27	90.76	91.91
Global contrastive	0.031972	92.07	90.52	91.47

Pending. Add the complete AFM initialization matrices and final learning curves after each registered condition finishes. Keep provisional validation snapshots separate from final held-out evaluation, and add new interaction results only after implementation and measurement.

F Qualitative Reconstructions and Failure Cases

Pending. Add full-resolution inputs, reference structures, predictions, and inverse-rotated overlays from a documented sample-selection rule. Include localization, atom-type, connectivity, and boundary failures. No qualitative panel is yet used here as evidence for the proposed interaction.