AutoMat: Enabling Automated Crystal Structure Reconstruction from Microscopy via Agentic Tool Use

Anonymous authors
Paper under double-blind review

Abstract

Machine learning models for interatomic potentials and force fields require high-quality structural data, yet experimental crystal structures remain limited, creating a critical gap between computational simulations and real structures. While atomic-resolution electron microscopy can provide valuable images, converting these into simulation-ready structures is time-consuming and error-prone. We present **AutoMat**, an agent-based framework that directly converts Scanning Transmission Electron Microscopy (STEM) images to atomistic crystal structures for property prediction. The framework adaptively calls tools, including pattern-adaptive denoising, physics-guided template retrieval, symmetry-aware atomic reconstruction, and property prediction, enabling closed-loop reasoning from "image → structure → property." For systematic evaluation, we introduce **STEM2Mat-Bench**, a benchmark dataset containing 450+ annotated samples. Performance is assessed using lattice root-mean-square deviation (RMSD), formation energy mean absolute error (MAE), and structure matching accuracy. Results demonstrate that AutoMat outperforms existing approaches including SOTA models, specialized domain tools, and closed-source multimodal large models. This work establishes a direct pathway from microscopic characterization to atomic-scale modeling, addressing a fundamental challenge in materials science.

1 Introduction

Machine learning interatomic potentials and force fields now approach *ab initio* accuracy in predicting atomic energies and forces (Yang et al., 2024; Batatia et al., 2024; Gasteiger et al., 2021; Liao et al., 2024; Lu et al., 2023) but remain limited by experimental validated crystal structures. Meanwhile, Scanning Transmission Electron Microscopy (STEM) can image atoms at sub-ångström resolution (Zhang et al., 2018; 2023; 2020; Ortalan et al., 2010), yet translating into quantitative crystal structures still relies on expert-driven, time-consuming annotation. This creates a gap between structural representation and theoretical validation in materials science (Kalinin et al., 2022; 2023).

Although recent advances in STEM image analysis have shown promise, most existing studies focus on individual components like denoising (Yang et al., 2025; Lin et al., 2020; 2021; Wang et al., 2020), atom localization (Ziatdinov et al., 2017; Borshon et al., 2024; Eliasson & Erni, 2024), reconstruction (De Backer et al., 2022; Stoops et al., 2025; Huang et al., 2025) and phase classification (Li et al., 2021; Maksov et al., 2019). These approaches remain fragmented and are not integrated into end-to-end system. Conventional image denoising techniques suppress noise and improve contrast at the pixel level but cannot yield periodic or chemically meaningful crystal structures (Gambini et al., 2023; Ihara et al., 2022; Wang et al., 2020; Yang et al., 2025). Atomic detection models can localize atomic peaks but cannot infer complete lattices[1] or identify atomic species. General-purpose multimodal models like GPT-4.1mini (Achiam et al., 2023) and Qwen2.5-VL (Bai et al., 2025) exhibit basic image understanding but lack the ability to produce simulation-ready formats such as crystallographic file (CIF). Even domain-specific tools like AtomAI (Ziatdinov et al., 2022) and the newly released domain SOTA model MicroscopyGPT (Choudhary, 2025) can only predict atomic coordinates and structural descriptions for minimal systems, without supporting

[1] In crystallography, a lattice refers to the periodic arrangement of atoms in a crystal structure.

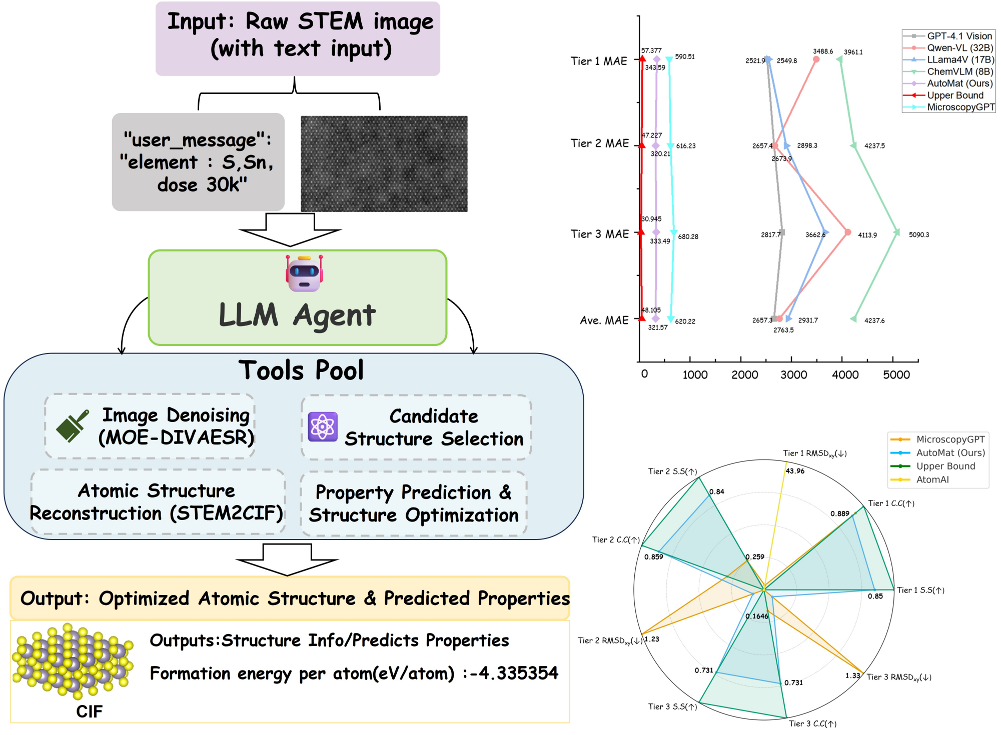


Figure 1: **Overview of AutoMat.** The left part shows an LLM agent how to manage four modules—pattern-adaptive denoising, template selection, atomic reconstruction (STEM2CIF), and ML-based property prediction. The right panel presents a line chart and radar plot comparing different models in terms of energy and structural errors.

complete CIF structure reconstruction or property prediction. Meanwhile, public datasets (e.g., JARVIS-STM) mainly target STM (Scanning Tunneling Microscopy) images, lack DFT-level energy labels, and are unsuitable for benchmarking structure–property pipelines. As a result, the field lacks a fully automated end-to-end system that can convert raw STEM images into reconstructed structures and simulated properties, along with the standardized benchmark for comprehensive evaluation.

To address this gap, as shown in Fig. 1, we introduce **AutoMat**, the first agent-based framework linking STEM imaging with atomistic simulation, along with a benchmark tailored to the task. At the core of AutoMat is an intelligent agent equipped with a suite of four powerful modular tools, which it can dynamically orchestrate to solve the task: 1) **Pattern-Adaptive STEM Image denoising**: We apply a pattern-adaptive mixture-of-experts network, MOE-Denoising Inference Variational Autoencoder Super-Resolution (Yang et al., 2025)(MOE-DIVAESR), to denoise and enhance raw iDPC-STEM[2] images. The ResNet-18-based gating network elects the most suitable expert network for each input image based on its estimated noise level, enabling joint denoising, inpainting, and super-resolution. 2) **Physics-Guided Template Retrieval**: Enhanced images are matched to a large-scale library of simulated STEM projections. Top candidate structures are retrieved using pixel similarity and filtered by elemental contrast patterns to produce strong structural priors. 3) **Symmetry-Constrained Structure Reconstruction**: Atomic peaks are detected via unsupervised clustering. We fit the lattice under symmetry constraints, assign atomic species based on the candidate, and generate the standard CIF file representing the periodic crystal structure. 4) **Energy Evaluation via Machine-Learned Potential**: The reconstructed structure is relaxed using the pretrained MatterSim potential to predict formation energy. The language agent autonomously orchestrates the workflow, planning tool execution and adaptively retrying failed steps based on quality checks.

To support rigorous evaluation, we curated 2,143 high-quality monolayer structures from C2DB (Haastrup et al., 2018), Materials Project (Jain et al., 2013), and OpenCrystal (Vaitkus et al., 2023), and

[2] iDPC-STEM stands for integrated Differential Phase Contrast Scanning Transmission Electron Microscopy, a technique that enhances light element contrast in atomic-resolution imaging.

simulated their corresponding iDPC-STEM images with abTEM (Madsen & Susi, 2020). From this pool we selected 450 representative image–structure pairs, which constitute our **STEM2Mat** benchmark and are used for all subsequent evaluations. Our evaluation metrics include projected lattice RMSD, formation energy MAE, and atom-wise structure matching success rate. Additionally, we introduce three fine-grained indicators to assess reconstruction quality (e.g., atomic recall/precision), robustness across noise levels, and computational efficiency. On this benchmark, AutoMat achieves a projected RMSD of 0.11 ± 0.03 Å, energy MAE below 350 meV/atom, and an atomic correspondence total success rate of 83.2%, outperforming GPT-4.1-mini, Qwen-VL, LLama4V, ChemVLM (Li et al., 2025), MicroscopyGPT and AtomAI by an order of magnitude. These results demonstrate AutoMat as a reproducible and accurate end-to-end solution for microscopy-driven materials modeling.

The contributions of this paper are summarized as follows:

• **Agent-based Autonomous Orchestration**. AutoMat proposes an intelligent agent that selectively orchestrates capabilities for denoising, template retrieval, reconstruction, and relaxation, dynamically charting a path from raw STEM images to material property predictions.

• **Algorithmic advancements**. We design MOE-DIVAESR as a pattern-adaptive denoiser for diverse STEM images, enabling efficient noise reduction, defect correction, and detail enhancement. We also develop STEM2CIF to reconstruct crystal structures by identifying the minimal repeating unit using symmetry heuristics and physical constraints, then converting the result into standard CIF format.

• **STEM2Mat benchmark & evaluation suite**. We release a dataset of 2,143 distinct supercell structures used to simulate large-field STEM images. In addition, we provide a 450-case test split with DFT-calculated energies and establish unified metrics to evaluate lattice reconstruction accuracy, energy fidelity, robustness, and computational efficiency.

• **State-of-the-art results with open access**. AutoMat outperforms leading large multimodal models by an order of magnitude. All code, data, and evaluation tools are publicly released to support reproducibility and future development.

2 Related Work

Automated Microscopy Image Analysis. In recent years, deep learning methods have been extensively applied to electron microscopy and scanning probe microscopy data analysis (Yang et al., 2025; Lin et al., 2020; 2021; Wang et al., 2020; Ziatdinov et al., 2017; Borshon et al., 2024; De Backer et al., 2022; Stoops et al., 2025; Huang et al., 2025). Current approaches range from unsupervised defect detection to supervised atomic column identification. AtomAI (Ziatdinov et al., 2022), for example, integrates microscopy images with computational simulations, but primarily focuses on atom segmentation and identification. More recently, SciLink (Yao et al., 2025) proposes an open-source *multi-agent* framework that closes the loop between microscopy/spectroscopy experiments, literature-based novelty assessment, and theory-in-the-loop simulations, and has demonstrated impressive performance in automated defect localization and atomic-scale analysis. However, its primary focus is on serendipity-aware, high-level experiment–theory orchestration. It does not yet provide a dedicated solution for reconstructing electron-microscopy images into explicit crystal structures that can be quantitatively compared with theoretical models, nor for using such reconstructed structures as direct inputs to downstream simulations and property prediction.

STEM Image to Structure Reconstruction. Existing methods for reconstructing crystal structures from STEM images typically rely on multiple images, prior structural information, or are limited to single-element systems. For instance, De Backer et al (De Backer et al., 2022) employed Bayesian algorithms for 3D reconstruction, but their approach was demonstrated only for simple single-element systems. Currently, few methods can generate standard crystallographic information files (CIF) from single experimental images, especially for complex multi-element 2D crystals (Stoops et al., 2025).

Vision-Language Models in Chemistry. Recently, multimodal large language models (ChemVLM (Li et al., 2025), GPT-4.1mini (Achiam et al., 2023) and Qwen-VL (Bai et al., 2025)) have started to be explored in chemistry and materials science. However, these models generally lack the capability to accurately handle detailed spatial structure tasks. Existing chemical agent tools (e.g., ChemCrow (Bran et al., 2023) and Chemagents (Tang et al., 2025a)) are limited to text-based descriptions and cannot process image-based inputs, significantly restricting their applicability in microscopy-based analyses.The field's SOTA multimodal model MicroscopyGPT (Choudhary, 2025) can generate structural descriptions from STEM images but cannot yet reconstruct CIFs or predict properties.

Material Property Prediction Models. Advances in machine learning-based interatomic potentials (such as MatterSim (Yang et al., 2024), M3GNet (Chen & Ong, 2022), and MACE (Batatia et al., 2023)) have significantly improved the accuracy of computational property predictions. Combining these models with experimental imaging provides a novel, digital twin-like approach for validating structural reconstructions. However, existing benchmarks predominantly evaluate models using theoretical datasets, lacking end-to-end assessments starting from experimental image inputs.

In summary, these approaches highlight progress and limitations in microscopy image analysis, structure reconstruction, multimodal modeling, and property prediction. **AutoMat addresses these gaps** by integrating an automated agent from STEM images to material property prediction and establishes the STEM2Mat benchmark for evaluating the robustness and scalability of automated material characterization.

3 Dataset Construction

3.1 Composition and Taxonomy

Our benchmark focuses on two-dimensional (2D) materials, whose atomic-scale thickness allows STEM to resolve individual columns with minimal multiple-scattering artifacts. Starting from nearly 10,000 candidate structures harvested from C2DB, Materials Project, and OpenCrystal,[3] we followed a two-stage curation process. First, automated filters removed non-stoichiometric, partially occupied, or 3D bulk entries. Second, domain experts inspected symmetry, cleavage energy, and substrate feasibility, yielding **2,143** high-confidence monolayer crystals. The collection spans six chemical families (Fig. 5): *(i)* classic 2D materials—graphene, MoS_2, h-BN, black phosphorus; *(ii)* emergent allotropes, e.g. silicene, borophene; *(iii)* conductive MXenes (23 distinct formulas); *(iv)* intrinsic 2D magnets such as CrI_3 and Fe_3GeTe_2; *(v)* Janus structures typified by MoSSe; and *(vi)* Ruddlesden–Popper–type 2D perovskites. Elemental diversity is broad: 67 unique elements appear, producing 76 unary 1,409 binary, and 658 ternary systems. Each structure is stored as a CIF file with validated lattice vectors and fractional atomic coordinates.

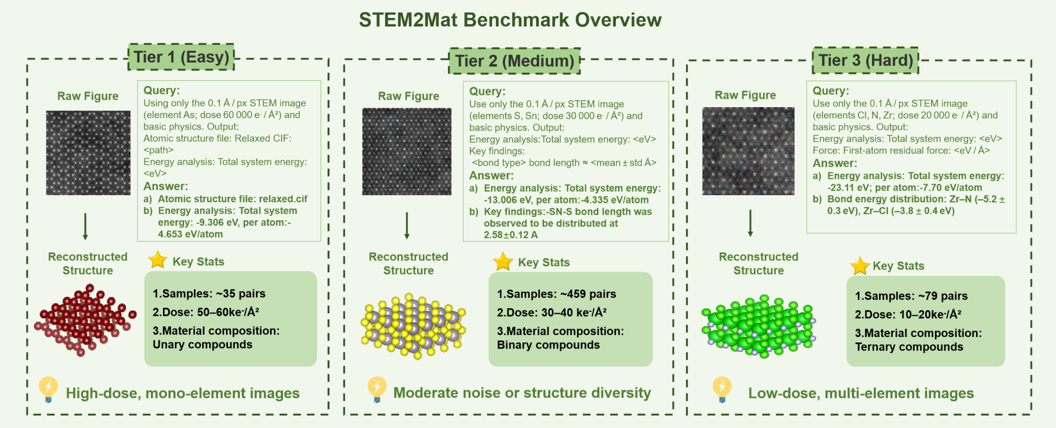
3.2 Image Simulation and Data Augmentation

To simulate realistic large-field STEM imaging conditions, we generated synthetic iDPC-STEM micrographs using the open-source abTEM simulation engine. For each structure, a random 12×12 to 16×16 supercell was constructed and projected at $0.1\,\text{Å}/\text{pixel}$ resolution. Five electron-dose settings (1–6 $\times$ $10^4\,\mathrm{e}^-/\text{Å}^2$) and realistic lens aberrations were sampled to mimic experimental conditions. Poisson detector noise was injected to match reported signal-to-noise ratios. To study model robustness, we applied Gaussian blurring and dose-specific shot noise to simulate additional imaging imperfections. Ground-truth atomic coordinates were rendered into Gaussian "atom masks" to enable supervised training of localisation models. Each sample thus forms an image–structure–property triplet: (i) the noisy STEM projection, (ii) the corresponding ground-truth CIF, and (iii) DFT-level formation energy (along with band gap and magnetic moment, if available). We conducted principal component analysis (PCA) on structural fingerprints, which revealed clear clustering patterns. Subsequently, k-means clustering (k=2, 6) was used to ensure balanced train/validation/test splits across the chemical diversity of the dataset.

3.3 STEM2Mat Benchmark Split and Tiering

To construct a representative and tractable benchmark dataset for STEM-based crystal modeling, we applied stringent geometric and chemical screening criteria to the 2,143 collected 2D material structures, ensuring their suitability for monolayer imaging and end-to-end reconstruction. Specifically, we retained only structures containing no more than three distinct elements. For those with multiple elements, we required a minimum atomic-number span of ten, i.e., $\max(Z_i) - \min(Z_i) \geq 10$, to ensure sufficient imaging contrast between heavy and light atoms. To guarantee monolayer geometry, we limited the z-axis thickness to no more than 3 Å. Each structure's atomic coordinates were then

[3] All sources were queried in April 2025 using identical monolayer filters.



Query:
Using only the 0.1 Å / px STEM image (element As; dose 60 000 e- / Å²) and basic physics. Output:
Atomic structure file: Relaxed CIF: <path>
Energy analysis: Total system energy: <eV>

Answer:
a) Atomic structure file: relaxed.cif
b) Energy analysis: Total system energy: -9.306 eV, per atom:-4.653 eV/atom

1.Samples: ~35 pairs
2.Dose: 50–60ke-/Å²
3.Material composition: Unary compounds

Query:
Use only the 0.1 Å / px STEM image (elements S, Sn; dose 30 000 e- / Å²) and basic physics. Output:
Energy analysis:Total system energy: <eV>
Key findings:
<bond type> bond length ≈ <mean ± std Å>

Answer:
a) Energy analysis: Total system energy: -13.006 eV, per atom:-4.335 eV/atom
b) Key findings:-SN-S bond length was observed to be distributed at 2.58±0.12 A

1.Samples: ~459 pairs
2.Dose: 30–40 ke-/Å²
3.Material composition: Binary compounds

Query:
Use only the 0.1 Å / px STEM image (elements Cl, N, Zr; dose 20 000 e- / Å²) and basic physics. Output:
Energy analysis: Total system energy: <eV>
Force: First-atom residual force: <eV / Å>

Answer:
a) Energy analysis: Total system energy: -23.11 eV; per atom:-7.70 eV/atom
b) Bond energy distribution: Zr–N (–5.2 ± 0.3 eV), Zr–Cl (–3.8 ± 0.4 eV)

1.Samples: ~79 pairs
2.Dose: 10–20ke-/Å²
3.Material composition: Ternary compounds

Figure 2: **Overview of the STEM2Mat Benchmark design**, illustrating the tiered classification of STEM samples by material complexity and imaging dose, which systematically stratifies reconstruction difficulty from simple unary to complex ternary compounds.

projected onto the (x, y) plane, discretized onto a 1 Å grid, and evaluated for overlapping projections. Only structures with a projected duplication ratio below 10%, i.e., $\frac{\text{Number of overlapping grid points}}{\text{Total grid points}} \leq 0.1$, were retained to avoid ambiguity in atomic interpretation.

Following this multi-criteria filtering process, we retained approximately **450** well-defined, unambiguous monolayer structures—21% of the original dataset—for blind end-to-end evaluation. The remaining 1,693 samples were split into training (80%) and validation (20%) sets to support model training and tuning.

To analyze model performance as a function of task difficulty, we stratified the test set into three tiers based on material composition and imaging noise (see Fig. 2):

- **Tier 1** – Unary materials acquired under high electron doses (5–6 $\times$ 10^4 $\mathrm{e}^-/\mathring{\mathrm{A}}^2$); these images are high contrast, low noise, and represent the easiest cases (35 samples).

- **Tier 2** – Binary materials or moderate electron dose conditions (3–4 $\times$ 10^4 $\mathrm{e}^-/\mathring{\mathrm{A}}^2$); these represent intermediate complexity in both noise level and atomic diversity (456 samples).

- **Tier 3** – Ternary compounds imaged at low dose (1–2 $\times$ 10^4 $\mathrm{e}^-/\mathring{\mathrm{A}}^2$); these samples exhibit high noise, complex contrast, and are the most challenging to reconstruct (79 samples).

While the tier sizes are imbalanced, this hierarchical structure reveals a clear gradient of reconstruction difficulty, establishing a principled ladder for evaluating robustness and scalability. With the STEM2Mat-Bench tiers defined, we now specify the quantitative metrics used throughout the paper. For detailed definitions and formulae of the evaluation metrics, see evaluation metrics 3.4.

3.4 Evaluation Metrics

To compare methods reproducibly, we report two *primary* metrics—energy error and lattice error—and two *holistic* metrics that reflect chemical and structural validity.

Mean Absolute Error (MAE). Average difference between predicted and DFT formation energies:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|E_i^{\mathrm{pred}} - E_i^{\mathrm{ref}}\right| \; [\mathrm{meV/atom}] \tag{1}$$

Projected Lattice RMSD. Deviation of in-plane lattice constants:

$$\mathrm{RMSD}_{xy} = \sqrt{\tfrac{1}{2}\Big[(a^{\mathrm{pred}} - a^{\mathrm{ref}})^2 + (b^{\mathrm{pred}} - b^{\mathrm{ref}})^2\Big]} \tag{2}$$

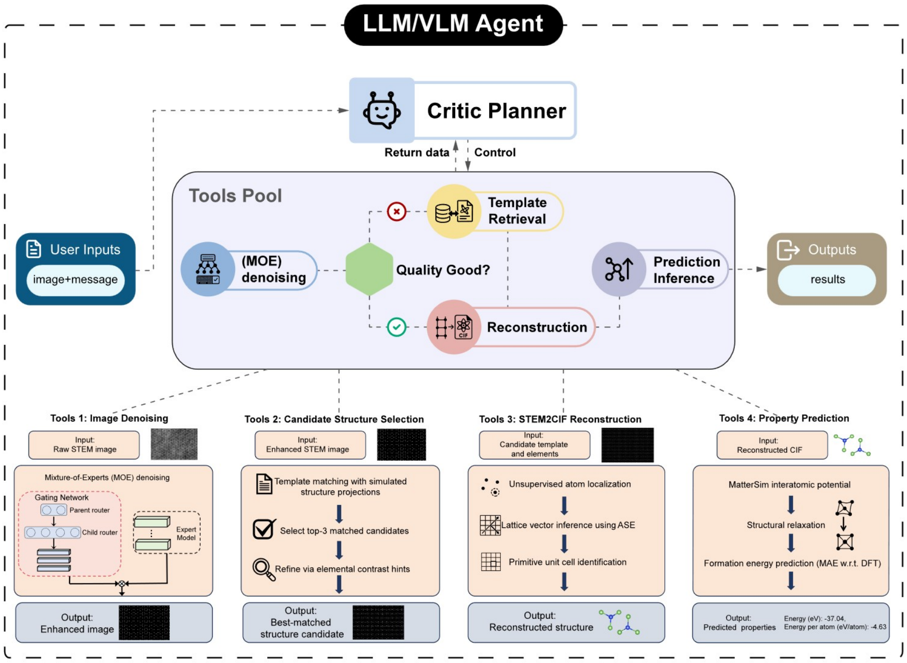

LLM/VLM Agent
Critic Planner
Return data
Control
Tools Pool
User Inputs
image+message
(MOE) denoising
Quality Good?
Template Retrieval
Reconstruction
Prediction Inference
Outputs
results
Tools 1: Image Denoising
Input: Raw STEM image
Mixture-of-Experts (MOE) denoising
Gating Network
Parent router
Child router
Expert Model
Output: Enhanced image
Tools 2: Candidate Structure Selection
Input: Enhanced STEM image
Template matching with simulated structure projections
Select top-3 matched candidates
Refine via elemental contrast hints
Output: Best-matched structure candidate
Tools 3: STEM2CIF Reconstruction
Input: Candidate template and elements
Unsupervised atom localization
Lattice vector inference using ASE
Primitive unit cell identification
Output: Reconstructed structure
Tools 4: Property Prediction
Input: Reconstructed CIF
MatterSim interatomic potential
Structural relaxation
Formation energy prediction (MAE w.r.t. DFT)
Output: Predicted properties
Energy (eV): -37.04,
Energy per atom (eV/atom): -4.63

Figure 3: **AutoMat's LLM agent orchestrates four stages**—denoising, template matching, structure reconstruction, and property prediction—from STEM image to relaxed crystal with properties.

Composition Correctness (C.C.). 1 if elemental types and counts match the reference; 0 otherwise.

Structure Success Rate (S.S.). A prediction is successful when it satisfies both chemistry and geometry. First define 2-D spatial similarity

$$S_{\text{spatial}} = \exp\big(-\text{MSE}_{\text{2D}}\big), \tag{3}$$

where MSE_{2D} is the mean-squared error of projected atomic positions after optimal element-wise matching. Then

$$\text{S.S.} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\Big[S_{\text{spatial}}^{(i)} \geq 0.8\Big] \times 100\%. \tag{4}$$

4 Overview of AutoMat

4.1 Limitations of Existing Large Multimodal Models and Domain Tools

Most current large multimodal models (e.g., GPT-4.1mini, LLama4V, Qwen2.5-VL, ChemVLM) focus on general-purpose image understanding tasks such as scene recognition, OCR, and molecular structure identification, but exhibit limited capability in interpreting scientific images like electron microscopy (EM). Domain-specific chemistry agents (e.g., ChemAgent, ChemCrow) remain predominantly text-based, executing tool calls for predefined tasks without closed-loop visual reasoning or image-guided decision-making. Specialized STEM data toolkits (e.g., AtomAI) provide atomic coordinate extraction and segmentation, yet their applicability is largely restricted to single-element nanoparticles and fail with complex, multi-element STEM images. MicroscopyGPT, the current state-of-the-art multimodal model in the domain, can generate structural descriptions from user-provided STEM images but still relies on known structural coordination and cannot directly reconstruct CIFs.

To date, existing agents are unable to autonomously generate simulation-ready CIFs and predict formation energies from a single real STEM image, thus completing a generalizable structure–property pipeline. To bridge this gap, we introduce **AutoMat**, an agent framework with advanced reasoning, enabling end-to-end modeling from pixel-level inputs to material property predictions.

4.2 Tools Pool

Within AutoMat—a general and efficient framework—we employ four core tools: MOE-DIVAESR, Image Template Matching, STEM2CIF, and MatterSim, responsible for image denoising, template matching, structure reconstruction, and property prediction, respectively. Further details are provided in Appendices A.2 and A.3.

MOE-DIVAESR is a structure-pattern–adaptive mixture-of-experts (MoE) model for STEM image denoising. Trained on an augmented STEM dataset, it contains multiple expert networks, each specializing in a distinct structural pattern, while a gating network dynamically selects the appropriate experts based on the input image features. This design enables denoising, defect correction, and fine-detail enhancement, producing sharply resolved atomic columns.

Image Template Matching compares an enhanced image against a database of known structures. By combining image features with elemental information, this technique efficiently narrows down potential candidates and identifies the structure that best fits experimental data. The process is even faster and more reliable when researchers can directly match against pre-calculated or known candidate structures.

STEM2CIF combines the selected template with the enhanced image to reconstruct a high-fidelity atomic model. It locates atomic column positions, infers lattice parameters, and identifies atomic species. Crystallographic symmetry heuristics and physical constraints then reduce the model to its minimal repeating unit, yielding a crystallographic CIF file.

MatterSim is a pretrained machine-learning interatomic potential for rapid structural relaxation and property evaluation. Trained on large-scale DFT datasets, it attains near-DFT accuracy for energies and properties. Integrated with ASE, it enables structure optimization and fast estimation of quantities such as total energy, formation energy, and elastic moduli, providing a swift and accurate alternative to conventional DFT.

4.3 Flexible Tool-Calling Framework

Building on a pool of specialized tools, **AutoMat** establishes a generalizable agentic framework for flexible tool orchestration, as illustrated in Figure 3. In AutoMat, the LLM/VLM backbone serves purely as a *text-based tool orchestrator*: it never directly "sees" raw STEM images, but instead reasons over structured outputs returned by the denoising, template-matching, and STEM2CIF modules and uses them to re-plan subsequent tool calls in a hybrid, state-dependent workflow. In practice, the input is a raw STEM image, and the agent dynamically selects appropriate tools at different steps according to the current input state, aiming to achieve accurate structural reconstruction and property prediction. Specifically, when the input contains substantial noise, such as in raw STEM images, the agent calls **MOE-DIVAESR** for denoising and structural enhancement. If a single denoising step does not meet the predefined threshold, MOE-DIVAESR can be repeatedly applied. Conversely, when the input quality is sufficiently high, the agent invokes **STEM2CIF** to localize atomic peaks, fit lattice parameters, and assign atomic species, thereby reconstructing an atomic model and producing a standard CIF file. For particularly challenging samples, if repeated attempts fail to converge, the agent resorts to **Image Template Matching** to directly retrieve candidate structures, select the most consistent template, and rapidly improve output quality. At the final step, **MatterSim** is automatically called to relax the reconstructed structure and predict properties such as total energy and formation energy. Throughout the process, the agent continuously monitors intermediate results, executing rollbacks and retries when necessary to ensure both stability and accuracy.

We adopt **DeepSeekV3** as a representative backbone. Even when driven purely by a text-based model, the framework demonstrates effective and flexible tool orchestration with competitive accuracy. Importantly, AutoMat enables robust end-to-end automation from raw images to material property insights, without depending on the capability of any single model.

5 Experiments and Results

5.1 Quantitative Evaluation

Baseline Overview To evaluate **AutoMat**, we compare it against three baselines targeting reasoning, reconstruction, and oracle performance:

Table 1: **Formation-energy MAE (meV/atom) across tiers.** "↓" indicates lower is better; "–" indicates the method does not provide an energy prediction.

Method	Tier 1 ↓	Tier 2 ↓	Tier 3 ↓	Avg. ↓
GPT-4.1 Vision	2521.9	2657.4	2817.7	2657.3
Qwen-VL (32B)	3488.6	2673.9	4113.9	2763.5
LLama4V (17B)	2549.8	2898.3	3662.6	2931.7
ChemVLM (8B)	3961.1	4237.5	5090.3	4237.6
MicroscopyGPT (11B)	590.51	616.23	680.28	620.22
AtomAI	–	–	–	–
AutoMat (Ours, DeepSeek)	**343.59**	**320.21**	**333.49**	**321.57**
AutoMat (Ours, GPT-4o)	**341.72**	**322.05**	**334.12**	**323.25**
Upper Bound	57.377	47.227	30.945	48.105

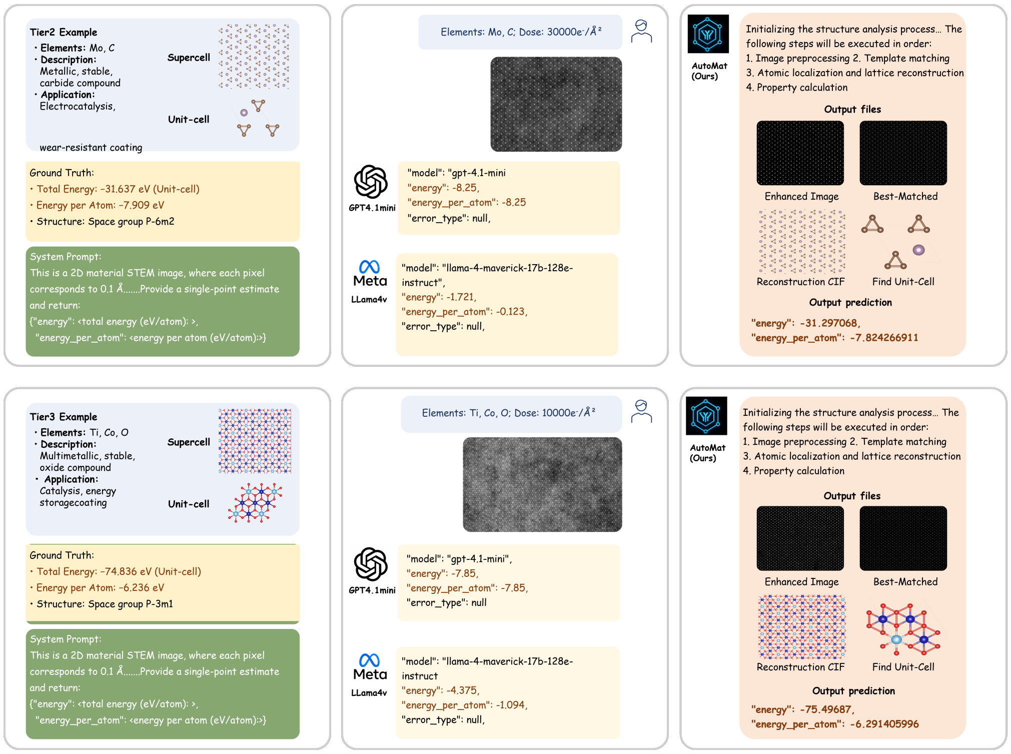
Table 2: **Structural-accuracy metrics for methods that output atomic models.** RMSD$_{xy}$↓: in-plane lattice root-mean-square deviation (lower is better); C.C.↑: composition correctness; S.S.↑: structure success rate.

Tier	Method	RMSD$_{xy}$ (Å)↓	C.C. (%)↑	S.S. (%)↑
1	AtomAI	43.96±0.31	2.70	0.0
	MicroscopyGPT	1.56±0.72	92.8	0
	AutoMat (Ours, DeepSeek)	**0.11±0.02**	**88.9**	**85.0**
	AutoMat (Ours, GPT-4o)	**0.11±0.02**	**89.2**	**85.3**
	Upper Bound	0 (def.)	100	100
2	AtomAI	N/A	0.0	0.0
	MicroscopyGPT	1.23±0.75	30.1	25.9
	AutoMat (Ours, DeepSeek)	**0.11±0.03**	**85.9**	**84.0**
	AutoMat (Ours, GPT-4o)	**0.12±0.02**	**86.4**	**84.7**
	Upper Bound	0 (def.)	100	100
3	AtomAI	N/A	0.0	0.0
	MicroscopyGPT	1.33±0.92	16.46	0
	AutoMat (Ours, DeepSeek)	**0.11±0.03**	**73.1**	**73.1**
	AutoMat (Ours, GPT-4o)	**0.11±0.02**	**72.8**	**72.8**
	Upper Bound	0 (def.)	100	100

- **Vision–Language Models (VLM).** GPT 4.1mini, Qwen2.5 VL (32B), LLama4V (17B), ChemVLM (8B) and MicroscopyGPT (11B) receive a fixed prompt, composition hints, and a STEM image to infer material properties, assessing multimodal reasoning.
- **AtomAI.** AtomAI's segmentation network detects atomic centers; relative coordinates plus image resolution are used to fit the lattice. This baseline measures reconstruction quality only.
- **Ground-Truth CIF + MLIP (Oracle Upper Bound).** The true CIF is fed directly to the MatterSim MLIP to benchmark formation-energy error under perfect-structure assumptions.

We summarize the performance of these baselines and our **AutoMat** system on the test set in Tables 1 and 2, which together cover samples across Tier 1–3 difficulty levels. As detailed in Appendix A.8, we further compare different LLM and VLM backbones as agent models on a test subset. Here we only present results using DeepSeek (LLM) and GPT-4o (VLM) as representative backbone models, but experiments show that regardless of which model is used as the backbone, AutoMat achieves comparable and consistent performance.

Discussion of Results. For energy prediction, **AutoMat** achieves a mean formation energy MAE of 332 ± 12 meV/atom, with tier-wise results of 343.59, 320.21, and 333.49 meV/atom. Although this is higher than the MLIP lower bound of 57 meV/atom, it is still far better than the multi-eV errors of vision–language models. Even compared with MicroscopyGPT—which may have been trained on data overlapping with our test set—AutoMat remains much stronger, with MicroscopyGPT reaching only about 620 meV/atom. As task difficulty increases, most baselines show higher MAE, confirming the soundness of our tiered benchmark. These results indicate that AutoMat's remaining errors come mainly from reconstruction rather than MLIP limits, and that its predicted structures are reliable for downstream property evaluation. Figure 4 shows two examples from Tier 2 and Tier 3, highlighting AutoMat's superior performance over GPT-4.1mini and LLaMA4, especially under low-dose and multi-element conditions.



Tier2 Example
- **Elements:** Mo, C
- **Description:** Metallic, stable, carbide compound
- **Application:** Electrocatalysis, wear-resistant coating

Ground Truth:
- Total Energy: −31.637 eV (Unit-cell)
- Energy per Atom: −7.909 eV
- Structure: Space group P-6m2

System Prompt:
This is a 2D material STEM image, where each pixel corresponds to 0.1 Å......Provide a single-point estimate and return:
{"energy": <total energy (eV/atom): >,
"energy_per_atom": <energy per atom (eV/atom):>}

Elements: Mo, C; Dose: 30000e⁻/Å²

```
"model": "gpt-4.1-mini
"energy": -8.25,
"energy_per_atom": -8.25
"error_type": null,
```

```
"model": "llama-4-maverick-17b-128e-
instruct",
"energy": -1.721,
"energy_per_atom": -0.123,
"error_type": null,
```

Initializing the structure analysis process... The following steps will be executed in order:
1. Image preprocessing 2. Template matching
3. Atomic localization and lattice reconstruction
4. Property calculation

```
"energy": -31.297068,
"energy_per_atom": -7.824266911
```

Tier3 Example
- **Elements:** Ti, Co, O
- **Description:** Multimetallic, stable, oxide compound
- **Application:** Catalysis, energy storagecoating

Ground Truth:
- Total Energy: −74.836 eV (Unit-cell)
- Energy per Atom: −6.236 eV
- Structure: Space group P-3m1

System Prompt:
This is a 2D material STEM image, where each pixel corresponds to 0.1 Å......Provide a single-point estimate and return:
{"energy": <total energy (eV/atom): >,
"energy_per_atom": <energy per atom (eV/atom):>}

Elements: Ti, Co, O; Dose: 10000e⁻/Å²

```
"model": "gpt-4.1-mini",
"energy": -7.85,
"energy_per_atom": -7.85,
"error_type": null
```

```
"model": "llama-4-maverick-17b-128e-
instruct
"energy": -4.375,
"energy_per_atom": -1.094,
"error_type": null,
```

Initializing the structure analysis process... The following steps will be executed in order:
1. Image preprocessing 2. Template matching
3. Atomic localization and lattice reconstruction
4. Property calculation

```
"energy": -75.49687,
"energy_per_atom": -6.291405996
```

Figure 4: Case studies on Tier 2 and Tier 3 samples comparing AutoMat, GPT-4.1mini and Llama-4-Maverick: enhanced images, best-matched templates, reconstructed unit cells and predicted energies.

For structural reconstruction, AutoMat achieves an average projected RMSD$_{xy}$ of about 0.11 Å, much lower than MicroscopyGPT (1.3–1.6 Å, requiring prior coordination knowledge) and AtomAI (43–44 Å). Most deviations can be corrected through final relaxation. In composition correctness, AutoMat averages 83% across tiers (88.9%, 85.9%, 73.1%), while MicroscopyGPT performs well only in Tier 1 ($\approx$92%) but drops sharply in Tier 2–3. AtomAI, by contrast, stays below 2.7%. For structure success rate, AutoMat achieves 83.2% overall (85.0%, 84.0%, 73.1%), compared with only 25.9% for MicroscopyGPT in Tier 2 and almost zero for AtomAI.

In summary, **AutoMat** not only outperforms all baselines but also maintains strong results in challenging Tier 3 cases with multi-element compositions and low imaging doses, demonstrating robustness and generalizability across the full benchmark.

5.2 Real iDPC-STEM Case Study and Human-Expert Effort

Real iDPC-STEM case. To complement the simulation-based benchmark, we also evaluate **AutoMat** on a real iDPC-STEM case: a ZSM-5 zeolite sample acquired on the same Cs-corrected STEM under dose-limited conditions. Without any manual tuning, the agent automatically performs denoising and structural reconstruction, and outputs a CIF structure whose projected lattice and channel framework qualitatively agree with the known crystallographic model. A concrete example is provided in Appendix A.10. This demonstrates that AutoMat has a certain ability to transfer the image–structure–property pipeline from abTEM simulations to real microscope images.

Human-expert effort. To provide an intuitive comparison, we consulted a senior electron-microscopy expert in our group. Based on their experience with Tier-2–level STEM images, manually interpreting a single large-field image (1024×1024 pixels), confirming lattice parameters and elemental species, and producing a simulation-ready CIF of comparable complexity typically requires about 6–8 hours per sample. In contrast, on our benchmark hardware, **AutoMat** processes similar Tier-2 test samples in roughly 2 minutes per case. Even when targeting expert-level structural quality, this corresponds to an orders-of-magnitude improvement in throughput.

5.3 Error Analysis

To better understand the failure modes of **AutoMat**, we analyzed representative failure cases across all three tiers and identified two primary types of errors (a detailed analysis with examples is provided in Appendix A.9):

(1) Template retrieval failure (39.3%): In these cases, AutoMat failed to retrieve the correct structure from the template database, resulting in mismatches in atomic arrangements and element types. This led to cascading errors in structure, composition, and property predictions. Incorrect atom counts further caused large energy estimation errors.

(2) Downstream failure despite correct template (60.7%): Even with the correct template, downstream steps failed due to projection ambiguity or elemental confusion. In 40% of these cases, atoms appeared too close in the 2D projection, and the lack of z-axis information led to poor relaxation and inaccurate energy estimates. In 20.7%, elements with similar atomic numbers (e.g., C and O) exhibited indistinguishable contrast, causing misclassification and full breakdowns in lattice fitting and CIF generation.

These findings highlight two key directions for improving **AutoMat**: (i) improving retrieval robustness via uncertainty-aware; and (ii) overcoming 2D projection limits through 3D-aware modeling Guo et al. (2023); Tang et al. (2024; 2025b) and enhanced modality integration. Together, these efforts can greatly improve structural fidelity and prediction reliability in complex systems.

6 Conclusion

We proposed **AutoMat**, an end-to-end agent system that automatically reconstructs material structures and predicts properties from STEM images. By integrating pattern-adaptive vision models, symmetry-constrained reconstruction, and LLM-driven orchestration, AutoMat achieves accurate alignment between microscopy data and atomic modeling, significantly outperforming existing methods in structural and energetic evaluation. Meanwhile, we propose STEM2Mat-Bench for this task. AutoMat advances autonomous materials research and AI-driven experimentation. Future work will focus on strengthening its role as a bridge between experimental characterization and theoretical computation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2023. URL `https://arxiv.org/abs/2206.07697`.

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, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, 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, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian

Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024. URL https://arxiv.org/abs/2401.00096.

Ishraque Zaman Borshon, Marco Ragone, Abhijit H Phakatkar, Lance Long, Reza Shahbazian-Yassar, Farzad Mashayek, and Vitaliy Yurkiv. Predicting column heights and elemental composition in experimental transmission electron microscopy images of high-entropy oxides using deep learning. *npj Computational Materials*, 10(1):275, 2024.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools, 2023. URL https://arxiv.org/abs/2304.05376.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Kamal Choudhary. Microscopygpt: Generating atomic-structure captions from microscopy images of 2d materials with vision-language transformers. *The Journal of Physical Chemistry Letters*, 16(27): 7028–7035, 2025. doi: 10.1021/acs.jpclett.5c01257. URL https://doi.org/10.1021/acs.jpclett.5c01257. PMID: 40590052.

Annick De Backer, Sandra Van Aert, Christel Faes, Ece Arslan Irmak, Peter D Nellist, and Lewys Jones. Experimental reconstructions of 3d atomic structures from electron microscopy images using a bayesian genetic algorithm. *npj Computational Materials*, 8(1):216, 2022.

Henrik Eliasson and Rolf Erni. Localization and segmentation of atomic columns in supported nanoparticles for fast scanning transmission electron microscopy. *npj Computational Materials*, 10(1):168, 2024.

Laura Gambini, Tiarnan Mullarkey, Lewys Jones, and Stefano Sanvito. Machine-learning approach for quantified resolvability enhancement of low-dose stem data. *Machine Learning: Science and Technology*, 4(1), 2023. ISSN 2632-2153. doi: 10.1088/2632-2153/acbb52.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6790–6802. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/35cf8659cfcb13224cbd47863a34fc58-Paper.pdf.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15372–15383, 2023.

Sten Haastrup, Mikkel Strange, Mohnish Pandey, Thorsten Deilmann, Per S Schmidt, Nicki F Hinsche, Morten N Gjerding, Daniele Torelli, Peter M Larsen, Anders C Riis-Jensen, et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials*, 5(4):042002, 2018.

Wenqiang Huang, Yucheng Jin, Zhemin Li, Lin Yao, Yun Chen, Zheng Luo, Shen Zhou, Jinguo Lin, Feng Liu, Zhifeng Gao, et al. Auto-resolving the atomic structure at van der waals interfaces using a generative model. *Nature Communications*, 16(1):2927, 2025.

Shiro Ihara, Hikaru Saito, Mizumo Yoshinaga, Lavakumar Avala, and Mitsuhiro Murayama. Deep learning-based noise filtering toward millisecond order imaging by using scanning transmission electron microscopy. *Scientific reports*, 12(1):13462, 2022. ISSN 2045-2322.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1), 2013.

Sergei V Kalinin, Colin Ophus, Paul M Voyles, Rolf Erni, Demie Kepaptsoglou, Vincenzo Grillo, Andrew R Lupini, Mark P Oxley, Eric Schwenker, and Maria KY Chan. Machine learning in scanning transmission electron microscopy. *Nature Reviews Methods Primers*, 2(1):11, 2022. ISSN 2662-8449.

Sergei V Kalinin, Debangshu Mukherjee, Kevin Roccapriore, Benjamin J Blaiszik, Ayana Ghosh, Maxim A Ziatdinov, Anees Al-Najjar, Christina Doty, Sarah Akers, Nageswara S Rao, et al. Machine learning for automated experimentation in scanning transmission electron microscopy. *npj Computational Materials*, 9(1):227, 2023.

J. Li, M. Telychko, J. Yin, Y. Zhu, G. Li, S. Song, H. Yang, J. Li, J. Wu, J. Lu, and X. Wang. Machine vision automated chiral molecule detection and classification in molecular imaging. *J Am Chem Soc*, 143(27):10177–10188, 2021. ISSN 1520-5126 (Electronic) 0002-7863 (Linking). doi: 10.1021/jacs.1c03091. URL `https://www.ncbi.nlm.nih.gov/pubmed/34227379`. Li, Jiali Telychko, Mykola Yin, Jun Zhu, Yixin Li, Guangwu Song, Shaotang Yang, Haitao Li, Jing Wu, Jishan Lu, Jiong Wang, Xiaonan eng Research Support, Non-U.S. Gov't 2021/07/07 J Am Chem Soc. 2021 Jul 14;143(27):10177-10188. doi: 10.1021/jacs.1c03091. Epub 2021 Jul 6.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 415–423, 2025.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, 2024. URL `https://arxiv.org/abs/2306.12059`.

Ruoqian Lin, Rui Zhang, Chunyang Wang, Xiao-Qing Yang, and Huolin L Xin. Temimagenet and atomsegnet deep learning training library and models for high-precision atom segmentation, localization, denoising, and super-resolution processing of atom-resolution scanning tem images. *arXiv e-prints*, pp. arXiv: 2012.09093, 2020.

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. ISSN 2045-2322.

Shuqi Lu, Zhifeng Gao, Di He, Linfeng Zhang, and Guolin Ke. Highly accurate quantum chemical property prediction with uni-mol+, 2023. URL `https://arxiv.org/abs/2303.16982`.

Jacob Madsen and Toma Susi. abtem: Ab initio transmission electron microscopy image simulation. *Microscopy and Microanalysis*, 26(S2):448–450, 2020.

Artem Maksov, Ondrej Dyck, Kai Wang, Kai Xiao, David B Geohegan, Bobby G Sumpter, Rama K Vasudevan, Stephen Jesse, Sergei V Kalinin, and Maxim Ziatdinov. Deep learning analysis of defect and phase evolution during electron beam-induced transformations in ws2. *npj Computational Materials*, 5(1):12, 2019. ISSN 2057-3960.

Volkan Ortalan, Alper Uzun, Bruce C Gates, and Nigel D Browning. Direct imaging of single metal atoms and clusters in the pores of dealuminated hy zeolite. *Nature nanotechnology*, 5(7):506–510, 2010. ISSN 1748-3387.

Tom Stoops, Annick De Backer, Ivan Lobato, and Sandra Van Aert. Obtaining 3d atomic reconstructions from electron microscopy images using a bayesian genetic algorithm: Possibilities, insights, and limitations. *Microscopy and Microanalysis*, 31(1):ozae090, 2025.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. Chemagent: Self-updating library in large language models improves chemical reasoning, 2025a. URL `https://arxiv.org/abs/2501.06590`.

Yiwen Tang, Ray Zhang, Jiaming Liu, Zoey Guo, Bin Zhao, Zhigang Wang, Peng Gao, Hongsheng Li, Dong Wang, and Xuelong Li. Any2point: Empowering any-modality large models for efficient 3d understanding. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2024.

Yiwen Tang, Zoey Guo, Zhuhao Wang, Ray Zhang, Qizhi Chen, Junli Liu, Delin Qu, Zhigang Wang, Dong Wang, Xuelong Li, et al. Exploring the potential of encoder-free architectures in 3d lmms. *arXiv preprint arXiv:2502.09620*, 2025b.

Antanas Vaitkus, Andrius Merkys, Thomas Sander, Miguel Quirós, Paul A. Thiessen, Evan E. Bolton, and Saulius Gražulis. A workflow for deriving chemical entities from crystallographic data and its application to the Crystallography Open Database. *Journal of Cheminformatics*, 15(1), Dec 2023. doi: 10.1186/s13321-023-00780-2. URL https://doi.org/10.1186/s13321-023-00780-2.

Feng Wang, Trond R Henninen, Debora Keller, and Rolf Erni. Noise2atom: unsupervised denoising for scanning transmission electron microscopy images. *Applied Microscopy*, 50(1):23, 2020. ISSN 2287-4445.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures, 2024. URL https://arxiv.org/abs/2405.04967.

Yaotian Yang, Hao Xiong, Zirong Wu, Zhiyao Luo, Xiao Chen, Xiaonan Wang, and Fei Wei. Deep learning-enabled stem imaging for precise single-molecule identification in zeolite structures. *Advanced Science*, 12(6):2408629, 2025. doi: https://doi.org/10.1002/advs.202408629. URL https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/advs.202408629.

Lance Yao, Suman Samantray, Ayana Ghosh, Kevin Roccapriore, Libor Kovarik, Sarah Allec, and Maxim Ziatdinov. Operationalizing serendipity: Multi-agent ai workflows for enhanced materials characterization with theory-in-the-loop, 2025. URL https://arxiv.org/abs/2508.06569.

Daliang Zhang, Yihan Zhu, Lingmei Liu, Xiangrong Ying, Chia-En Hsiung, Rachid Sougrat, Kun Li, and Yu Han. Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials. *Science*, 359(6376):675–679, 2018. ISSN 0036-8075.

Hui Zhang, Guanxing Li, Jiaxing Zhang, Daliang Zhang, Zhen Chen, Xiaona Liu, Peng Guo, Yihan Zhu, Cailing Chen, and Lingmei Liu. Three-dimensional inhomogeneity of zeolite structure and composition revealed by electron ptychography. *Science*, 380(6645):633–638, 2023. ISSN 0036-8075.

Qing Zhang, Alvaro Mayoral, Junyan Li, Juanfang Ruan, Viveka Alfredsson, Yanhang Ma, Jihong Yu, and Osamu Terasaki. Electron microscopy studies of local structural modulations in zeolite crystals. *Angewandte Chemie*, 132(44):19571–19581, 2020. ISSN 0044-8249.

M. Ziatdinov, O. Dyck, A. Maksov, X. Li, X. Sang, K. Xiao, R. R. Unocic, R. Vasudevan, S. Jesse, and S. V. Kalinin. Deep learning of atomically resolved scanning transmission electron microscopy images: Chemical identification and tracking local transformations. *ACS Nano*, 11(12):12742–12752, 2017. ISSN 1936-086X (Electronic) 1936-0851 (Linking). doi: 10.1021/acsnano.7b07504. URL https://www.ncbi.nlm.nih.gov/pubmed/29215876. Ziatdinov, Maxim Dyck, Ondrej Maksov, Artem Li, Xufan Sang, Xiahan Xiao, Kai Unocic, Raymond R Vasudevan, Rama Jesse, Stephen Kalinin, Sergei V eng Research Support, Non-U.S. Gov't 2017/12/08 ACS Nano. 2017 Dec 26;11(12):12742-12752. doi: 10.1021/acsnano.7b07504. Epub 2017 Dec 14.

Maxim Ziatdinov, Ayana Ghosh, Chun Yin Wong, and Sergei V Kalinin. Atomai framework for deep learning analysis of image and spectroscopy data in electron and scanning probe microscopy. *Nature Machine Intelligence*, 4(12):1101–1112, 2022.

A Appendix

A.1 Dataset Overview

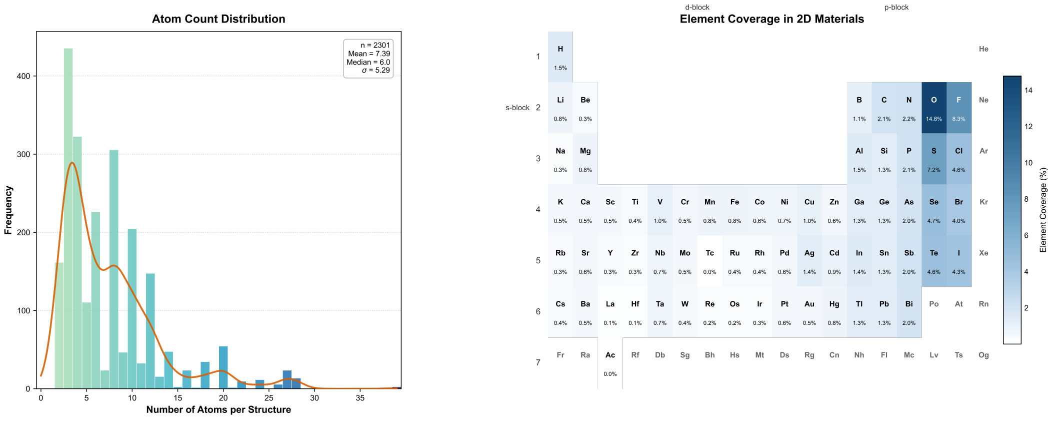


Figure 5: Distribution of per-unit-cell structure atom counts (left) and elemental coverage percentages in our curated 2D materials dataset (right).

A.2 MoE–DIVAESR: Training Pipeline & Architecture

A. Training pipeline

1. **Dataset partitioning.** A total of 1,693 denoised STEM images were projected to a PCA-reduced feature space and clustered using k-means (k=8). The resulting clusters were grouped into **2 parent** and **6 child** routing categories, defining the hierarchical MoE.
2. **Prototype selection & simulation.** For each child cluster, the top 25% most representative samples (424 total) were re-simulated into iDPC-STEM projections. These images were augmented with random rotation, flipping, Poisson noise, Gaussian blur, elastic deformation, and sliding-window cropping, resulting in a total of **756,042** training patches of size 128×128.
3. **Stage 1: Router pretraining.** A ResNet-18 model was trained as a routing network for 30 epochs to predict the appropriate expert index. It achieved a top-1 routing accuracy of **99%** on the validation set. The weights were then frozen for expert training.
4. **Stage 2: Expert training.** Six DIVAESR experts (EDSR backbone with 12 residual blocks and 64 feature maps) were trained jointly under fixed routing for 300 epochs. Compared to a single DIVAESR baseline, the MoE model reduced reconstruction error by approximately 50%.

B. Model Architecture and Objective

Each expert E_k is modeled as a two-stage cascade of a denoising module and a super-resolution module:

$$E_k = g^{\text{SR}}_{\phi_k} \circ f^{\text{DIVAE}}_{\theta_k}, \qquad k = 1, \dots, K \tag{5}$$

where $f^{\text{DIVAE}}_{\theta_k}$ is the denoising network and $g^{\text{SR}}_{\phi_k}$ is the super-resolution network.

The router R_ψ produces a one-hot routing mask $\mathbf{r} \in \{0,1\}^K$ that selects expert k^* for each input:

$$k^* = R_\psi(x^{\text{noisy}}) \tag{6}$$

The output of the MoE model is then computed as:

$$\hat{x}^{\text{HR}} = \sum_{k=1}^{K} r_k \cdot g^{\text{SR}}_{\phi_k}\left(f^{\text{DIVAE}}_{\theta_k}(x^{\text{noisy}})\right) \tag{7}$$

Joint loss objective. The total training loss for a mini-batch is given by:

$$\mathcal{L}_{\text{MoE}} = \sum_{k=1}^{K} r_k \cdot \left(\mathcal{L}_{\text{DIVAE}}^{(k)} + \mathcal{L}_{\text{SR}}^{(k)}\right) + \lambda \sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k \tag{8}$$

where:

$$\mathcal{L}_{\text{DIVAE}}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left\|\hat{x}_i^{\text{clean}} - x_i^{\text{clean}}\right\|_2^2 + \beta \cdot \text{KL}\left[q(z|x_i^{\text{noisy}}) \parallel \mathcal{N}(0, I)\right] \tag{9}$$

is the β-VAE denoising loss, and

$$\mathcal{L}_{\text{SR}}^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left\|\hat{x}_i^{\text{HR}} - x_i^{\text{HR}}\right\|_1 \tag{10}$$

is the L1 reconstruction loss for super-resolution. Here, $\bar{p}_k$ denotes the empirical routing frequency for expert k, and the KL divergence term regularizes the latent encoding.

This formulation jointly minimizes noise and detail loss through specialized expert modules. The use of KL regularization in Eq. equation 9 ensures stability in latent representation, while Eq. equation 10 preserves fine image details. The entropy term in Eq. equation 8 promotes balanced expert usage. Altogether, the MOE-DIVAESR system delivers higher-fidelity reconstruction and improved downstream performance (e.g., energy MAE) compared to single-model baselines.

A.3 Template Matching and STEM2CIF Conversion

A. Image Template Matching

1. **Feature extraction.** Atomic centroids $\{\mathbf{p}_i\}_{i=1}^N$ are detected in each denoised image. An m=32-bin radial-distribution histogram $\mathbf{g} = (g_1, \ldots, g_m)$ is computed:

$$g_b = \frac{1}{N} \sum_{i<j} \mathbb{I}[r_{b-1} \leq \|\mathbf{p}_i - \mathbf{p}_j\| < r_b], \quad b = 1, \ldots, m, \tag{11}$$

   where $\{r_b\}$ are uniformly spaced radial bin edges, and $\mathbb{I}[\cdot]$ is the indicator function.

2. **Candidate retrieval.** All template images are similarly encoded as RDF descriptors $\{\mathbf{g}^{(t)}\}$. The cosine distance is computed as:

$$d_{\cos}\left(\mathbf{g}, \mathbf{g}^{(t)}\right) = 1 - \frac{\langle \mathbf{g}, \mathbf{g}^{(t)} \rangle}{\|\mathbf{g}\| \cdot \|\mathbf{g}^{(t)}\|}, \tag{12}$$

   and the top-k nearest templates ($k = 10$ by default) are retrieved.

3. **Element filtering.** Retain only templates whose elemental composition $\mathcal{E}^{(t)}$ matches that of the input:

$$\mathcal{E}^{(t)} = \mathcal{E}^{(\text{input})}. \tag{13}$$

   From the remaining candidates, select the one with the smallest $d_{\cos}$.

4. **Output.** The top-1 matched template is copied to the output directory for use in downstream structure reconstruction.

Equation equation 11 encodes pairwise geometric statistics that are invariant to translation and rotation, facilitating global matching. The cosine similarity in Eq. equation 12 offers scale-invariant structural comparison. Combined with elemental filtering, this method reduces false matches in complex multi-element systems and provides high-quality priors for reconstruction.

B. STEM2CIF Conversion

1. **Peak localization.** Atomic column centers were localized using weighted mean-shift clustering followed by sub-pixel 2D Gaussian fitting.

2. **Lattice fitting.** An initial lattice was fitted using least-squares optimization based on known pixel resolution (0.1 Å/pixel), constrained by the space group symmetry of the matched template.
3. **Element assignment.** Atomic column intensities were compared to simulated iDPC-STEM contrast values, and the most probable element was assigned to each atomic site.
4. **CIF generation.** The reconstructed periodic structure was reduced to its primitive unit cell and exported as a standard CIF file for downstream MatterSim relaxation and property prediction.

A.4 Ablation Study

The `STEM2CIF` module is the only component in our system that converts microscopy images into atomic structures in the form of `CIF` files. Removing this module would prevent the generation of any structural outputs, which in turn makes key evaluation metrics—such as RMSD, composition correctness, or formation-energy MAE—undefined. Therefore, an ablation study that omits `STEM2CIF` is technically infeasible and would not yield meaningful results.

To isolate the specific contribution of `STEM2CIF`, we conducted an "Upper Bound" experiment (see Table 2 in the main text), where ground-truth structures were directly fed into the module, bypassing earlier steps. This experiment quantifies the module's best-case reconstruction accuracy and demonstrates its essential role in the pipeline.

To evaluate the contributions of individual modules in the AutoMat pipeline, we conduct an ablation study by selectively disabling (i) pattern-adaptive denoising (MOE-DIVAESR), and (ii) physics-guided template matching, while keeping the rest of the pipeline unchanged. We report the mean absolute error (MAE) of formation energy per atom across all three evaluation tiers(Tables 3).

Table 3: Energy per Atom MAE (meV/atom) under different ablation settings. Lower is better.

Method	**Tier 1**	**Tier 2**	**Tier 3**
No Denoising (w/o MOE-DIVAESR)	6584	2616	938
No Template Matching	617	608	672
Full Pipeline (AutoMat)	**344**	**320**	**333**

The results show that removing MOE-DIVAESR significantly increases the error in Tier 1, indicating its critical role in enhancing low-noise images. In contrast, removing template matching severely affects Tier 3, where accurate prior guidance is essential for complex, low-dose structures. AutoMat achieves the best results across all tiers, demonstrating that both components are complementary and necessary.

A.5 Broader Impact

AutoMat offers the potential to significantly accelerate the discovery and validation of novel materials by automating the labor-intensive process of structure reconstruction from electron microscopy images. This can reduce the reliance on expert annotations and lower the barrier to entry for scientific exploration, particularly in under-resourced research settings. The open-access nature of our dataset and code also promotes transparency and reproducibility in materials science, fostering collaboration and inclusive participation.

A.6 Experimental Settings

Framework. Model training and experiment management were implemented using the `PyTorch Lightning` framework on a 4x NVIDIA H100 NVLink server (80GB per GPU). Hyperparameter tuning for MOE-DIVAESR required approximately 4–5 days(See Tables 4 for more configuration details.).

Table 4: Configuration of MOE-DIVAESR and related submodules

Gating Network	
Parent Router Modules	2
Child Router Modules	6
Latent Prior Module (DIVAESR)	
Input Channels	1
Latent Dimension	128
Super-Resolution Module (DIVAESR)	
Residual Blocks	12
Feature Maps per Block	64
Activation Function	ReLU
Patch Size	128
Scale Factor	×2
Image Channels	1 (grayscale)
Precision	FP32
Optimization and Training Settings	
Optimizer	Adam
Initial Learning Rate	0.001
Learning Rate Scheduler	Exponential decay (gamma = 0.95)
Weight Decay	0.0
KL Divergence Weight	0.00025
Manual Seed	22
Training Epochs	300
Training Batch Size	128
Validation Batch Size	64
Number of Workers	8
Active GPU Device	GPU ID 3

A.7 MicroscopyGPT Baseline

In our experiments, MicroscopyGPT is used as an off-the-shelf vision–language baseline without any additional fine-tuning on top of the released checkpoint. We rely on the model introduced by Choudhary *et al.*, which fine-tunes an 11B LLaMA-3.2–Vision model with QLoRA on a chemically and structurally diverse collection of simulated STEM images generated with AtomVision from the JARVIS-DFT-2D and C2DB two-dimensional materials databases. The training set comprises on the order of 5 000 2D materials; each example pairs a 2D STEM projection (mainly along the (001) direction) with a structured textual description containing lattice parameters, lattice angles, element types, and fractional atomic coordinates, and the model is evaluated on a 10% held-out test split using distributional metrics (KLD, EMD) and coordinate RMSE for atomic positions.

Because the MicroscopyGPT training data already cover a large class of simulated 2D monolayers similar to those in our benchmark, we deliberately do *not* further fine-tune the model on STEM2Mat-Bench. Additional fine-tuning on overlapping data could lead to overfitting or catastrophic forgetting and would make the baseline less comparable across studies. Instead, we treat MicroscopyGPT as a frozen black-box predictor and evaluate it under the same metrics as AutoMat. Concretely, for each benchmark image we (i) feed the STEM image to MicroscopyGPT with a standard structure-generation prompt, (ii) parse the textual output (lattice constants, angles, species, fractional coordinates, and Miller index) into a structured representation, and (iii) convert this representation into a CIF file using ASE and related crystallographic utilities. The resulting CIFs are then fed into the same evaluation pipeline as AutoMat to compute RMSD, C.C., S.S., and energy-related errors. The complete inference and text-to-CIF conversion scripts will be released in our public code repository.

A.8 Robustness Across LLM/VLM Agents and Retry Mechanism

Retry Mechanism. We implement **in-context learning from environmental feedback**, where error messages and traceback logs are automatically fed into subsequent conversation turns. This enables the agent to refine function calls until either the maximum retry limit is reached or the tool call succeeds. The following example illustrates a failed attempt to call the `stem2cif_tool` due to parameter format errors, followed by an automatic correction and successful execution:

```
[TOOL_CALL] stem2cif_tool
{
  "img_path": ...,
  "elements": ["C", "Y"]
}
[TOOL_RESPONSE] stem2cif_tool
An error occurred: ValueError: Unexpected
[ANSWER]
Tool call failed due to a parameter format error.
Correct the parameters and Retry.
[TOOL_CALL] stem2cif_tool
{"img_path": ..., "elements": ["C", "Y"]}
[TOOL_RESPONSE] stem2cif_tool
{success: true, cif_path: ...}
```

This retry mechanism demonstrates the system's robustness: failures are not final but provide useful signals to guide subsequent decisions.

Overall Sensitivity to the Choice of LLM/VLM Agent. To address concerns that AutoMat's performance may depend on a specific backbone model (e.g., DeepSeek V3), we conducted controlled experiments replacing the LLM/VLM with Qwen3, GPT-4o, and Qwen-VL (72B)(refer to Tables 5 and Tables 6). Using a stratified random sample of 230 test cases, we observed that performance remained stable across all variants, with the average formation-energy MAE within $\pm 3\%$. This result confirms that the innovation lies in the framework design and agentic tool-use paradigm rather than reliance on a particular model.

Table 5: Formation-Energy MAE (meV/atom) Across Tiers for Different LLM and VLM Agents

Method	**Tier 1 ↓**	**Tier 2 ↓**	**Tier 3 ↓**	**Avg. ↓**
Qwen-VL-agent (72B)	261.6	308.5	446.1	313.84
Qwen3-agent (2025-04-28)	245.2	306.2	435.1	310.54
GPT-4O-agent (2025-03-27)	240.7	304.2	435.0	308.54
DeepSeekV3-agent	243.8	305.3	433.5	309.58

Decision-Making Flexibility and Tool Invocation Statistics. AutoMat functions as an agentic, end-to-end controller rather than a fixed pipeline. It dynamically decomposes user instructions into subtasks (denoising, template matching, structure reconstruction, property prediction) and adapts tool invocation depending on intermediate outcomes. If a tool fails (e.g., invalid reconstruction), the system triggers rollback-and-retry until success or termination. Moreover, the agent can skip, re-order, or re-rank templates, enabling robust performance on ill-posed problems.

To quantify this behavior, Table 7 reports detailed tool usage and retry statistics across different agents.

These results show that although success and retry rates vary across models, the overall pipeline remains stable and reliable, highlighting AutoMat's robustness and generalizability.

Table 6: Structural Accuracy Metrics for Different LLM and VLM Agents Across Tiers. RMSDxy↓: in-plane lattice RMSD; C.C.↑: composition correctness; S.S.↑: structure success rate.

Tier	Method	RMSDxy (Å) ↓	C.C. (%) ↑	S.S. (%) ↑
1	Qwen-VL (72B)	0.12 ± 0.02	100	100
	Qwen3-agent	0.12 ± 0.01	100	100
	GPT-4O-agent	0.12 ± 0.01	100	100
	DeepSeekV3-agent	0.11 ± 0.02	100	100
2	Qwen-VL (72B)	0.11 ± 0.03	83.1	79.2
	Qwen3-agent	0.11 ± 0.02	83.3	79.4
	GPT-4O-agent	0.11 ± 0.02	84.2	80.1
	DeepSeekV3-agent	0.11 ± 0.02	85.9	79.8
3	Qwen-VL (72B)	0.12 ± 0.02	58.3	58.3
	Qwen3-agent	0.11 ± 0.02	66.7	66.7
	GPT-4O-agent	0.11 ± 0.02	66.7	66.7
	DeepSeekV3-agent	0.11 ± 0.03	66.7	66.7

Table 7: Tool Invocation, Success, and Retry Statistics of Different LLM/VLM Agents in the AutoMat Toolchain.

LLM Agent	Tool	Total Calls	Callbacks	Successes	Retries	Success Rate (%)
Qwen-VL (72B)	Denoising	293,417	259,217	259,217	34,200	88.34
	Property Prediction	165,594	121,153	68,532	97,062	41.39
	STEM2CIF	314,001	252,801	133,282	180,719	42.45
	Template Matching	258,987	205,762	205,762	53,225	79.45
Qwen3 (2025-04-28)	Denoising	64,929	58,142	58,142	6,787	89.55
	Property Prediction	40,040	32,563	32,563	7,477	81.33
	STEM2CIF	47,806	40,270	40,270	7,536	84.24
	Template Matching	57,912	48,036	48,036	9,876	82.95
GPT-4O (2025-03-27)	Denoising	58,053	56,950	56,950	1,103	98.10
	Property Prediction	38,705	36,563	36,563	2,142	94.46
	STEM2CIF	46,600	43,507	43,507	3,093	93.36
	Template Matching	59,506	55,951	55,951	3,555	94.02

A.9 Failure Analysis

Failure Analysis. We provide a detailed breakdown of downstream failure cases to clarify the system's limitations. Failure analysis was conducted with rollback statistics using the Qwen3 agent, with similar outcomes observed across other LLM/VLM backbones. Failures can be grouped into two major categories:

- **Template-retrieval failures (39.3%).** These arise when the correct structure is absent from the template database or visually indistinguishable within it. *Representative case:* Tier 1, `2dm-5936` (Se-oxide vs. nitride). High-Z Se columns resemble N, while light elements are suppressed in Z-contrast, leading to the selection of a P-based template and subsequent errors.
- **Downstream failures despite correct template (60.7%).** In these cases, the correct template is retrieved, but downstream reconstruction fails due to projection effects or extreme contrast imbalance. Two common sub-modes are summarized in Table 8.

In summary, this analysis highlights two important aspects. First, the indispensability of `STEM2CIF`, without which structural reconstruction and quantitative evaluation are impossible. Second, failure

Table 8: Breakdown of downstream failure sub-modes despite correct template retrieval.

Sub-mode	Share	Core Problem	Representative Case
Projection overlap	40%	Atoms in different z-layers project to the same (x, y), confusing structural relaxation	Tier 2, `2dm-1014`: B/I overlap; I dominates IDPC, B is lost
Extreme Z-contrast	20.7%	Heavy elements dominate contrast and mask lighter atoms, treated as noise	Tier 3, `2dm-5199`: U contrast hides O/F, leading to $>$ 3 eV energy error

modes are not random but arise from well-defined imaging and physical conditions, such as projection overlap and extreme Z-contrast. These insights provide clear directions for future work on improving robustness and addressing systematic limitations in the pipeline.

A.10 Real STEM Image Reconstruction on ZSM-5

Due to experimental limitations, our current setup does not yet support systematic preparation and iDPC-STEM imaging of 2D materials, and it is difficult to obtain high-quality 2D samples (with both suitable morphology and stability) within a short time frame. To nevertheless provide an honest and informative evaluation of *AutoMat* on real experimental data under our present conditions, we chose ZSM-5 zeolite — a system that our group has extensively studied and can reliably synthesize — as a representative test case. Using ZSM-5 allows us to assess how well *AutoMat* generalizes from simulated iDPC-STEM images to real microscopy images in a realistic laboratory scenario.

ZSM-5 iDPC-STEM imaging experiments. The iDPC-STEM experiments were carried out on a Cs-corrected scanning transmission electron microscope (FEI Titan Cubed Themis G2 300) equipped with a four-quadrant detector and operated at an accelerating voltage of 300 kV. The convergence semi-angle was 15 mrad, the beam current was kept below 0.5 pA, the collection angle ranged from 4 to 22 mrad, and the probe dwell time was 32–64 μs.

Sheet-like ZSM-5 single crystals were ultrasonically dispersed in pure pyridine or thiophene for 1–2 h, then deposited onto microgrids and heated at 130 °C for 2 h to remove physically adsorbed molecules, leaving only those strongly interacting with the acid sites. In view of the intrinsic stochastic nature of single-molecule adsorption and dynamics, the present work mainly focuses, in the quantitative analysis, on assessing the robustness of the framework reconstruction pipeline. To this end, we selected image regions with little or no visibly adsorbed small molecules, cropped these relatively "clean" framework areas, and used them to evaluate the reconstruction performance of the ZSM-5 zeolite framework.

AutoMat reconstruction on real ZSM-5 images. We then fed these relatively clean ZSM-5 framework regions directly into *AutoMat* for end-to-end reconstruction. The results (Fig. 6) show that *AutoMat* can effectively recover the ZSM-5 surface framework from real iDPC-STEM images. On the one hand, the reconstructed CIFs exhibit continuous and well-defined ten-membered-ring channels and the expected framework topology, in close qualitative agreement with the canonical DFT ZSM-5 structure. On the other hand, after aligning the reconstructed structure with the DFT reference, the deviation of framework atomic positions is within approximately 0.2 Å, indicating that *AutoMat* maintains good structural accuracy even under realistic experimental noise and contrast conditions.

Although this test system differs from our main task (2D monolayers) in both material class and imaging conditions, the ZSM-5 experiment demonstrates that *AutoMat* can already exhibit a certain degree of generalization to real experimental data without any task-specific fine-tuning, providing preliminary evidence for its potential deployment on a broader range of real materials systems.

A.11 Limitations and Future Directions

While AutoMat demonstrates the feasibility of an agentic pipeline for microscopy-to-structure reconstruction, our present scope is intentionally restricted to **2D monolayer systems imaged with**

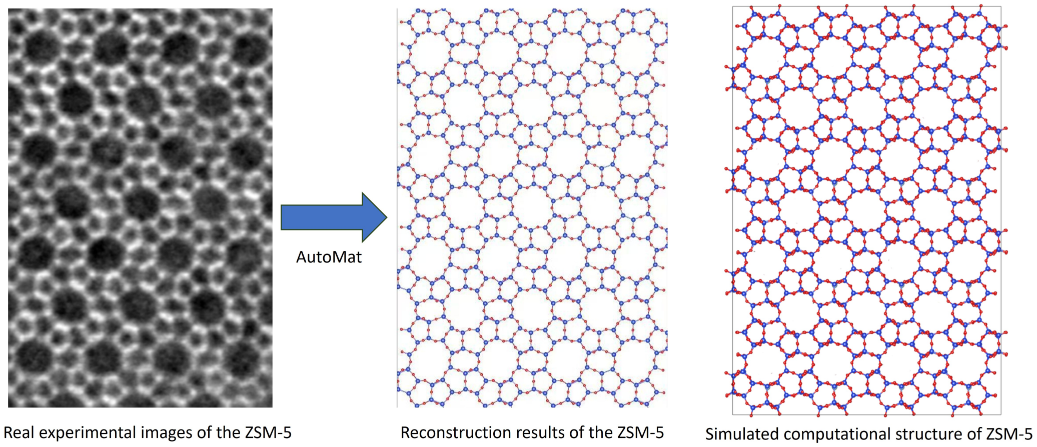

AutoMat
Real experimental images of the ZSM-5
Reconstruction results of the ZSM-5
Simulated computational structure of ZSM-5

Figure 6: **AutoMat reconstruction on real iDPC-STEM images of ZSM-5 zeolite.** (Left) Real experimental iDPC-STEM image of the ZSM-5 framework, showing the characteristic ten-membered-ring pore arrangement. (Middle) Atomic structure reconstructed by *AutoMat* directly from the real STEM image, where the surface framework and pore topology are largely recovered. (Right) DFT-relaxed reference model of ZSM-5.

iDPC-STEM. This design yields a stable and physically interpretable testbed, but also introduces several limitations:

- **2D scope and single-projection depth ambiguity.** The current pipeline assumes a single 2D iDPC-STEM projection of a monolayer, exploits the approximately linear relationship between iDPC contrast and atomic number (Z), and leverages this for a deterministic, physics-guided reconstruction. This setting is well suited to 2D materials, but inherently discards z-depth information and therefore cannot yet be directly applied to 3D bulk crystals or multilayer stacks.
- **Element classification uncertainty.** Even in 2D, element assignment still relies on contrast cues and template priors. When elements have similar Z, or when contrast is degraded by low dose or aberrations, different species may become indistinguishable in STEM images. Such misclassification is a major source of reconstruction error and can propagate to downstream property prediction failures.
- **Template retrieval sensitivity to acquisition conditions.** In the current release, the retrieval library is generated at a fixed pixel scale and spans only a limited range of imaging conditions; template matching is therefore sensitive to mismatches in pixel scale, defocus, and aberrations, which reduces robustness when acquisition conditions deviate from those assumed in the simulations.
- **Heuristic agent policy instead of a learnable one.** At present, the agent uses state-dependent rules and prompt scheduling to select tools and refine parameters. Given the small, semantically stable tool pool and the fact that evaluation signals (RMSD, C.C., S.S., energy errors) are costly to obtain and delayed, this rule-based design already meets our target metrics while preserving interpretability, traceable rollback, and safe self-correction. However, it may not remain optimal once the tool set and task space become significantly larger and more branching.

Future directions. To address these limitations, we plan to: (i) integrate **3D-aware imaging and modeling**, such as multi-tilt STEM tomography, to recover depth information and extend AutoMat toward 3D and multilayer structures; (ii) adopt **contrast-sensitive recognition models** and fuse complementary spectroscopy (e.g., EELS/EDX) to reduce element confusion and improve robust element discrimination; (iii) explore **learnable policy optimization**, including reinforcement-learning–based tool selection, once the tool pool and task complexity have grown to a level where purely rule-based scheduling is no longer sufficient; and (iv) **expand the template-retrieval space and introduce a lightweight pre-retrieval calibration stage** (e.g., for pixel scale and defocus) to reduce dependence on exact acquisition matching and improve robustness across different instruments and imaging settings.

As the first end-to-end framework in this field, AutoMat establishes a foundation for future work in extending to 3D systems, improving atomic species identification, and ultimately enhancing structural fidelity, retrieval robustness, and predictive reliability.

A.12 Use of LLMs

In this work, large language models (LLMs) were used solely for text polishing and grammar checking of the manuscript. They were not involved in research design, implementation, or analysis.