Automated Atom-Resolved Defect and Element Classification in 2D MoWSSe HAADF-STEM Dataset

Cheng-Yu Chen[1], Swarnendu Das[1], George Hollyer[1], and Pawan Vedanti[1]

[1]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Reliable automated defect identification in 2D materials is essential for linking atomic-resolution microscopy to structure–property relationships and extracting robust statistics across large datasets. We develop an end-to-end pipeline for an open-source HAADF-STEM dataset [1] of quaternary MoWSSe (ORNL) that detects atomic sites, computes local nearest-neighbor descriptors, and classifies defect environments using transparent, physics-motivated rules. Atomic positions are obtained using Difference of Gaussians (DoG) blob detection, and a KDTree ($K = 3$) provides the three nearest-neighbor distances per atom. Defects are assigned sequentially by deviations from an ideal bond length, with vacancy-related motifs labeled on atoms adjacent to vacancies. Element identification is performed using a sublattice-aware intensity classifier on locally normalized HAADF contrast, enabling atom-resolved defect–composition maps. We report defect- and element-labeled overlays and dataset-level statistics, including composition-dependent defect distributions, on representative MoWSSe images.

Methodology

We implemented an automated analysis pipeline for an open-source HAADF-STEM dataset of quaternary MoWSSe. First, raw microscopy data (single images or stacks) are loaded; for stacks, frames are combined into a single high signal-to-noise image for consistent visualization. Image pixel calibration is read from metadata when available, with manual overrides applied for known cases. We then extract atom-site coordinates using Difference of Gaussians (DoG) blob detection with a tuned scale range and detection threshold to target atom-sized features.

Next, local geometric descriptors are computed for each atom by finding its $K = 3$ nearest neighbors using a KDTree. The three nearest-neighbor distances are used as compact, interpretable descriptors of the local bonding environment. To avoid boundary artifacts from truncated neighborhoods, atoms within a set margin of the image edge are excluded prior to classification.

Defects are classified per atom using sequential, physics-motivated rules based on deviations from an ideal bond length. [2, 3] We label vacancy-related motifs using atoms adjacent to vacancies rather than vacancy sites (which are absent from the detected-atom list), and this atom-resolved representation is also directly compatible with the per-atom element labels planned for the next stage of the project. The current pipeline outputs defect-labeled overlays on the enhanced images, per-class counts, class-separated distributions of local distance metrics, and per-atom CSV files containing atom indices, positions, and defect class labels. Additional implementation details and parameter choices are provided in the project GitHub repository.

Element labels are assigned using a sublattice-aware, intensity-based classifier. After local tile-based normalization of the HAADF-STEM image to suppress background variations, we sample a normalized intensity for each detected atom. Atoms are then separated into A and B sublattices using a nearest-neighbor reciprocity criterion. For each sublattice, the intensity distribution is modeled with Gaussian peaks (two for A sites and three for B sites), fit using aggregated data to obtain robust peak locations. Each atom is labeled by the closest peak for its sublattice, producing per-atom element maps that are directly index-matched to the defect labels for subsequent composition-dependent statistics. Additional implementation details and parameter choices are also provided in the project GitHub repository.

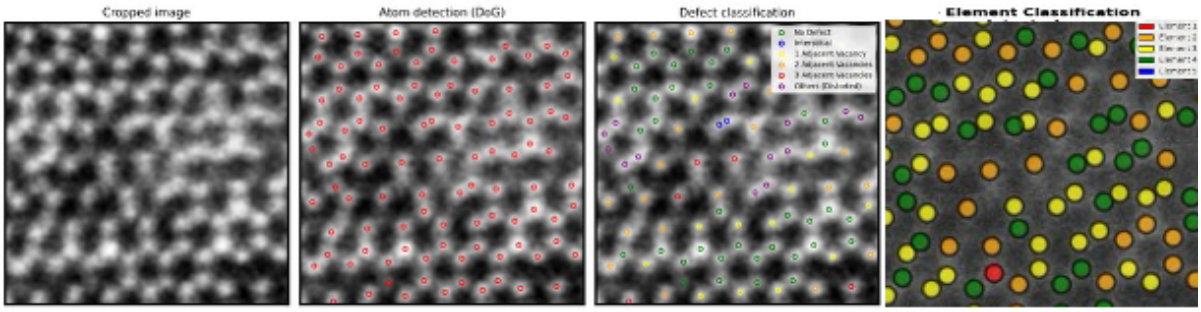

Figure 1: Workflow visualization. (a) HAADF-STEM image (stack-summed and contrast-enhanced). (b) Atom detection result on the same crop using Difference of Gaussians (DoG), shown as detected atomic sites overlaid on the image. (c) Defect classification overlay for the detected atoms, where each atom is assigned exactly one class based on its local nearest-neighbor geometry (three nearest-neighbor distances) using sequential decision rules. (d) Element identification overlay, where each atom is assigned an element label using a sublattice-aware, intensity-based classifier on locally normalized HAADF contrast.

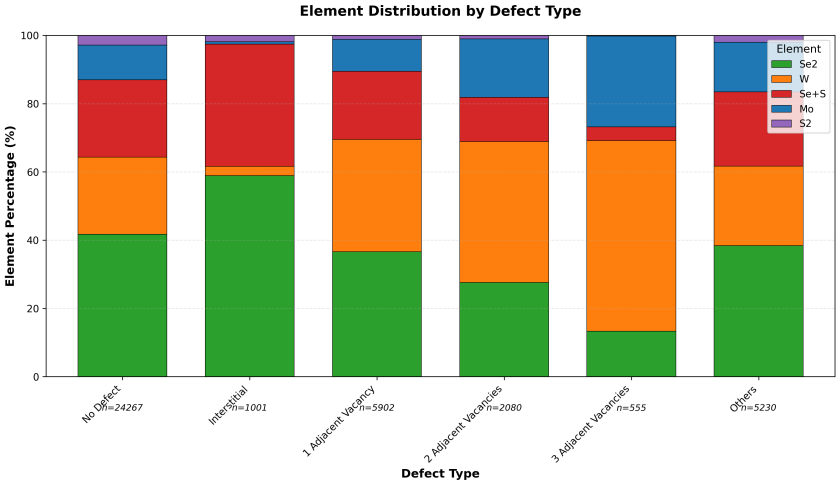

Figure 2: Elemental dependence of different types of defects

Results and Discussion

Figure 2 summarizes the element composition conditioned on local defect class. Clear, class-dependent trends are observed. First, the no-defect population is broadly mixed, consistent with a chemically disordered alloy in regions that retain near-ideal nearest-neighbor geometry. In contrast, vacancy-related environments show progressively stronger enrichment of the metal species as the inferred vacancy order increases. Notably, the three-adjacent-vacancy class is dominated by W with a substantial Mo fraction, while Se_2-type sites are strongly suppressed. This behavior is consistent with the fact that our vacancy classes are assigned to atoms adjacent to missing neighbors: as the local coordination decreases (one to three adjacent vacancies), the remaining atoms in that neighborhood increasingly correspond to metal-site atoms and/or metal-rich local configurations.

The interstitial-like class shows a distinct composition signature compared to vacancy-related classes, indicating that locally crowded geometries are not simply a subset of vacancy motifs but reflect different structural environments. The “Others” (distorted) class also exhibits a composition distribution closer to the no-defect baseline, suggesting that a large fraction of these sites correspond to moderate lattice distortions, imperfect neighborhoods, or detection/classification edge cases rather than a single dominant defect motif.

Overall, these results demonstrate that combining atom-resolved defect labeling with atom-resolved element identification enables composition-dependent defect statistics. Such joint analysis provides a quantitative pathway to test whether specific defect environments preferentially occur on particular sublattices or are stabilized in metal-rich or chalcogen-rich local regions, and it motivates future validation against simulated contrast and first-principles predictions.

Associated Content

- GitHub repository (code, documentation, and full output files): https://github.com/vnpawan/Mic_Hack_2025_Penn
- Project walkthrough video (YouTube): https://youtu.be/ryvBRTgC4_c

References

[1] Rama Vasudevan and Jordan A Hachtel. *MoWSSe HAADF Image Dataset*. Tech. rep. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States). Oak Ridge . . ., 2021.

[2] Hannu-Pekka Komsa and Arkady V Krasheninnikov. "Native defects in bulk and monolayer MoS 2 from first principles". In: *Physical Review B* 91.12 (2015), p. 125304.

[3] KC Santosh et al. "Impact of intrinsic atomic defects on the electronic structure of MoS2 monolayers". In: *Nanotechnology* 25.37 (2014), p. 375703.