Supplementary Information

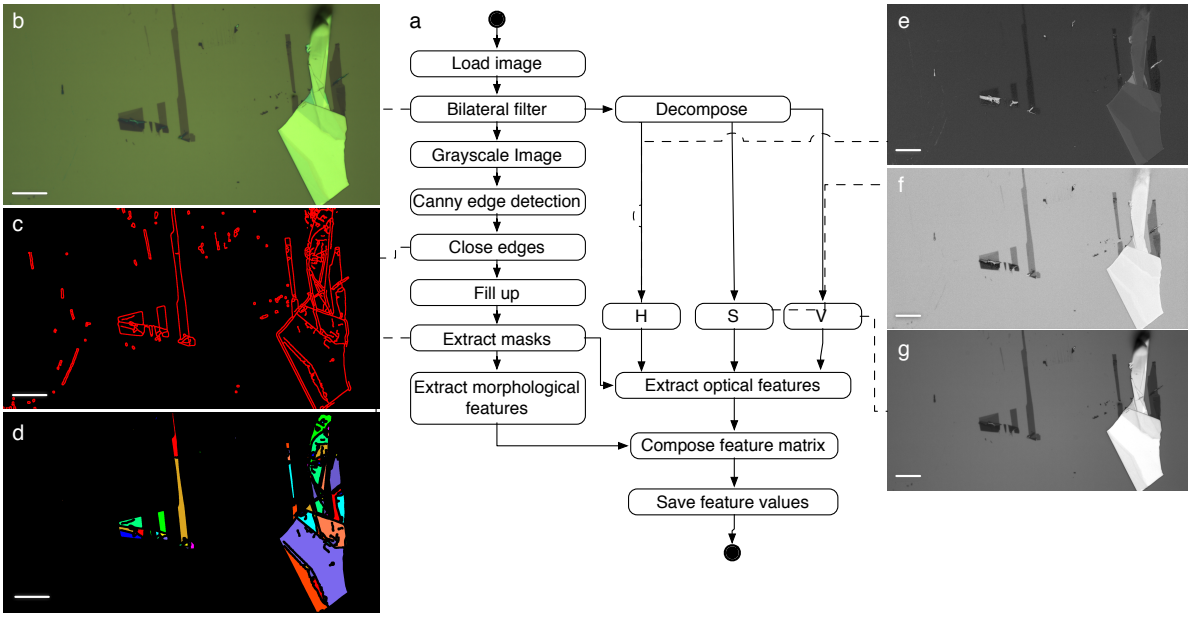
Classifying optical microscope images of exfoliated graphene flakes by data-driven machine learning

Satoru Masubuchi[1, *] and Tomoki Machida[1, *]

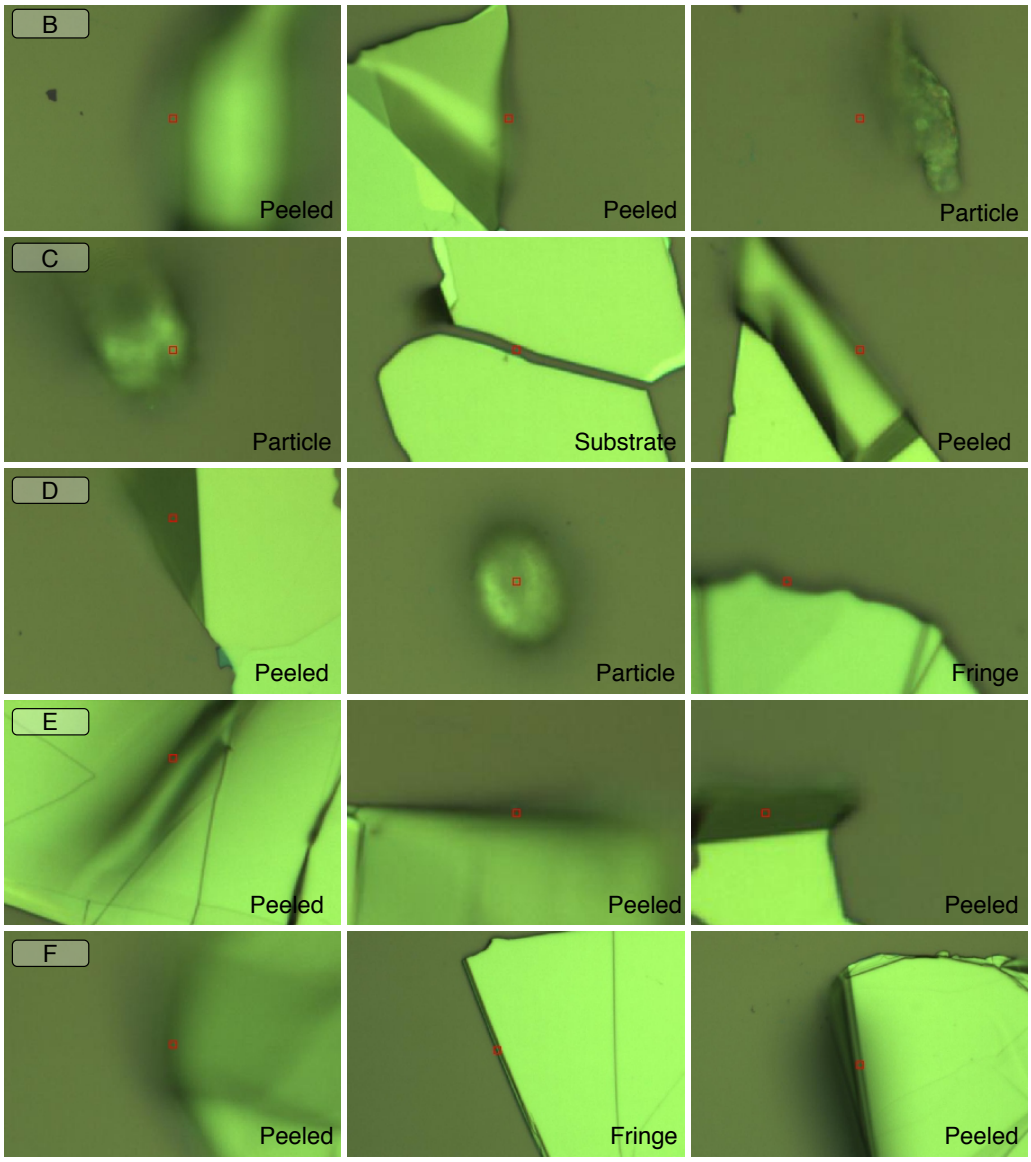
[1]*Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505 Japan*

*Correspondence: msatoru@iis.u-tokyo.ac.jp, tmachida@iis.u-tokyo.ac.jp

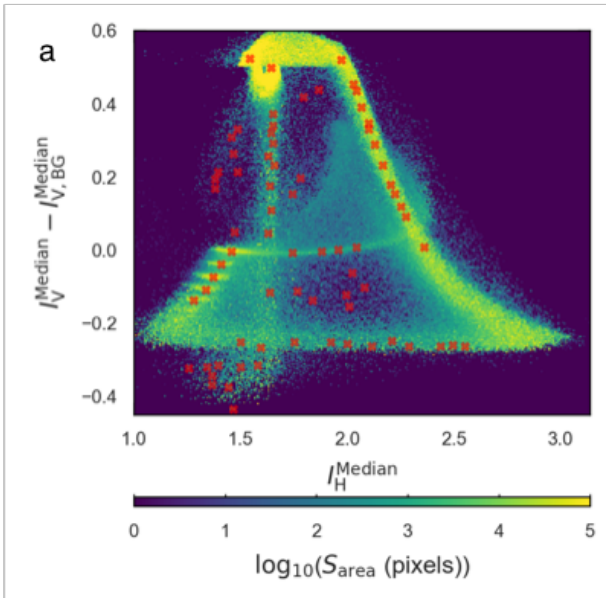
Supplementary Figures



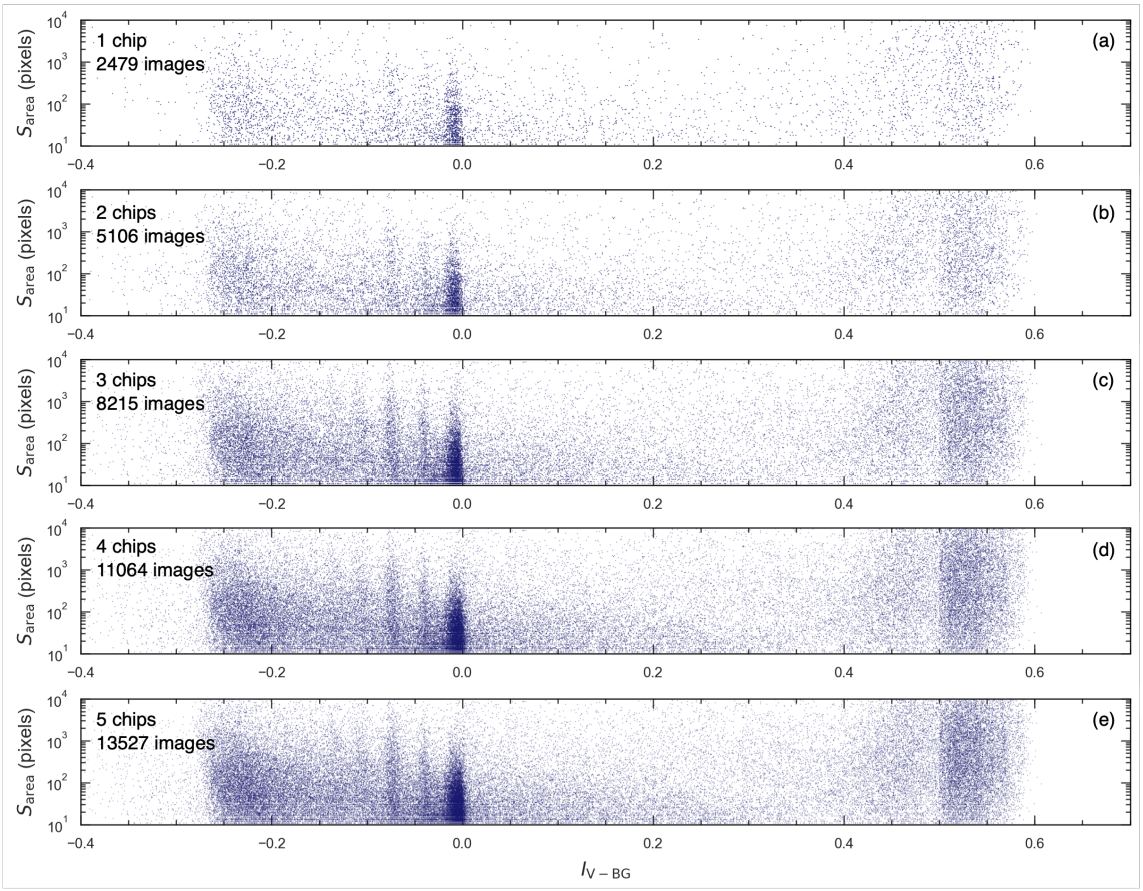
Supplementary Figure 1. Image processing pipeline to extract morphological and optical features. (a) Schematic of image processing algorithm for extracting features of atomically thin 2D crystals from the optical microscope images. (b) Optical microscope image of exfoliated graphene flakes on SiO_2/Si substrate, where the scale bar represents a length of 20 µm. The image includes bilayer graphene, trilayer graphene, and thick graphite flakes. (c)-(g) Representative images generated at each step. The correspondence between images and blocks are indicated by the dashed lines. The scale bar represents a length of 20 µm.



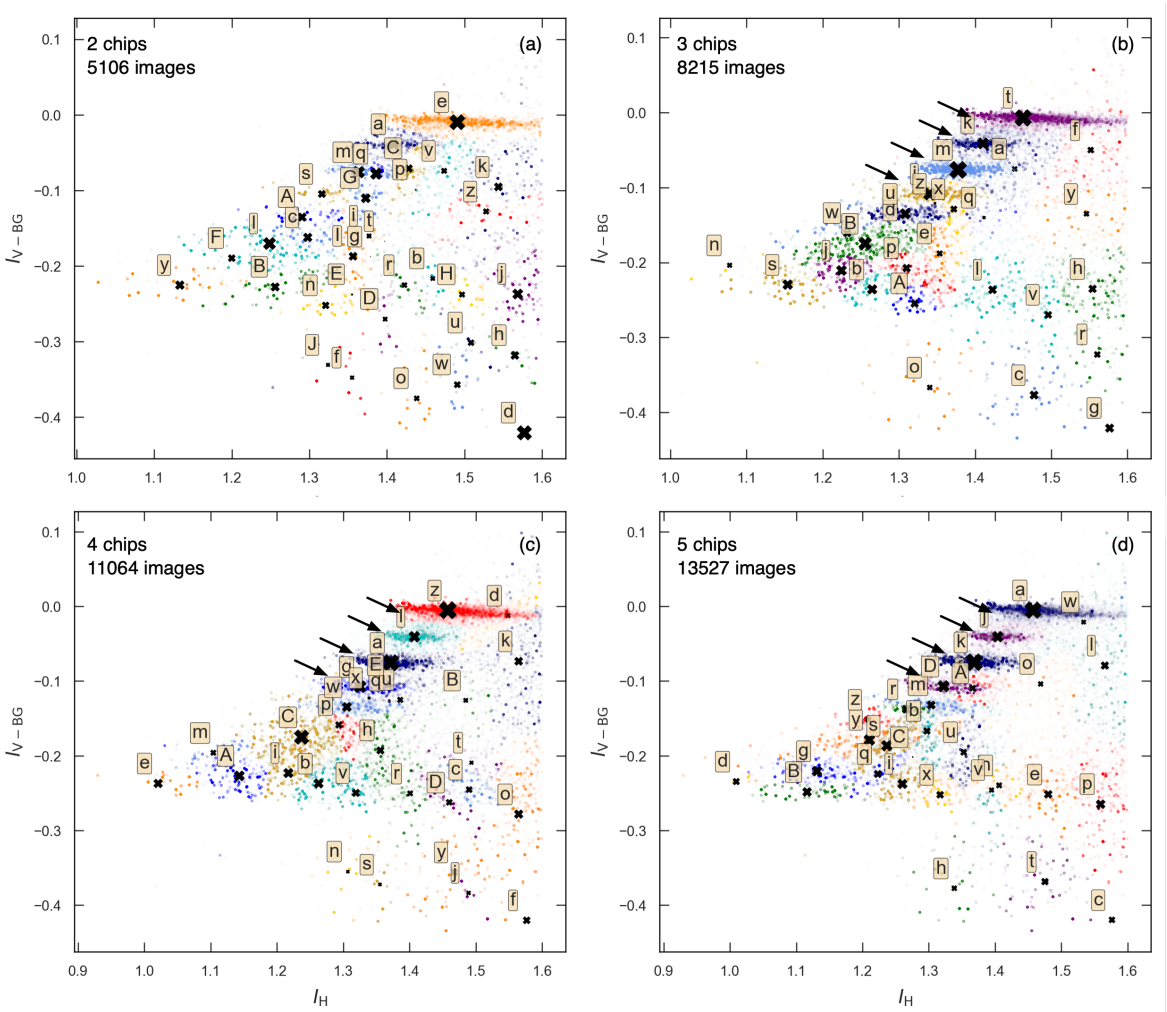
Supplementary Figure 2. Typical examples of optical microscope images that lead to misclassification. For all the clusters B-F, the graphite flakes peeled from the substrate (Peeled), fringes of graphite (Fringe), the silicon substrate surrounded by the graphite flakes (Substrate), and particles (Particle) are the main sources of the misclassification.



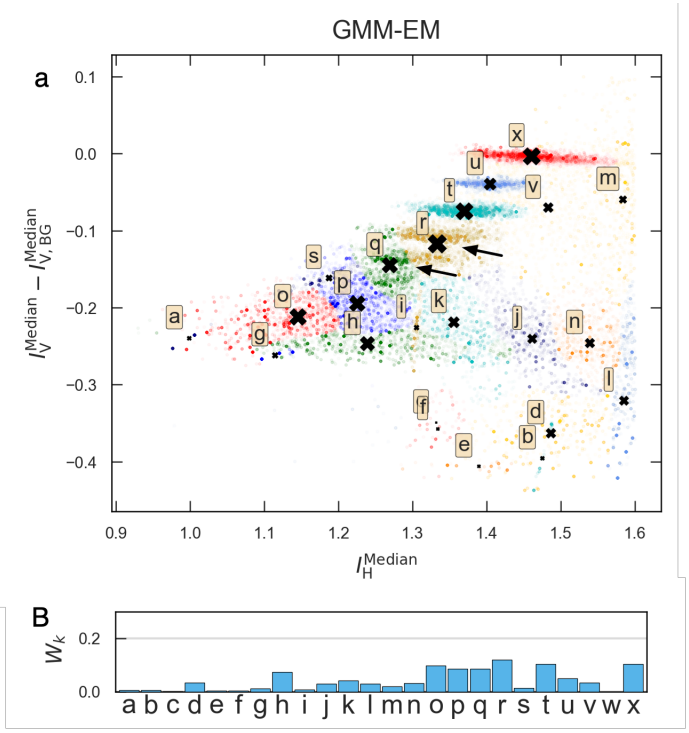
Supplementary Figure 3. The results of peak detection applied to the area histogram S_{area}. (a) Cumulative area histogram as a function of (I_H, I_{V-BG}). The positions of the detected peaks are indicated by the red crosses.



Supplementary Figure 4. Scatter plots of the feature values for varying number of analyzed images. The scatter plots of S_{area} vs. $I_{\mathrm{V-BG}}$ obtained by analyzing (a) 1 silicon chip and 2479 images, (b) 2 silicon chips and 5106 images, (c) 3 silicon chips and 8215 images, (d) 4 silicon chips and 11064 images, and (e) 5 silicon chips and 13527 images.



Supplementary Figure 5. Clustering results of the Bayesian Gaussian mixture model with the Dirichlet process for varying number of analyzed images. (a) 2 silicon chips and 5106 images, (b) 3 silicon chips and 8215 images, (c) 4 silicon chips and 11064 images, and (d) 5 silicon chips and 13527 images.



Supplementary Figure 6. Clustering results of Gaussian mixture model with expectation maximization (GMM-EM). (a) The cluster centers are indicated by filled crosses. The weights of each cluster are indicated by the sizes of cross marks. (b) The extracted weights for the clusters of GMM-EM.

Supplementary tables

Morphology features

- Circularity
- Compactness
- Contlength
- Convexity
- Rectangularity
- Anisometry
- Bulkiness
- Struct Factor
- Outer Radius
- Inner Radius
- Inner Height
- Inner Width
- Dist Mean
- Dist Deviation
- Roundness
- Num Sides
- Connect Num
- Holes Num
- Area Holes
- Max Diameter
- Orientation
- Euler Number
- Area
- Column
- Height
- Row
- Width

Optical features

- BG_mean_(h, s, v, g)
- BG_median_(h, s, v, g)
- Mean_(h, s, v, g)
- Median_(h, s, v, g)
- Entropy
- Anisotropy

Basic information

- Filenames

Supplementary Table 1. Feature values extracted by the presented algorithm.

Supplementary notes

Supplementary note 1

Here we discuss the minimum number of optical microscope images required to make good clustering results. In Supplementary Figure 4, we show the scatter plots of the feature values S_{area} vs. $I_{\mathrm{V-BG}}$ obtained by analyzing the differing number of the silicon chips thereby that of the optical microscope images. When the analysis was conducted for one silicon chip and 2479 images (Supplementary Figure 4 (a)), no vertical stripe patterns were emerged. When the number of silicon chips (n) and optical microscope images (m) were increased as $(n, m) =$ (b)$(2, 5106)$, (c)$(3, 8215)$, (d)$(4, 11064)$, and (e) $(5, 13527)$, the vertical stripe patterns became gradually discernible. From these observations, we could state that around $5000 - 10000$ optical microscope images are required for obtaining the vertical stripe patterns deriving from monolayer-, bilayer-, and trilayer graphene flake as discussed in the main text.

When the clustering analysis were conducted using the BGMM-DP, the successful clustering started from $(n, m) = (3, 8215)$ [Black arrows in Supplementary Figure 5 (b)], and the clustering in small number of optical microscope images $(n, m) = (2, 5106)$ tended to fall into misclassification. From these observations, we can estimate that at least 10000 optical microscope images are needed to obtain the successful unsupervised clustering results.

Supplementary note 2

Gaussian mixture model with expectation maximization

We describe the results of the clustering analysis using the Gaussian mixture model with expectation maximization optimization. In this model, one needs to specify the number of clusters *K* and the initial guess of cluster centers for performing the optimization. Here, to estimate the cluster numbers and the centers, we applied a local peak detection algorithm to the cumulative histogram of the area S_{area} of the regions [Red crosses in Supplementary Figure 3]. The detected peak positions were supplied to the EM algorithm and optimization iteration

was performed. The results are presented in Supplementary Figure 6(a). Compared to the GMM-DP, the clusters consisting of trilayer-graphene and tetralayer-graphene flakes (clusters r and q, respectively, indicated by the black arrows in Supplementary Figure 6(b)) were confused with the different clusters. Even when we changed the initialization conditions, the GMM-EM results did not exhibit the correct clustering results. These results indicate that when applied to our feature matrix, the GMM-EM algorithm tend to result in overfitting; therefore, GMM-DP gives a better performance compared to GMM-EM.