Predicting Peptide Interface Parameters using Image Processing & Machine Learning

Demi Liu[1,4], Sarthak P. Jariwala[2,4], Christopher Nyambura[3], Sid Rath[2,5]

1. Chemistry 2. Material Science and Engineering 3. Chemical Engineering 4. CEI 5. GEMSEC

Background Subtraction

The background approximation and subtraction function uses Morphological Filtering from scikit-image. It returns all bright spots in the image that are smaller than the structuring element that the user defines (rectangular, square or ellipsoidal/circular). It outputs a foreground with some threshold, deletes the background, while preserving the relative height values to use later.
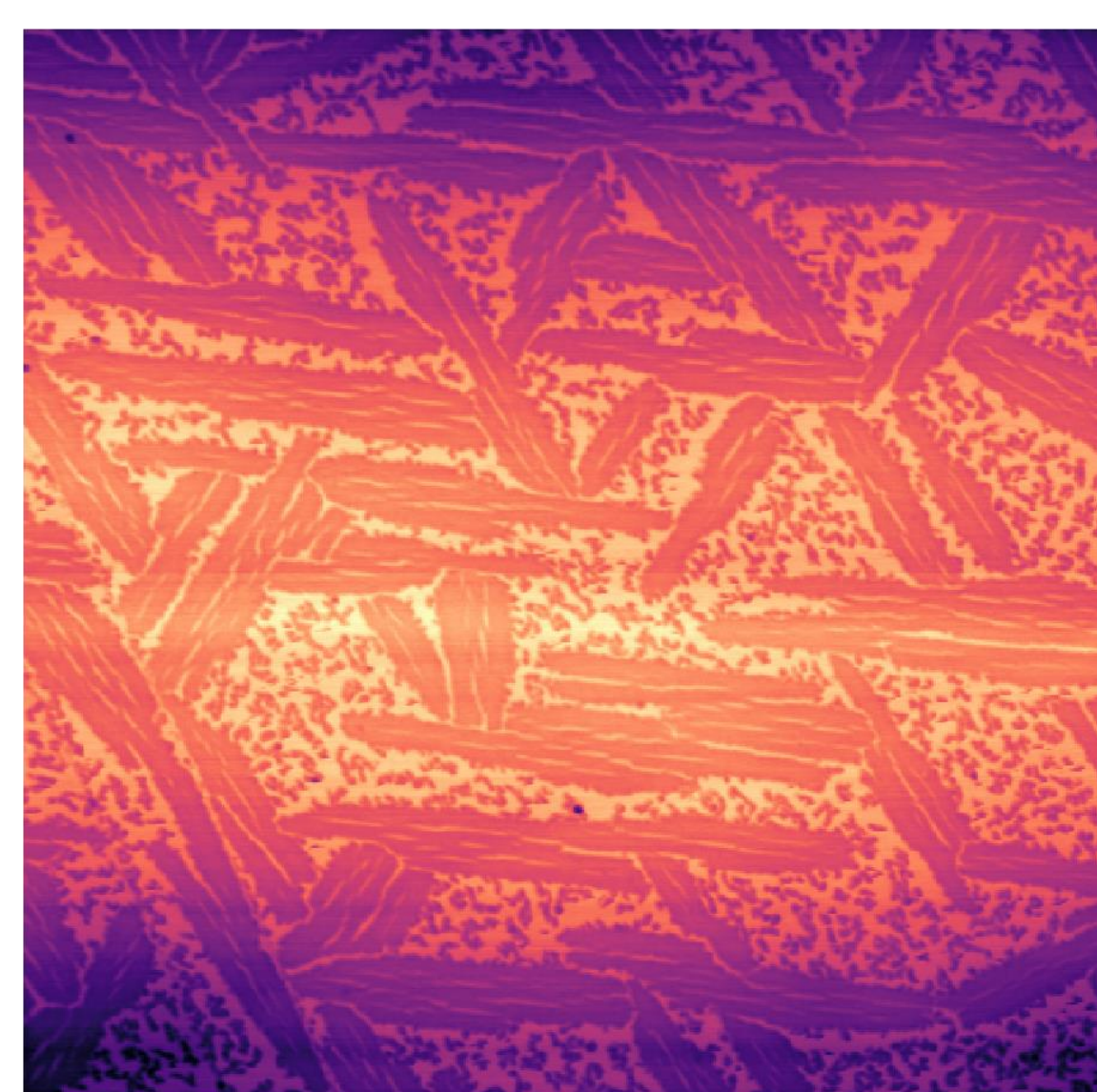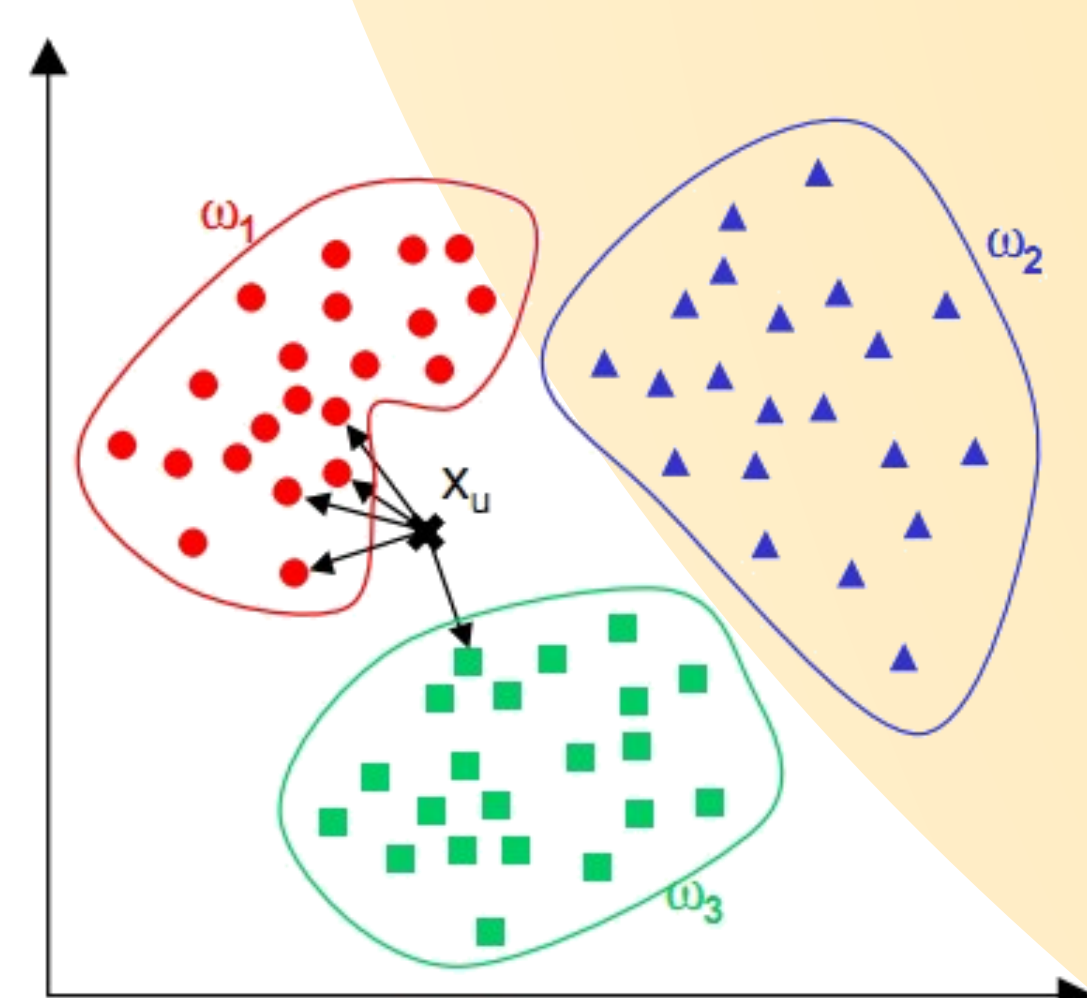
Image Segmentation

Random walker segmentation from scikit-image was employed for this step. Random walk solves an anisotropic diffusion equation with tracers starting from a pre-determined pixel with a certain label (initial marker position). Diffusivity is high, if neighboring pixels have similar values, whereas widely varying pixels values results in low diffusion. Thus, diffusion has an inverse relationship to the gradient height.
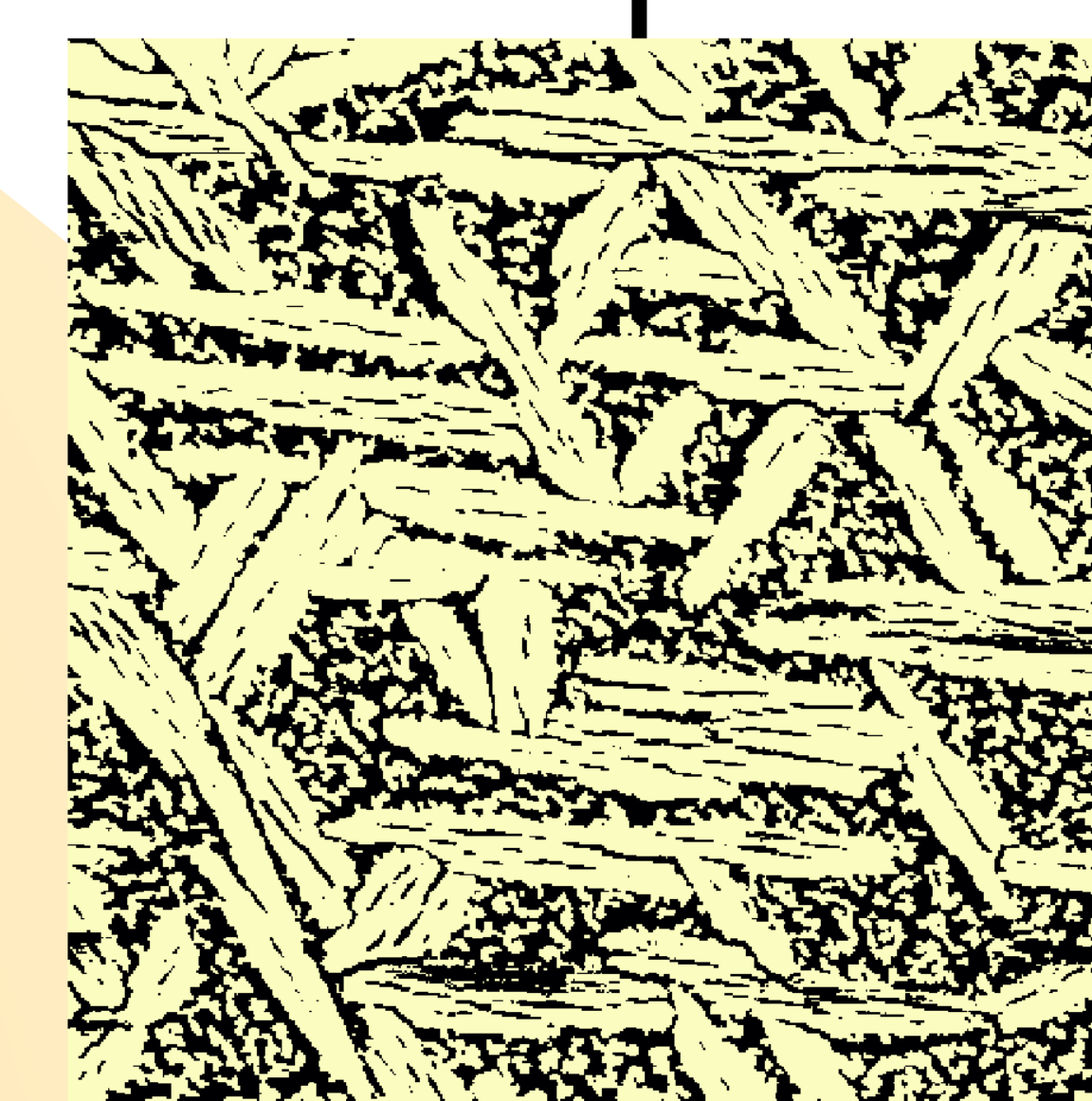
Prediction of Bio/Nano Interface Parameters using KNN Clustering and Regression

K-Nearest Neighbor Clustering was performed on 100 images, split into random train/test datasets. Images were clustered based on pH, concentration of the peptide solution used and image descriptors generated from segmented images. Next, regression was used to predict order to disorder ratio and percent coverage for a set of different input parameters.

Ordered Regions
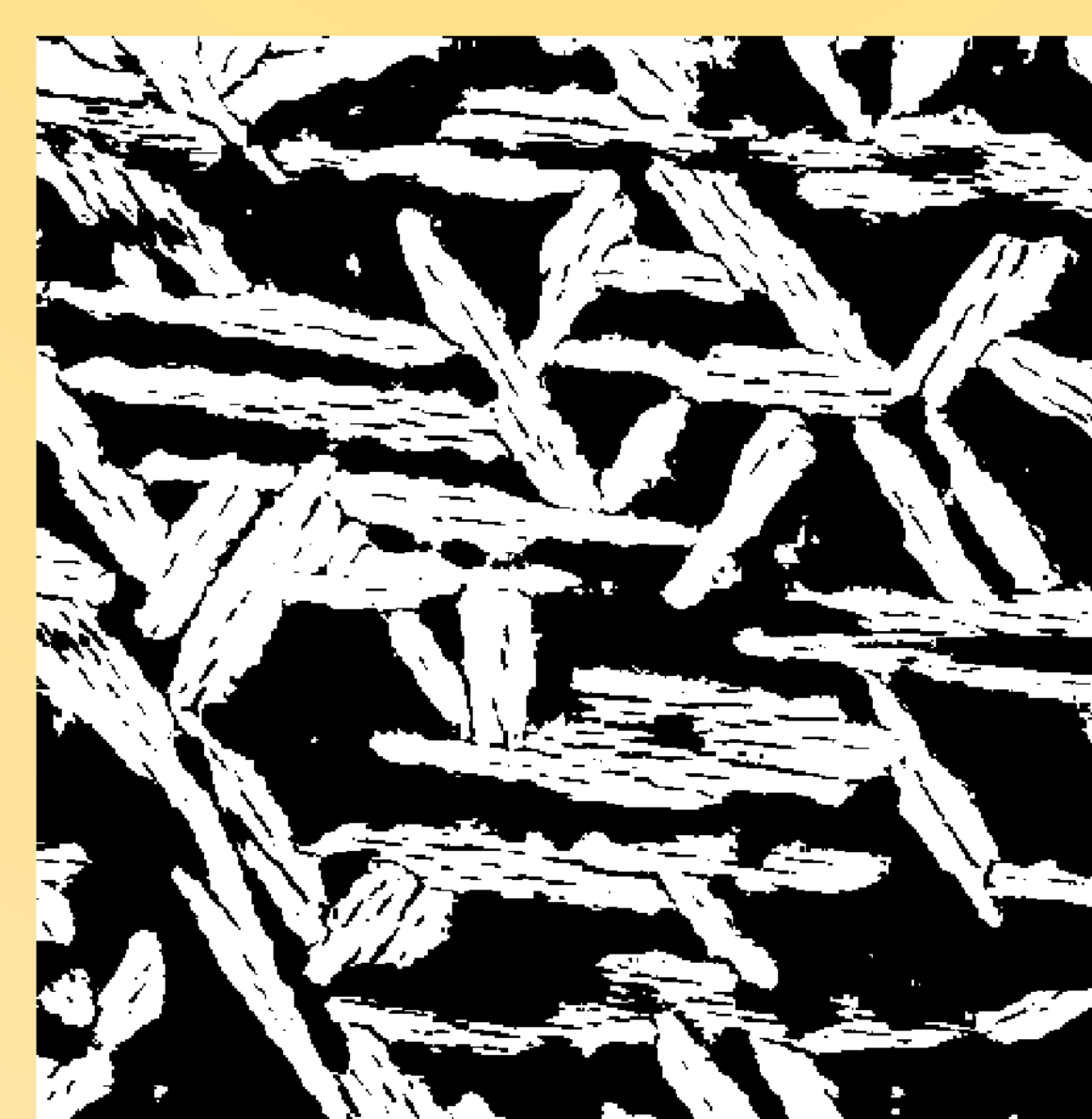
Generation of Image Descriptors from Segmented Images

Percentile Filter from scipy.ndimage was used to separate the ordered and disordered regions. Parameters such as percent coverage, order to disorder ratio, order and disorder percentage were calculated. These parameters serve as image descriptors that are specific to the experimental conditions.

Disordered Regions

Sarikaya Research Group
Molecular Biomimetics for Technology & Medicine

GEMSEC
Genetically Engineered Materials
Science & Engineering Center

NSF

DIRECT
Data Intensive Research
Enabling Clean Technologies

W