

## Supporting Information

# A High Throughput Bubble Screening Method for Combinatorial Discovery of Electrocatalysts for Water Splitting

*Chengxiang Xiang<sup>1\*</sup>, Santosh K. Suram<sup>1</sup>, Joel A. Haber<sup>1</sup>, Dan W. Guevarra<sup>1</sup>, Jian Jin<sup>2</sup>,  
John M. Gregoire<sup>1\*</sup>*

<sup>1</sup>Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena  
CA 91125

<sup>2</sup>Engineering Division and Joint Center for Artificial Photosynthesis, Lawrence Berkeley  
National Laboratory, Berkeley CA 94720

## Support Vector Machines

The objective of Support Vector Machine (SVM) classification is to identify hyperplanes that maximize the separation between various classes of data.

The mathematics of SVM in its' primal form can be expressed as:

$$\arg \min_{(w,b)} \frac{1}{2} \|w\|^2, \text{ subject to } y_i(w \cdot x_i - b) \geq 1. \quad [1]$$

Wherein,  $\|w\|$  is inversely proportional to the margin of the SVM classifier.

Using Lagrange multipliers, the primal form can be represented as

$$\arg \min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(w \cdot x_i - b) - 1] \right\} \quad [2]$$

The dual form of the above equation expressed below is the basis for non-linear SVM classification.

$$L(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j (x_i \bullet x_j), \text{ subject to } \alpha_i \geq 0 \quad \forall i \in [1, 2, \dots, n] \quad [3]$$

Wherein  $x_i \bullet x_j$  is the dot product of feature vectors  $x_i$  and  $x_j$ .

Eq. 3 represents a linear classifier as indicated by the presence of dot product of the feature vectors. Cortes et al. (Cortes & Vapnik, 1995) suggested replacing the dot product with a nonlinear kernel function  $k(x_i, x_j)$  to construct a nonlinear classifier.

In this work, we use a radial basis kernel (rbf-kernel) function (eq. 4)

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad [4]$$

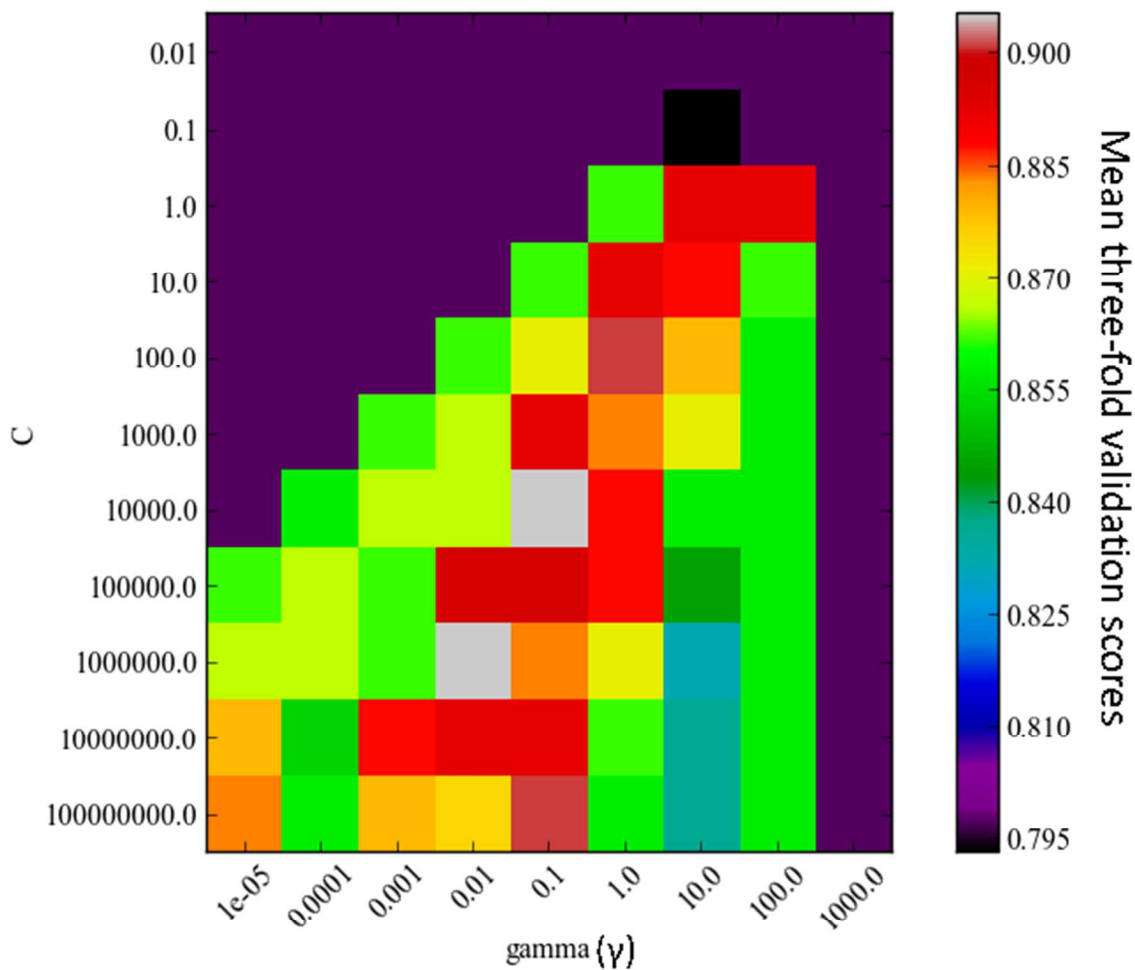
where the effect of neighboring compositions decreases with increasing gamma. In the SVM framework, the implementation of this kernel with a soft-margin parameter C defines the complexity and the generalizing capabilities of the SVM

classifier. For a given value of  $\gamma$ , larger C values increase the complexity of the decision surface of the SVM classifier.

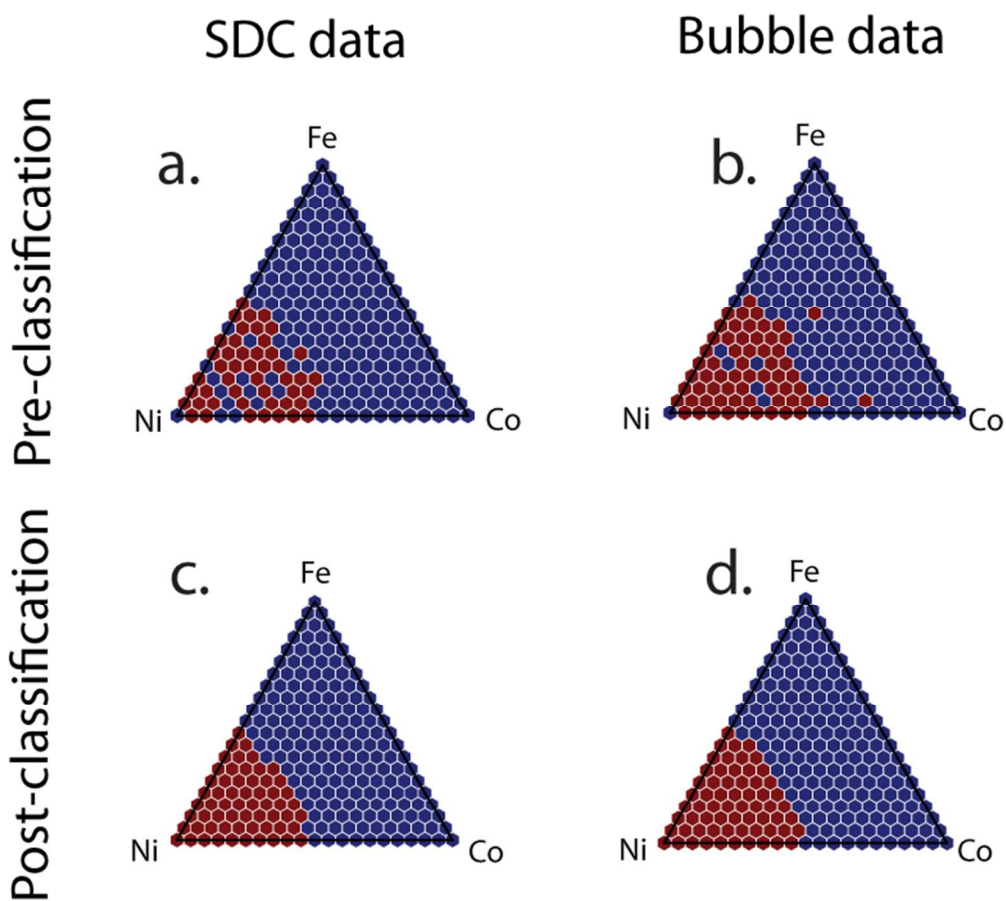
### **Application**

In this article, SVM classification is applied to compositional maps of catalyst performance, provided by both the SDC and bubble screening techniques. Here, we use the SDC dataset as an example to explain the SVM classification procedure. The training data is obtained by labelling the top 20 percentile of SDC data as “good” and the rest as “bad” catalysts. The composition of each component in the ternary Ni-Fe-Co library represents the feature vectors of this training dataset. The feature vectors are normalized by transforming them into zero mean, unit standard deviation vectors to remove any variance bias in the dataset. Three training datasets are created from this dataset by randomly choosing one-third of the samples for testing the performance of the SVM classifier. The mean of the validation scores (fraction of correct predictions) of prediction of the SVM classifier in these three datasets is used as a measure to quantify the performance of the SVM classifier. Fig. S1 shows the mean three-fold validation scores for the SVM classifier as a function of a range of values for the parameters C and  $\gamma$ . In our analysis, we observed that a general value of  $C = 1$  and  $\gamma = 10$  is appropriate for all the datasets while minimizing the deleterious effects of a complex decision surface and constricting the effect of adjacent compositions to a local neighborhood.

Fig. S2 shows the pre-classified and post-classified labels of the compositions in the Ni-Fe-Co library. The generalization capability of the SVM classifier enables it to identify a cluster of composition region wherein the catalysts are active while rejecting the outliers present in the original dataset. Due to the outlier rejection achieved by the SVM classification, comparison of the post-classified SDC and bubble data is a more reliable method to quantify the correlation between these techniques.



**Figure S1.** A false color representation of mean three-fold validation scores as a function of the SVM classification parameters  $C$  and  $\gamma$ . A high validation score while minimizing the complexity of the SVM classifiers' decision surface (i.e., small value of  $C$ ) and constricting the effect of adjacent compositions to a local neighborhood (i.e., small value of  $\gamma$ ) is desired. In this analysis, parameters  $C=1$  and  $\gamma=10$  satisfy the above mentioned criteria.



**Figure S2.** Ternary compositional maps for the scanning droplet cell measurements before (a) and after (c) applying a support-vector machine (SVM) based classification method. Ternary compositional maps for the bubble screening data before (b) and after (d) applying a support-vector machine (SVM) based classification method. Pre-classification acceptance rates of 20 and 27 percentile were chosen for SDC data and bubble data respectively (red represents good catalysts). The SVM classified datasets show rejection of outliers.

### References:

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. doi:10.1007/BF00994018