Supplementary Information to "Active learning for regression of structure-property mapping: the importance of sampling and representation "

Hao Liu[1] Berkay Yucel[2] Baskar Ganapathysubramanian[3]
Surya R. Kalidindi[2] Daniel Wheeler[4] Olga Wodo[1]

[1]Materials Design and Innovation Department, University at Buffalo, Buffalo, NY, USA
{olgawodo}@buffalo.edu
[2]The School of Materials Science and Engineering, the School of Computational Science and Engineering, Georgia Institute of Technology, GA, USA. [3]Mechanical Engineering Department, Iowa State University, IA, USA.
[4]Materials Science and Engineering Division, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA.

1 List of descriptors

Table 1 lists the descriptors used in this work.

2 Feature importance score

In this work, we calculate the feature importance score based on random forest (RF) algorithm [1, 2]. RF is constructed using a set of decision trees. Each decision tree has a set of internal nodes and leaves and uses these to calculate the feature importances based on how much the feature reduces the impurity (e.g. MSE) when it split a node. These importance scores are accumulated over all nodes. Finally, the feature importance scores are averaged across all tree models.

3 Quantitative analysis of sampling strategy–three metrics

In this work, we use three metrics to compare and contrast different sampling strategies:

(i) Wasserstein distance, also known as earth mover distance, which is derived from optimal transport theory [3]. In this work, for each iteration, the Wasserstein distance metric is used to calculate the distance between the microstructure distribution, μ, at each iteration and the complete microstructure distribution, ν. Following work [4], the Wasserstein distance is defined as:

$$W_2(\mu,\nu) = \left(\min_{\Gamma \in \mathbb{R}_+^{m\times n}} \sum_{i,j} \Gamma_{i,j} \|\gamma_{x_i} - \gamma_{y_j}\|_2 \right)^{1/2} \quad \text{s.t.} \quad \Gamma\mathbf{1} = \mu, \Gamma^T\mathbf{1} = \nu, \Gamma \geq 0 \tag{1}$$

where, μ and ν are two probability distributions, Γ is the optimal transport matrix and the $L2$ norm is the distance between samples (microstructures) γ_{x_i} and γ_{y_j} from two distributions.

(ii) Entropy is used to measure the diversity of the data pool at each iteration. Formally, the Shannon entropy of a probability distribution, H, is defined as:

$$H = -\int p(\gamma)\log(p(\gamma))\, d\gamma \tag{2}$$

where $p(\gamma)$ is the probability density function of variable γ. Here, the kernel density estimator determines p of microstructure feature space, γ.

Table 1: The list of descriptors used in this work, the names of descriptors, and the corresponding `GraSPI` names. The abbreviations D, A, Ca, An, CC correspond to donor phase, acceptor phase, cathode, anode, and connected components, respectively. Note that the abbreviation D in this table is used for consistency with the software and should not mixed with the vector of descriptors from the main manuscript.

d_i	Descriptor	GraSPI name
d_1	Fraction of D voxels	ABS_f_D
d_2	Weighted fraction of D voxels in 10 distance to interface	DISS_wf10_D
d_3	Interfacial area	STAT_e
d_4	Number of D voxels	STAT_n_D
d_5	Number of A voxels	STAT_n_A
d_6	Number of D CCs	STAT_CC_D
d_7	Number of A CCs	STAT_CC_A
d_8	Number of D CCs connected to An	STAT_CC_D_An
d_9	Number of A CCs connected to Ca	STAT_CC_A_Ca
d_{10}	Weighted fraction of D	ABS_wf_D
d_{11}	Fraction of D voxels in 10 distance to interface	DISS_f10_D
d_{12}	Fraction of interface with complementary paths to An and Ca	CT_f_e_conn
d_{13}	Fraction of D voxels connected to An	CT_f_conn_D_An
d_{14}	Fraction of A voxels connected to Ca	CT_f_conn_A_Ca
d_{15}	Interfacial area with complementary paths	CT_e_conn
d_{16}	Number of D interfacial voxels with path to An	CT_e_D_An
d_{17}	Number of A interfacial voxels with path to Ca	CT_e_A_Ca
d_{18}	Fraction of D voxels with straight rising paths (t=1)	CT_f_D_tort1
d_{19}	Fraction of A voxels with straight rising paths (t=1)	CT_f_A_tort1
d_{20}	Number of D voxels in direct contact with An	CT_n_D_adj_An
d_{21}	Number of A voxels in direct contact with Ca	CT_n_A_adj_Ca

(iii) Informativeness is measured by calculating the average variance of the unselected data pool at each iteration. The definition is shown in equation 4 of the main text.

4 Comparison of sampling for setting 2

Figure S1 depicts the distribution of query points for five sampling strategies. The features of query points for each sampling mimics the results presented in the main document. GSx and uncertainty sampling tend to choose query points from the boundary of the distribution in the latent space. GSy chooses query points to span the range of properties. Finally, iGS balances the sampling uniformity in input and output spaces.

5 Active learning combined with feature selection

The salient features selected based on currently labeled data will be changing along with the active learning process. However, salient features will be stable as sufficient data samples have been learned.

Table S2 and Figure S3 list salient features for the selected iterations of the AL workflow. Initially, 16 features are required to meet the 0.98 accumulated importance score criterion. With subsequent iterations, the number and the list of salient features converge. Table S2 provides the summary of the selected feature. Interestingly, the features used in setting 1 ($d_3, d_{11}, d_{20}, d_{21}, d_2$) are consistently being selected beyond 10 iterations. The value of feature importance for these descriptors increases with iterations, as shown in Figure S3. For example, d_3 scores high even in the early iterations. While the score of d_{20} gradually increases to become the third most important feature at iteration 490.

References

[1] Hao Liu, Berkay Yucel, Daniel Wheeler, Baskar Ganapathysubramanian, Surya R Kalidindi, and Olga Wodo. How important is microstructural feature selection for data-driven structure-property mapping? *MRS Communications*, pages 1–9, 2022.

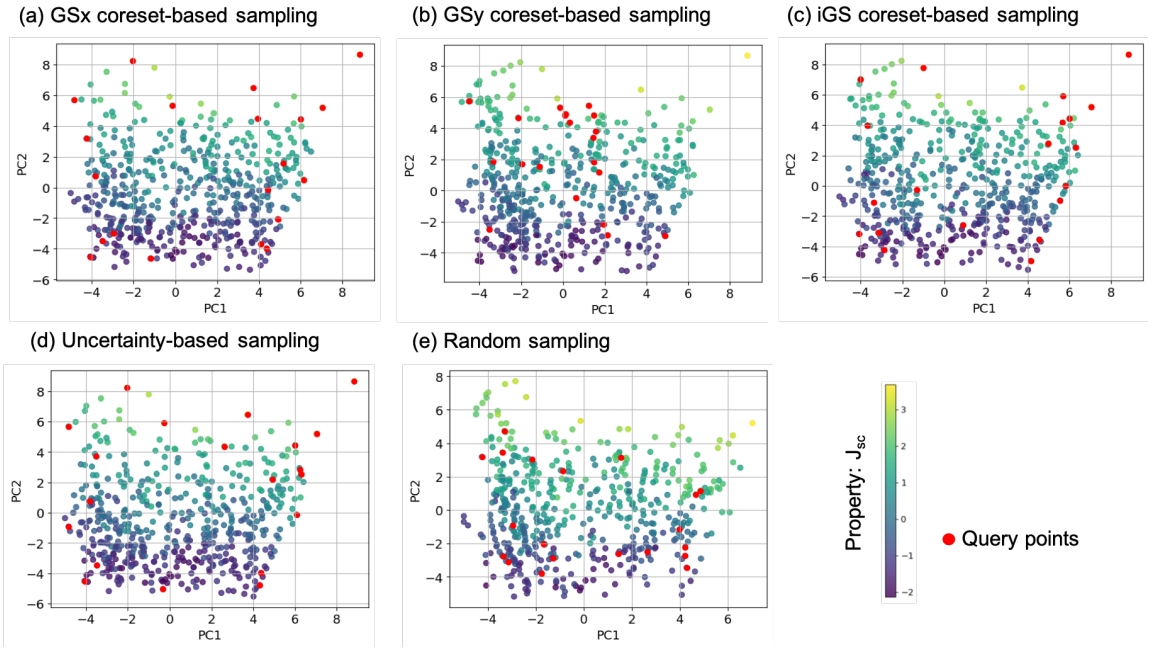

Figure S1: Visualization of query point selection using different sampling strategies: GSx, GSy, iGS, uncertainty sampling, and random sampling with unknown salient features. Each panel highlights 20 points selected using a given strategy (marked red), and also includes remaining points that are color-coded using the property of interest-J_{sc}. Note that each point corresponds to one microstructure projected into the first two principal components of descriptor-based representation.

Iterations	Number of features with accumulated importance > 0.98	Selected features (ordered based on importance score)
1	16	d19, d17, d3, d8, d11, d13, d12, d6, d2, d9, d20, d10, d14, d15, d16, d1
10	13	d3, d2, d11, d8, d17, d19, d15, d18, d21, d13, d12, d6, d9
100	11	d11, d3, d8, d2, d20, d16, d15, d21, d19, d18, d5
490	9	d3, d11, d20, d21, d2, d8, d19, d18, d16

Figure S2: The selected features in the active learning process

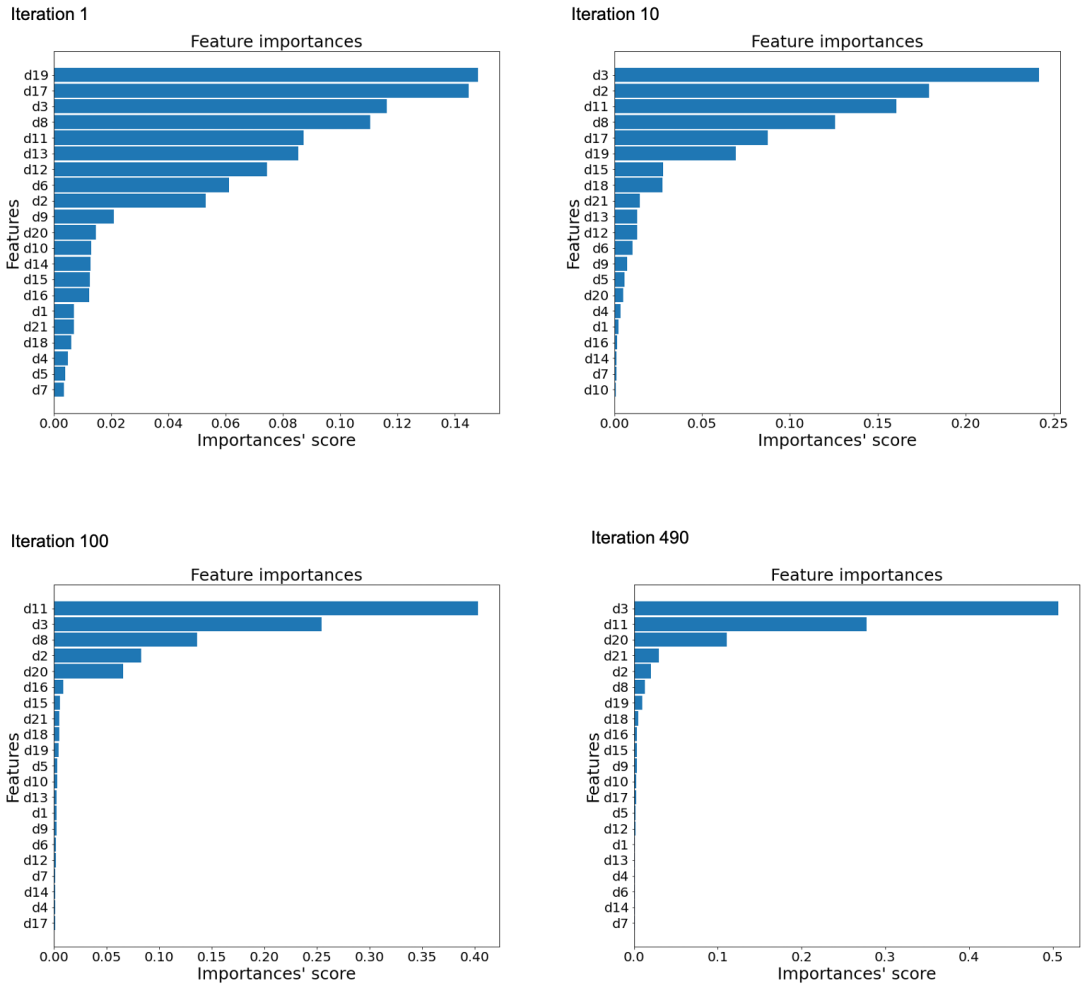

Figure S3: The feature selection results in the active learning process

[2] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Found. Trends Mach. Learn.*, 11:355–607, 2018.

[4] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.