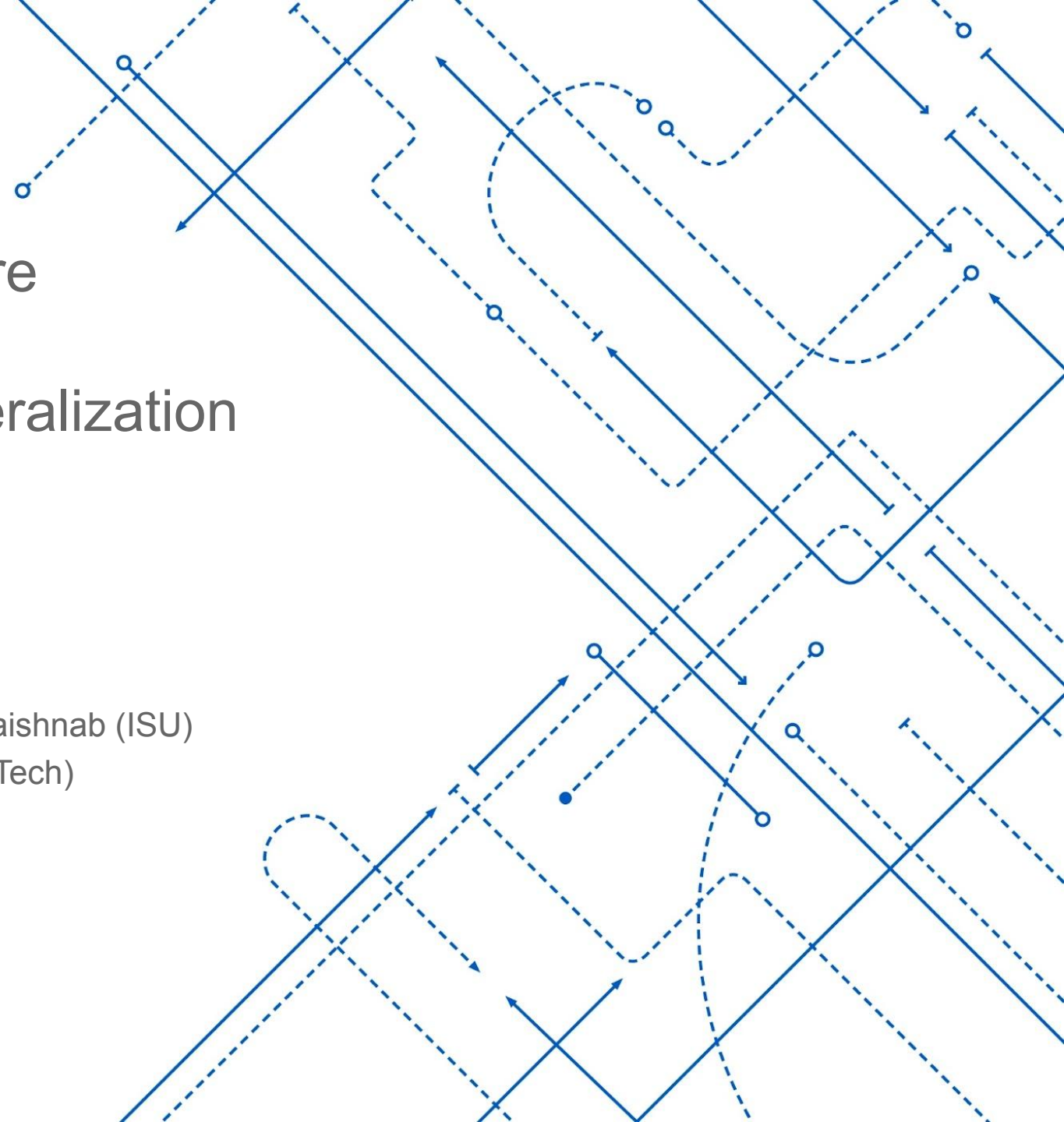


The importance of microstructure representation, and implications for model generalization

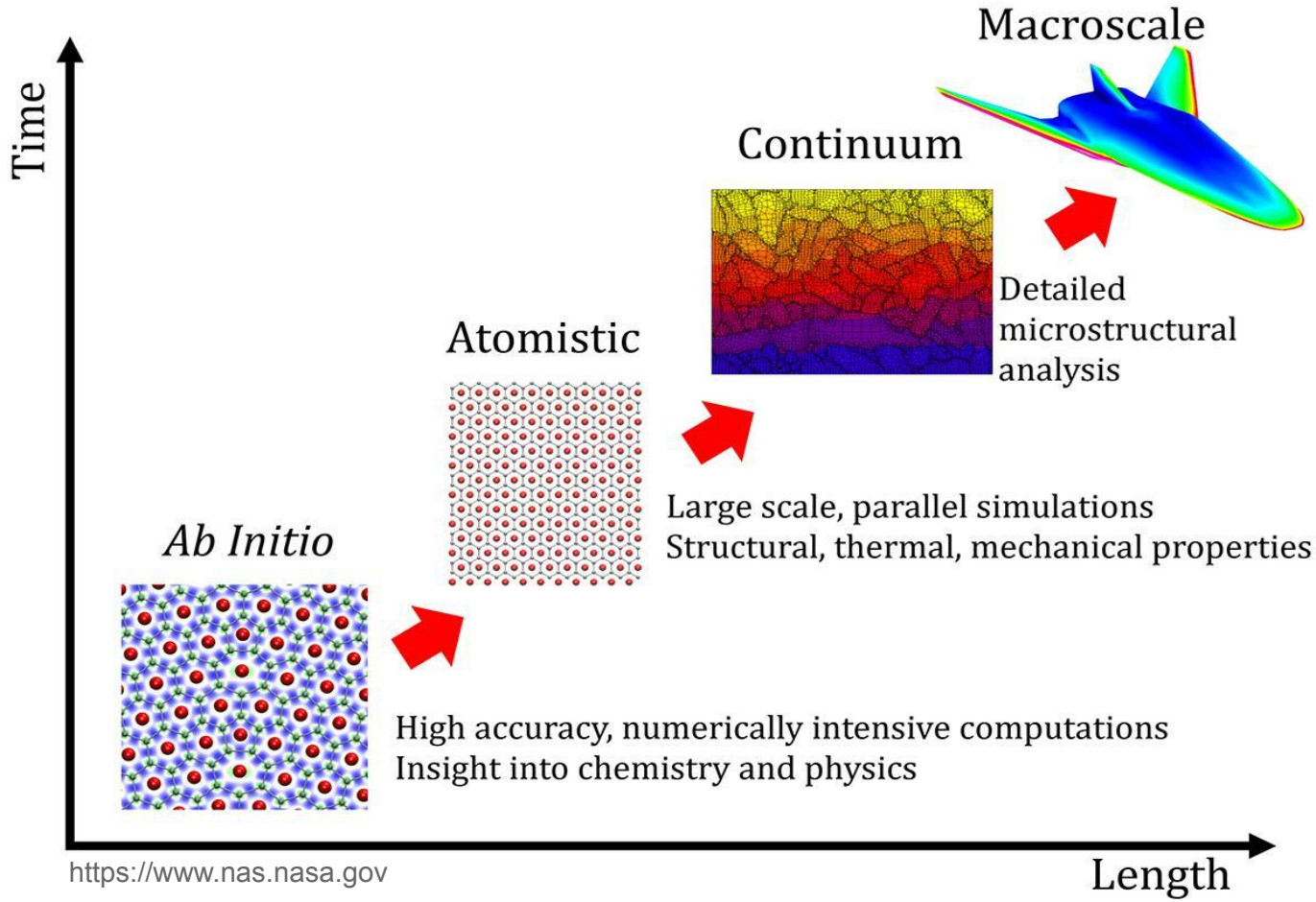
Olga Wodo

H. Liu (UB), N. Parikh (CMU)
B. Ganapathysubramanian (ISU), B. Pokuri (ISU), N. Baishnab (ISU)
D. Wheeler (NIST), S. Kalidindi (GaTech), B. Yucel (GaTech)

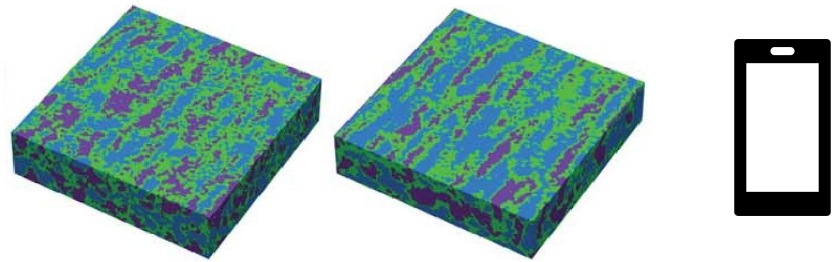
University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Materials behavior depends on scale

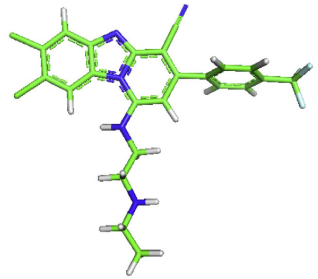


https://www.nas.nasa.gov



One micrograph: 50 M voxels (A. Moule group)

S.N.	Type of Molecular Descriptor	Total Number
1	Constitutional • Functional groups • Molecular weight • Simple Atom counts • Ratio of various types of atoms	235
2	Geometric • Molecular surface area (MSA) • Solvent accessible molecular surface area (SASA) • Ratio of MSA and SASA of various types of atoms	212
3	Circular fingerprint • Presence/Absence of different types of atom pairs at specific spatial distance	2650
4	Quantum chemical • Charges	3548
5	Topological • Atom-pairs	4500



Myriad of software for atomistic and molecular featurization: pyDescriptors, ADAPT, Pentacle, Codessa, Dragon, Magpie (e.g., pyDescriptor calculates 11,145 molecular descriptors)

[60% data curation, 30% features/data representation, 10% machine learning method]

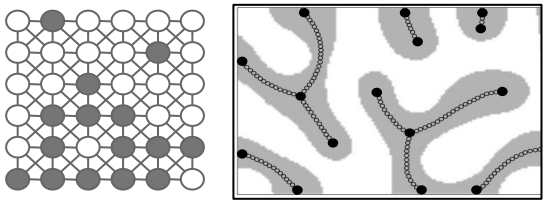
University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Various microstructure representation layers (RL)

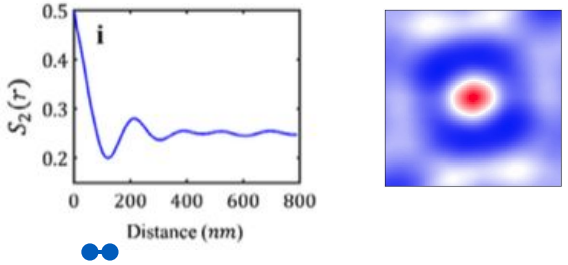
RL0:
Raw data

Image/matrix
size: $N=n_x n_y n_z$

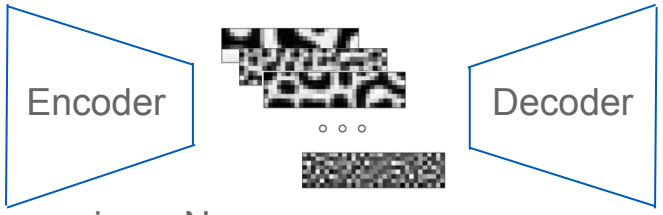
RL1:
Mathematical representation



Graph: size ~N



Statistical function, size: ~N



Neural Network, latent space size: <N

RL2:
Reduced representation

Descriptors:
$[d_1, d_2, \ldots, d_n]$

Size: $n \ll N$

RL3:
Salient features

$[\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_m]$
$m \ll n$

$[\widetilde{PC}_1, \widetilde{PC}_2, \ldots, \widetilde{PC}_l]$
Size: $l \ll N$

$[\widetilde{PC}_1, \widetilde{PC}_2, \ldots, \widetilde{PC}_l]$
Size: $l \ll N$

SP model

$$P = \sum_{i=1} a_i \tilde{d}_i$$

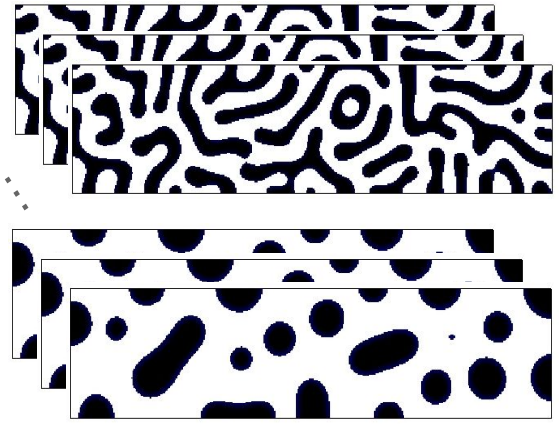
$\{1, \tilde{d}_1, \tilde{d}_2, \tilde{d}_1^2, \tilde{d}_2^2, \tilde{d}_1\tilde{d}_2\}$
....
Neural Networks (with large # coefficients)

$$P = \sum_{i=1} a_i \widetilde{PC}_i$$

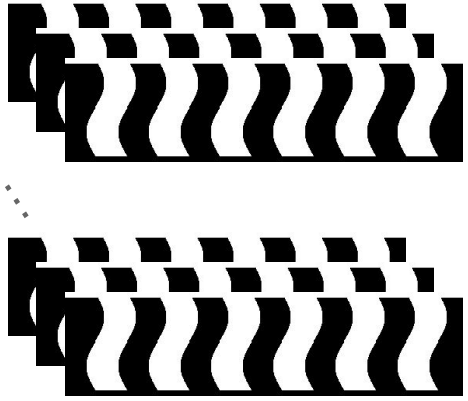
$\{1, \widetilde{PC}_1, PC_2, \widetilde{PC}_1^2, PC_2^2, \widetilde{PC}$

$$P = \sum_{i=1} a_i \widetilde{PC}_i$$

Three representations, three types of morphology and three questions



~2k microstructures



~50 microstructures

~200 microstructures

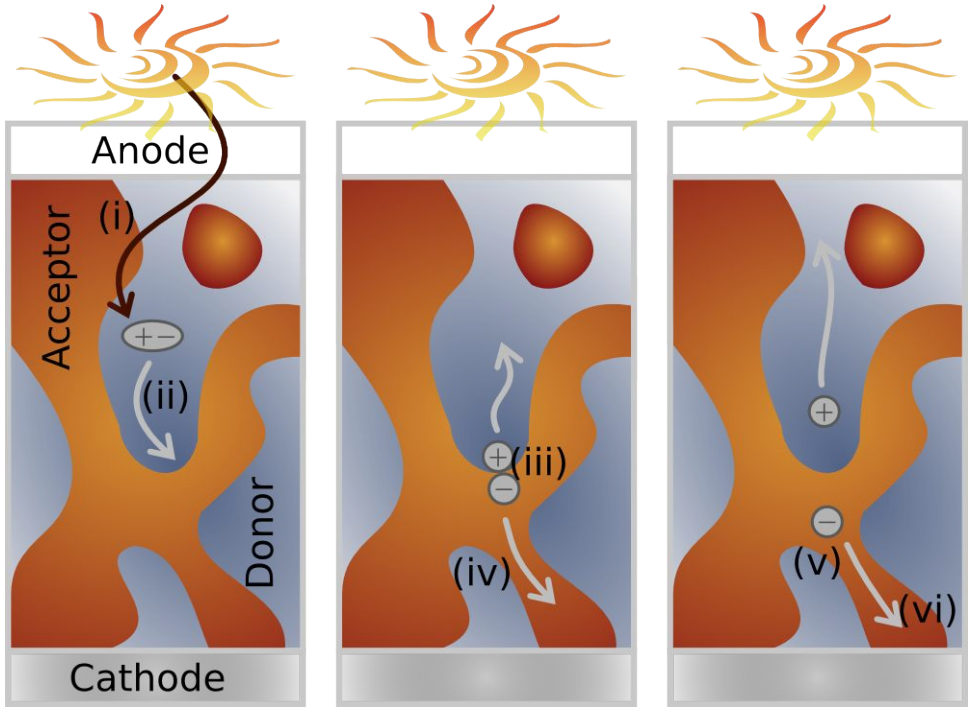
Q1: Does representation matter ?
(purely data-driven vs expert-enriched)
Q2: How generalizable are models built using three representation?
Q3: How generalizable are models across materials systems?

$$P = f(\widetilde{d_1}, \widetilde{d_2}, \widetilde{d_3})$$

H. Liu, B. Yucel, D. Wheeler, B. Ganapathysubramanian, S. Kalidindi, O. Wodo, "How important is microstructural feature selection for data-driven structure-property mapping?", MRS Communications, 2022, 12, 95-103.

University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

From physical intuition to the set of descriptors

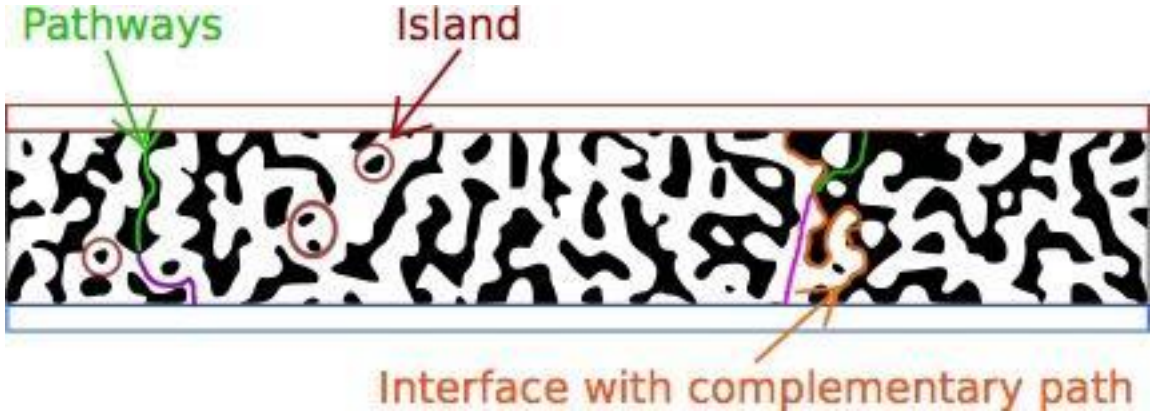


Characteristics of good morphology:

- bicontinuous interpenetrated network of domain: donor (D) and acceptor (A),
- domain size comparable with exciton diffusion length (10 nm),
- long interface between donor and acceptor,
- short and continuous pathways for charges to travel.

We need two types of information:

- Connectivity
- Distances



Welcome to GraSPI's documentation!

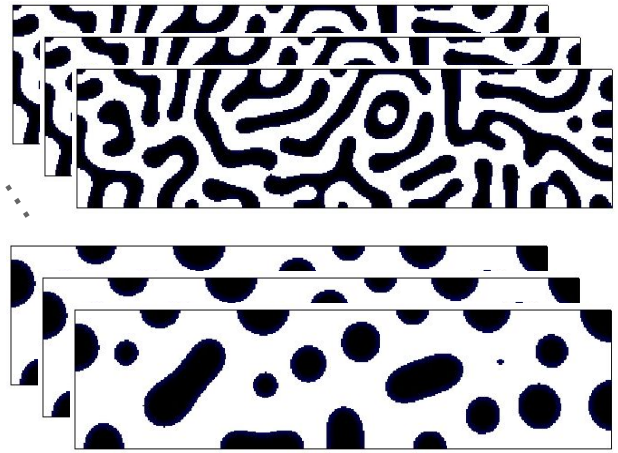
`GraSPI` is graph-based structure property identifier software implemented as a C/C++ package. `GraSPI` computes a library of descriptors for a segmented microstructure and computes descriptors that are relevant for organic solar cells performance. The package represents microstructure as a graph and harnesses the graph-based theory to compute wide range of descriptors at low computational cost. A suite of tools for data conversion between various formats and for post-processing the raw results from the graph analysis is provided.

Contents:

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

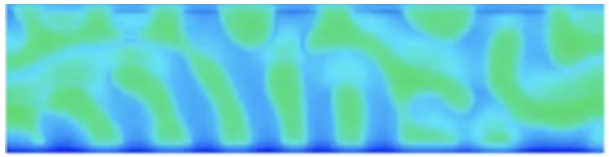
University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Three descriptors are sufficient to predict performance of microstructure

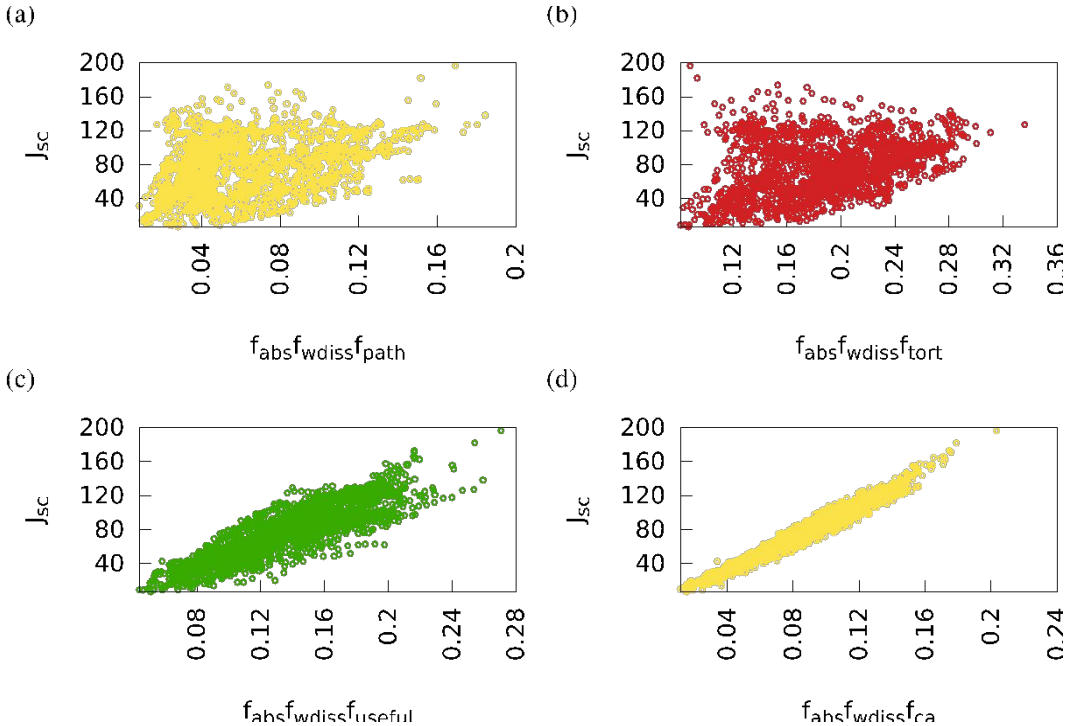


2k microstructure

$$\begin{bmatrix} d_1^1, d_2^1, \dots, d_{21}^1, P^1 \\ d_1^2, d_2^2, \dots, d_{21}^2, P^2 \\ \dots \\ d_1^N, d_2^N, \dots, d_{21}^N, P^N \end{bmatrix}$$



Property: J_{sc} from excitonic drift diffusion model



$$P = f(\widetilde{d_1}, \widetilde{d_2}, \widetilde{d_3})$$

$$J_{sc} = a_o + a_1 f_D f_D^{wdiss} \min(A_{An}^D, A_{Ca}^A)$$

$$\log(J_{sc}) = a_0 + a_1 \log(f_D) + a_2 \log(f_D^{wdiss}) + a_3 \log(\min(A_{An}^D, A_{Ca}^A))$$

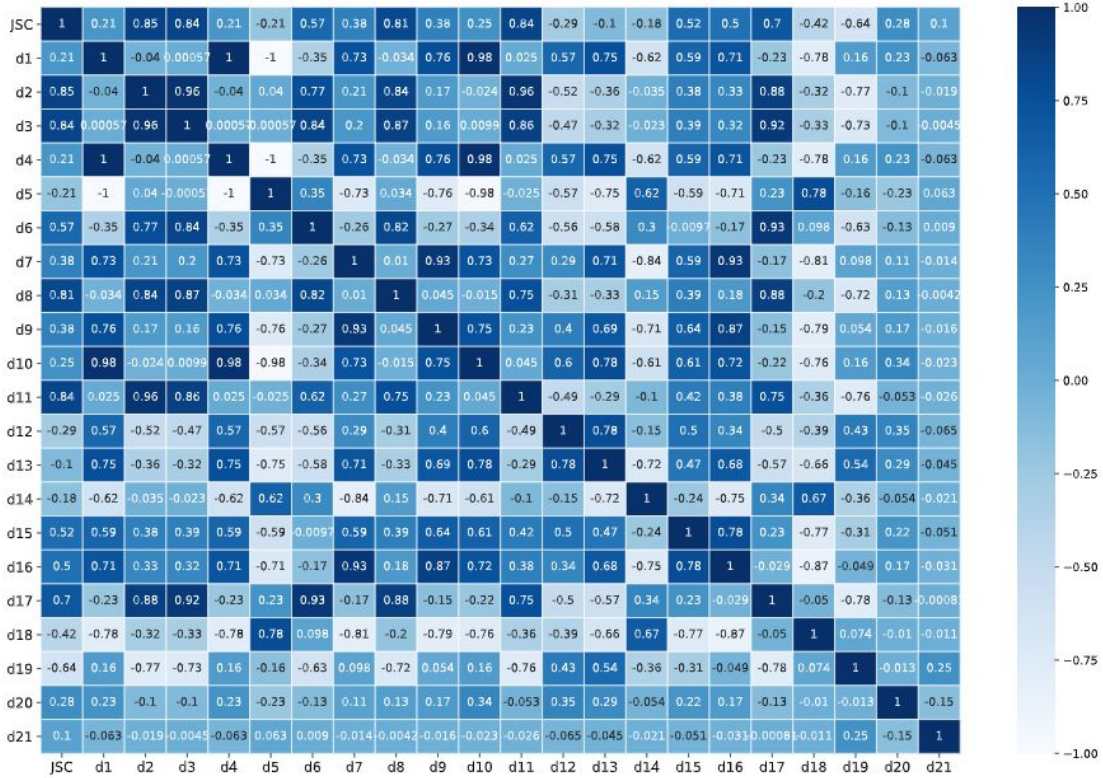
University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Descriptors are correlated with each other

d

d
,

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

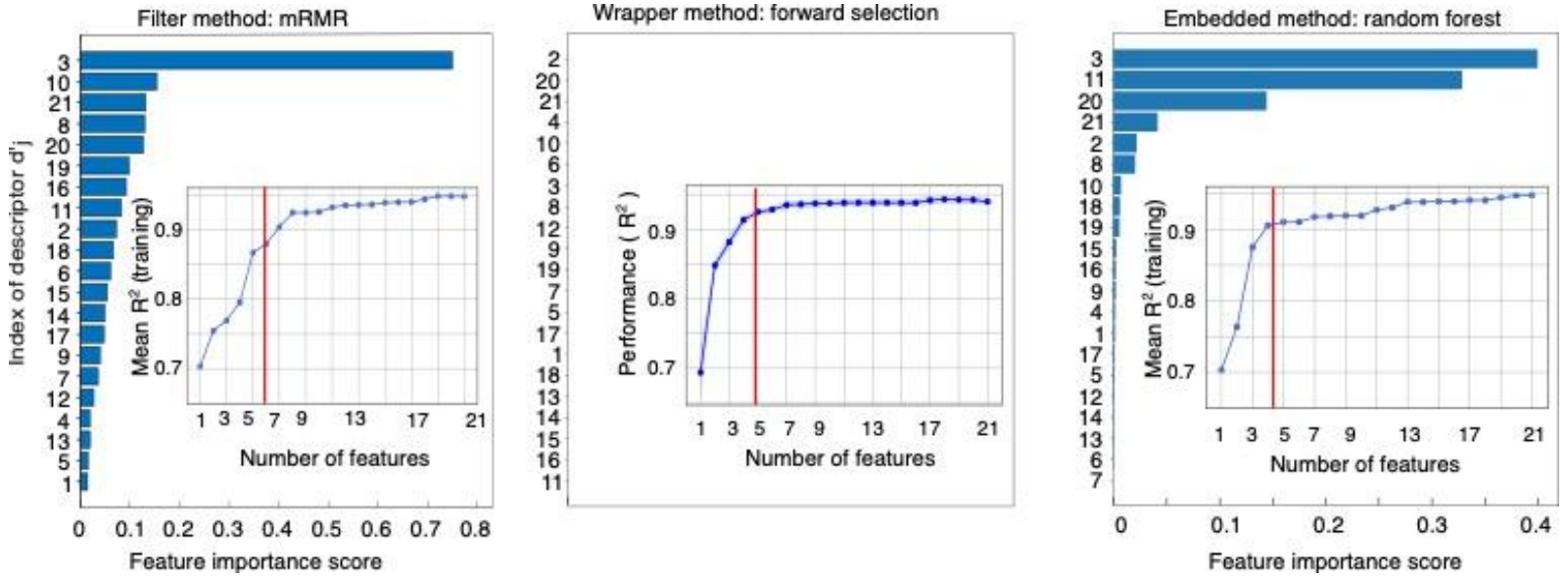


H. Liu, B. Yucel, D. Wheeler, B. Ganapathysubramanian, S. Kalidindi, O. Wodo, "How important is microstructural feature selection for data-driven structure-property mapping?", MRS Communications, 2022, 12, 95-103.

University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Enriching features with physics matters the most.

1700 mst (80/20), 21 desc, Jsc	Expert derived model	mRMR (human-derived)	mRMR (expert-enriched)	Feature engineering (2pt correlation of phase)	Feature engineering (2pt correlation+expert)
Features	d_{10}, d_2, $\min(d_{20},d_{21})$	d_3, d_{10}, d_8, d_{15}, d_{19}, d_{11}	d_3, d_{10}, d_{21}, d_8, d_{20}	f	f'
Number of features	4	6 (28)	5 (21)	7 (36)	7(36)
Model	M_E				
Performance (R^2)	0.97	0.81 (0.83)	0.89 (0.98)	0.75 (0.85)	0.87 (0.95)

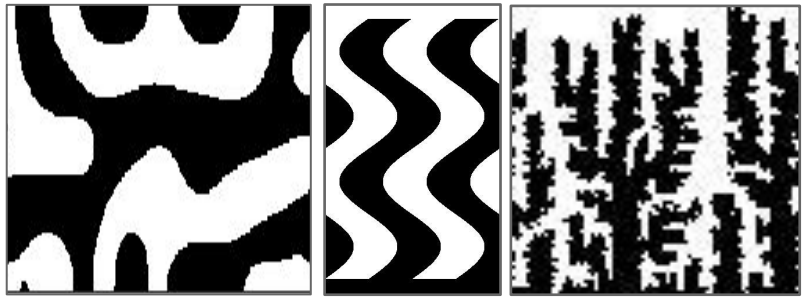


We improve the accuracy model by adding more features (expert informed features vs generic features)

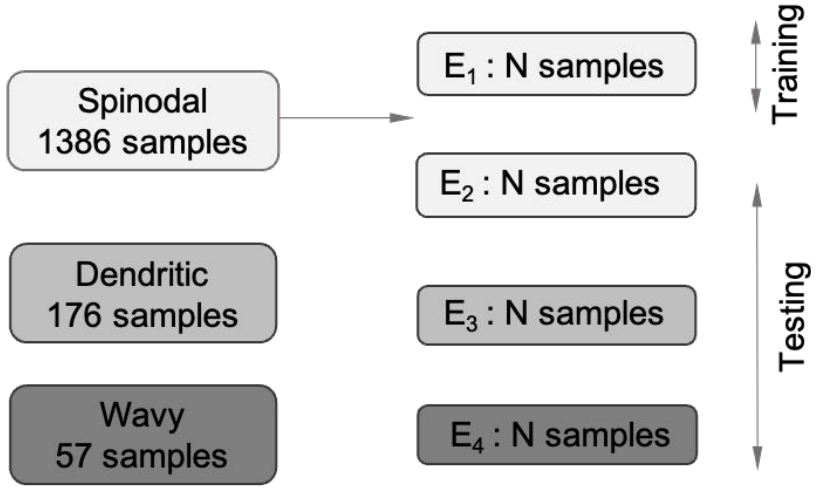
H. Liu, B. Yucel, D. Wheeler, B. Ganapathysubramanian, S. Kalidindi, O. Wodo, "How important is microstructural feature selection for data-driven structure-property mapping?", MRS Communications, 2022, 12, 95-103.

Is the model generalizable? Are salient features invariant?



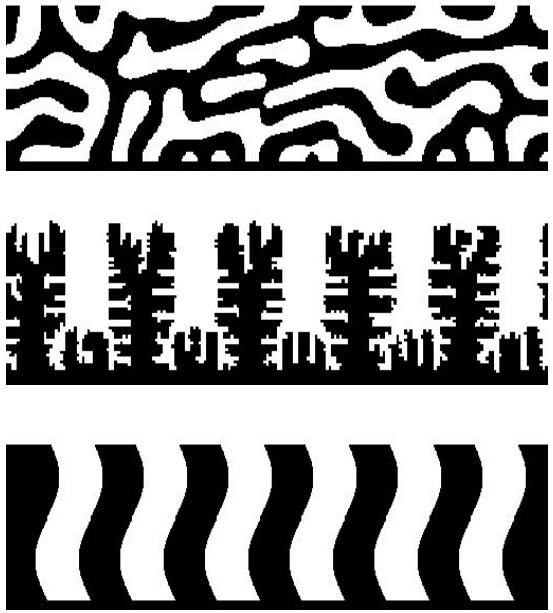



- **In-distribution generalization**: Training and testing distributions very similar - i.i.d. assumption (independent and identically distributed datasets)
- **Out-of-distribution generalization:** all other cases.



Can we find microstructural features that are invariant regardless of their context (datasets/environment)?

Three types of microstructure and three types of microstructure representation



Property:

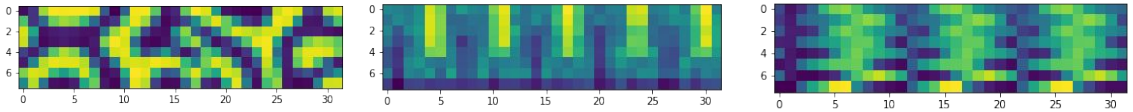
J_{sc}

"Computer simulation of heterogeneous polymer photovoltaic devices" HK Kodali, B Ganapathysubramanian Modelling and Simulation in Materials Science and Engineering 20 (3), 035015

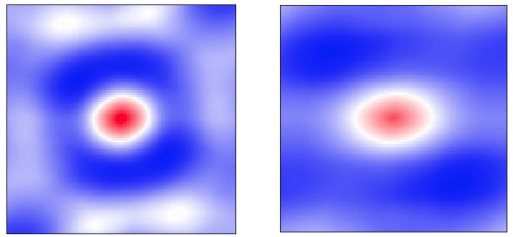
Descriptors

$[d_1,\ldots, d_{21}]$ □ 3PCs

Autoencoder latent space
32x8 → 3, 12, 30PCs



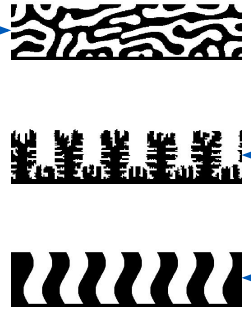
Two-point correlation function
512x128 → 3, 12, 30PCs

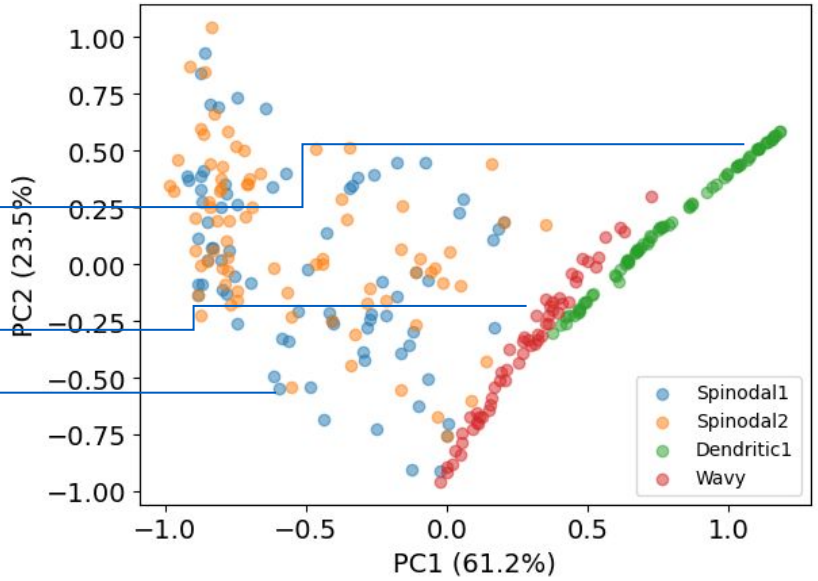


University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

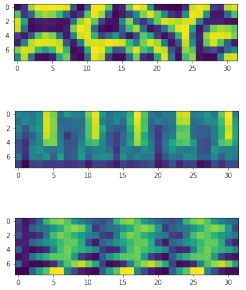
Visual assessment offers insight into generalizability of model

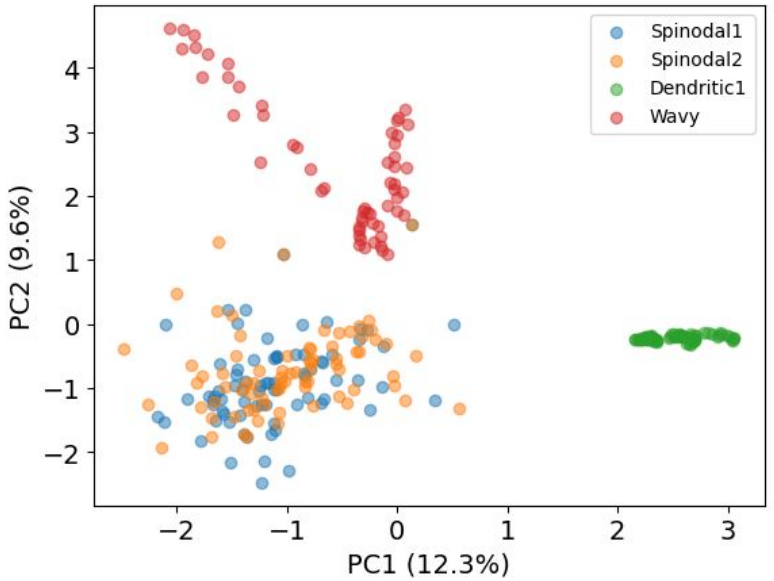
Descriptors



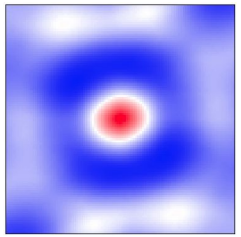


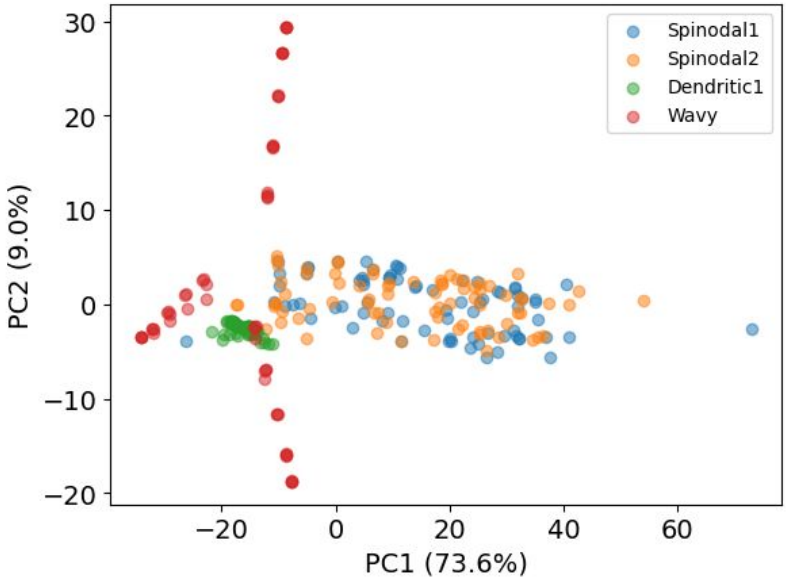
AE latent space





Two-point correlation function

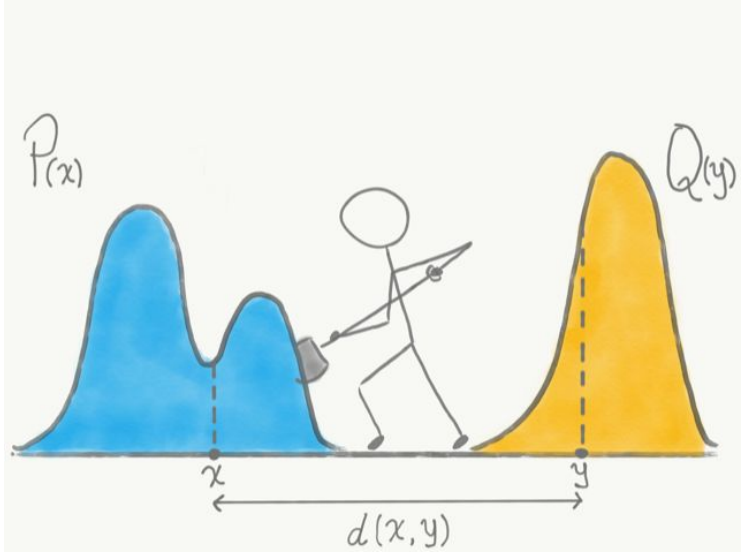




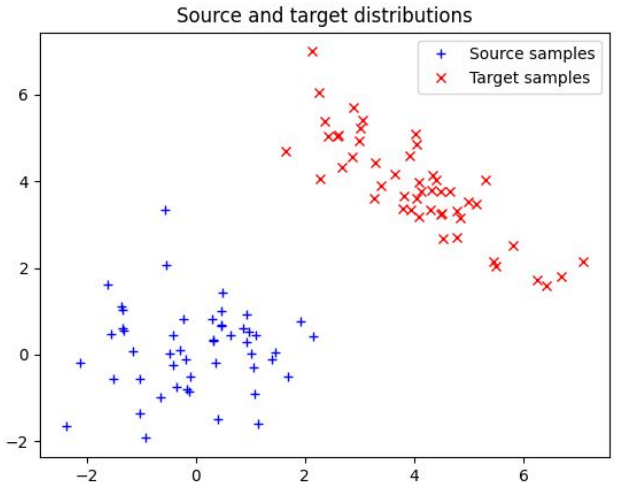
Capability to generalize

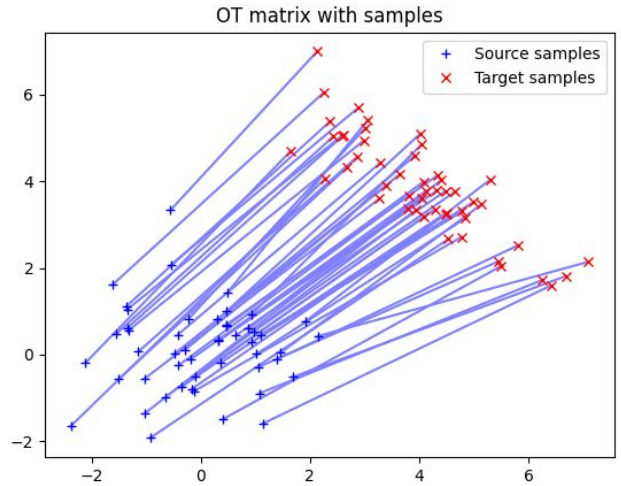
Descriptors > auto-encoder and 2pt correlation

Wasserstein distance between distributions offers quantitative insight into generalizability



https://tinyurl.com/u4khfdjd



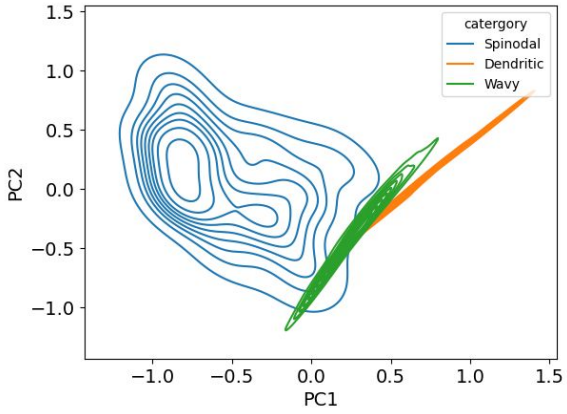


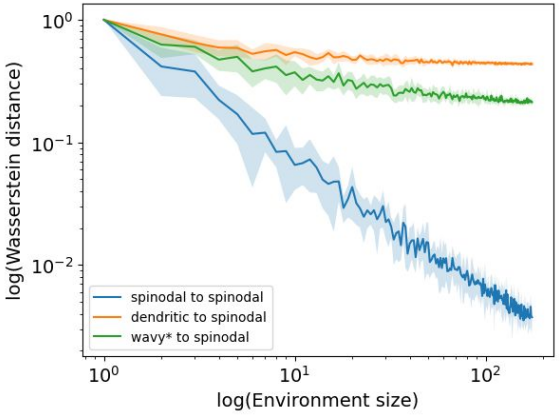
Wasserstein distance quantifies the distance between two distributions.
Optimal transport problem - the minimum work or cost required to transform one distribution into another

University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

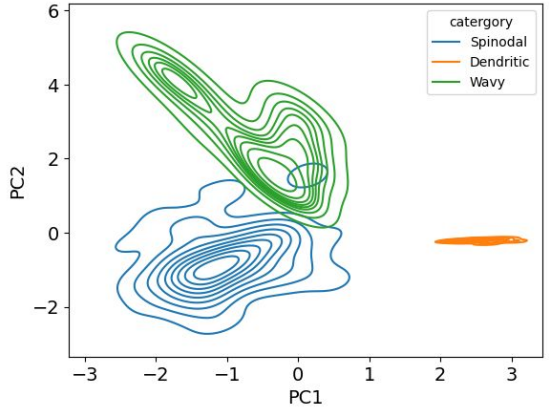
Descriptors show the highest potential to generalize the model for wavy but not dendritic type of microstructure

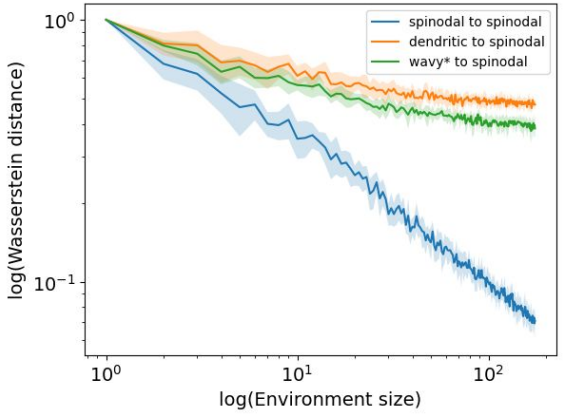
Descriptors



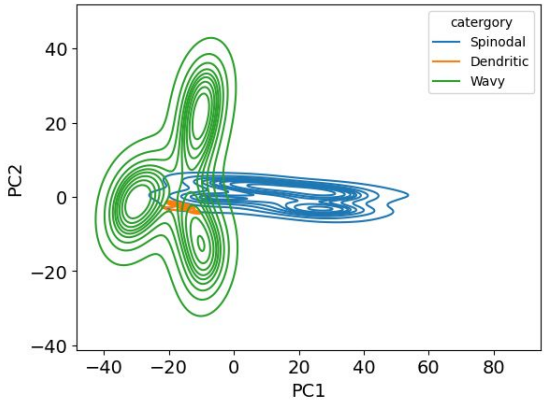


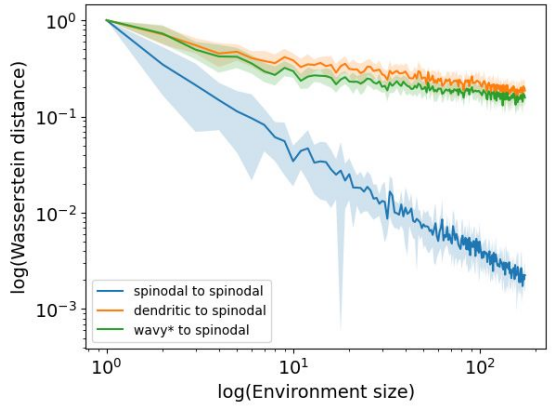
Autoencoder





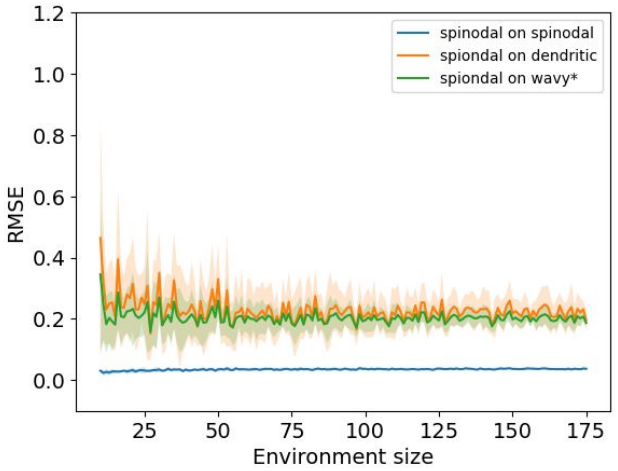
Two-point correlation function



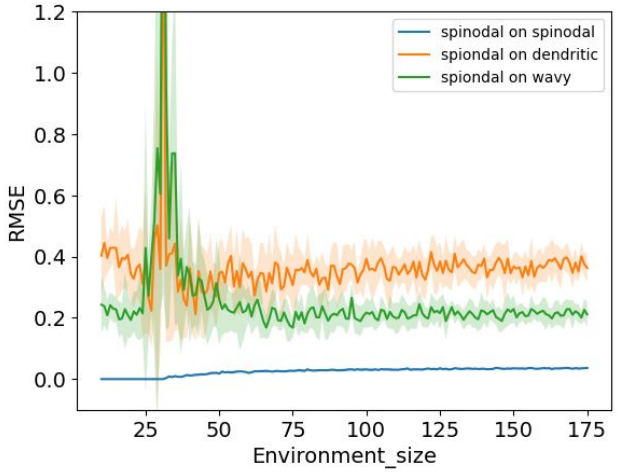


Model accuracy confirms our initial assessments

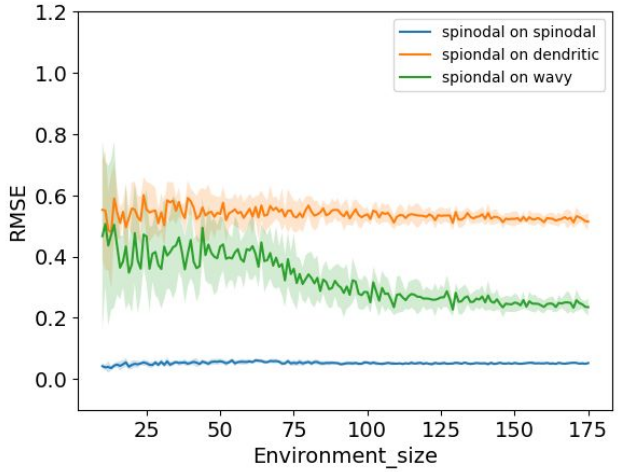
Descriptors



Autoencoder



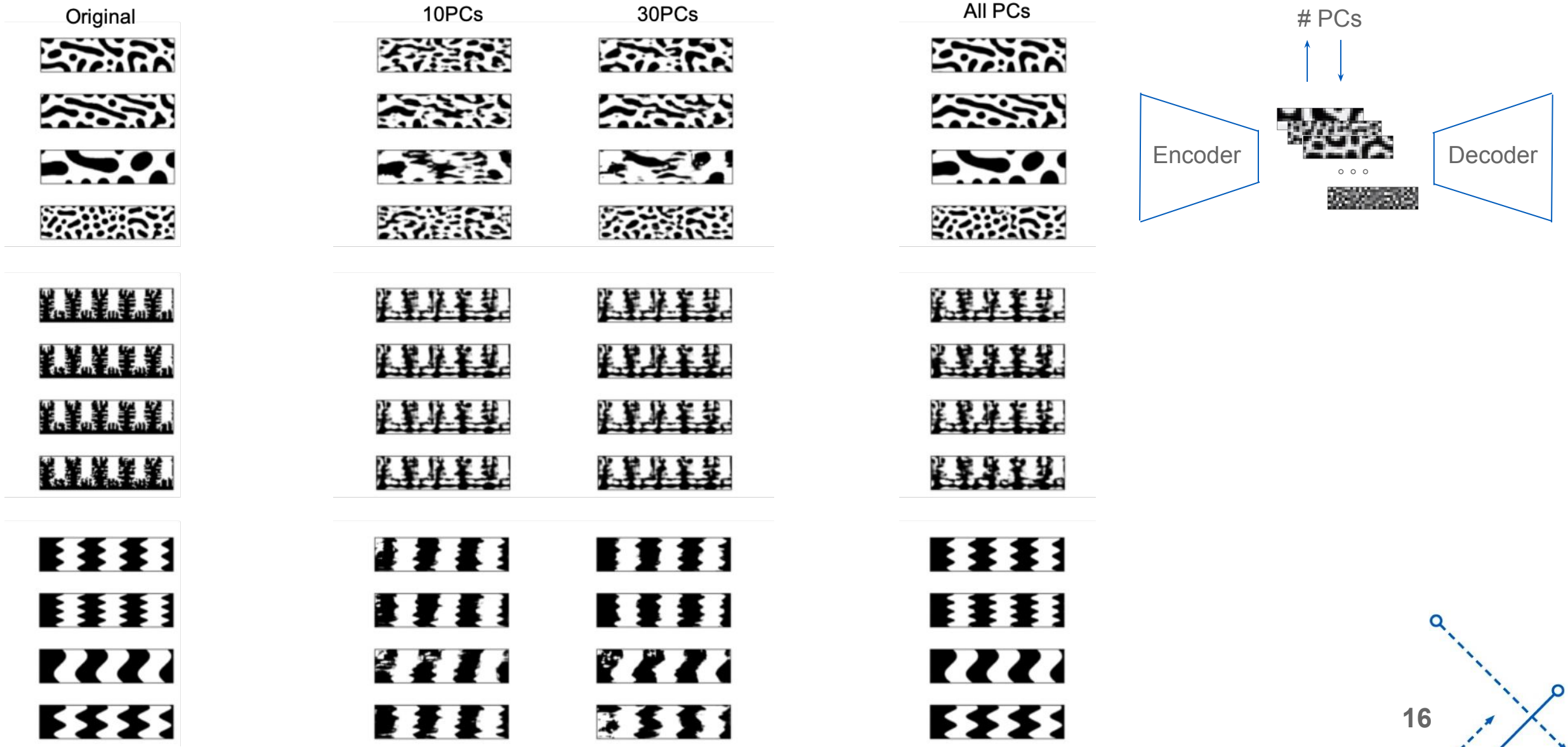
Two-point correlation function



Representation	Data type	Performance (RMSE, Environmental size=57)	Performance (RMSE, Environmental size=175)
Descriptors	Dendritic	0.26	0.20
	Wavy	0.22	0.19
Autoencoder (30PCs)	Dendritic	0.38	0.36
	Wavy	0.22	0.21
2pt correlation (3PCs)	Dendritic	0.52	0.51
	Wavy	0.38	0.23

University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Reconstruction error explain poor generalizability of autoencoders



University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Salient descriptors remain the same for wide range of material properties

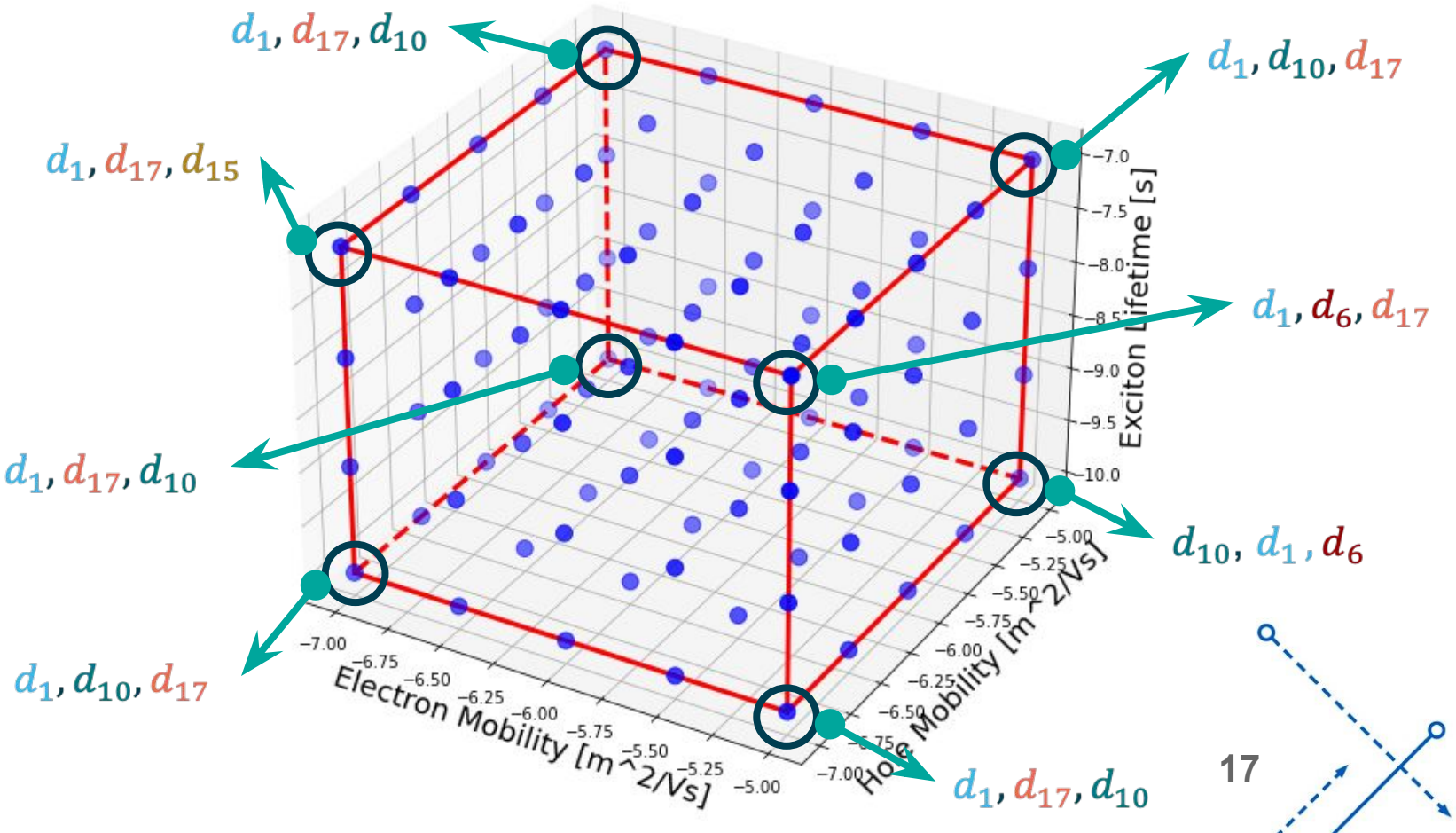
- Similar salient descriptors for wide range of properties
- 3 descriptors typically sufficient

Area of Interface

Distance to interface

Electrode connectivity

d_1: Interfacial area
d_6: Area of donor-anode contact
d_{10}: Fraction of donor within certain distance to interface
d_{15}: Interfacial area with paths to both electrodes
d_{17}: Volume of acceptor with straight path to cathode



University at Buffalo
Department of Materials Design and Innovation
School of Engineering and Applied Sciences
College of Arts and Sciences

Microstructure informatics for P=f(structure)

Challenge: micrographs (image) are high dimensional points (~10^6 dimensions)

[60% data curation, 30% features/data representation, 10% machine learning method]

Take home message:

- Physics-informed features matter more than representation.
- Descriptors as a representation offers the highest generalizability of SP map.
- Generalizable model across types of microstructure and materials system can be calibrated.